UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

BIOLOGIA ESTRUTURAL DE UREASES:

FILOGENIA, ATIVAÇÃO E PEPTÍDEOS DERIVADOS

Porto Alegre – Brasil

Março de 2014

Biologia Estrutural de Ureases:
Filogenia, Ativação e Peptídeos Derivados

Rodrigo Ligabue Braun

Tese de doutorado elaborada no Grupo de Bioinformática Estrutural do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul sob orientação do professor doutor

Hugo Verli

e no Laboratório de Proteínas Tóxicas do Departamento de Biofísica e do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul sob orientação da professora doutora

Célia Regina Ribeiro da Silva Carlini

Porto Alegre – Brasil

Março de 2014

Biologia Estrutural de Ureases:
Filogenia, Ativação e Peptídeos Derivados

Rodrigo Ligabue Braun

Tese submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul como parte dos requisitos necessários para a obtenção do grau de Doutor em Biologia Celular e Molecular.

BANCA EXAMINADORA:

Hugo Verli
Centro de Biotecnologia - UFRGS
Orientador

Célia Regina Carlini
Centro de Biotecnologia – UFRGS
Orientadora

Augusto Schrank
Centro de Biotecnologia
Universidade Federal do Rio Grande do Sul

Eduardo Eizirik
Centro de Biologia Genômica e Molecular
Pontifícia Universidade Católica do Rio Grande do Sul

Barbara Zambelli
Dipartimento di Farmacia e BioTecnologie
Università di Bologna

Nelson Jurandi Rosa Fagundes
Departamento de Genética
Universidade Federal do Rio Grande do Sul
[revisor]

Esta tese foi desenvolvida sob a orientação dos Professores Doutores Hugo Verli e Célia Regina Carlini, com o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Pró-Reitoria de Pós-Graduação (UFRGS), como requisito para obtenção do grau de Doutor em Biologia Celular e Molecular, junto ao Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul.

FICHA CATALOGRÁFICA

CIP - Catalogação na Publicação

Ligabue-Braun, Rodrigo
Biologia Estrutural de Ureases: Filogenia, Ativação e Peptídeos Derivados / Rodrigo Ligabue-Braun. -- 2014.
214 f.

Orientadora: Célia Regina Carlini.
Orientador: Hugo Verli.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia do Estado do Rio Grande do Sul, Programa de Pós-Graduação em Biologia Celular e Molecular, Porto Alegre, BR-RS, 2014.

1. ureases. 2. filogenia. 3. níquel. 4. peptídeos tóxicos. I. Carlini, Célia Regina, orient. II. Verli, Hugo, orient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo autor.

AGRADECIMENTOS

Em 1948, Mário Quintana disse que "além do controlado Dr. Jekyll e do desrecalcado Mr. Hyde, há também um chinês dentro de nós, o Sr. Wong. Nem bom, nem mau: gratuito. Entremos, por exemplo, neste teatro. Tomemos este camarote. Pois bem, enquanto o Dr. Jekyll, muito compenetrado, é todo ouvidos, e Mr. Hyde arrisca um olho e a alma no decote da senhora vizinha, o nosso Sr. Wong, descansadamente, põe-se a contar carecas na platéia…". Assim, quero agradecer, antes de qualquer coisa, a todos que conviveram comigo nos últimos anos, por aceitarem o meu Sr. Wong, sempre substituindo o Dr. Jekyll e o Mr. Hyde (que saíram de férias em algum momento e ainda não voltaram).

Agradeço à Célia e ao Hugo pelo exemplo, pela dedicação, pela compreensão, pelas segundas chances, pelas oportunidades e por todas as revelações sobre o lado B da ciência no Brasil e no mundo. Nem sempre é fácil servir a dois senhores, mas vocês facilitaram para mim. Obrigado!

Aos membros da minha comissão de acompanhamento no PPGBCM, Dra. Marilene Vainstein e Dr. Hubert Stassen, agradeço por estarem sempre disponíveis para discutir o andamento do trabalho (além de todas as outras dúvidas aleatórias que surgiram pelo caminho). Agradeço também aos membros da banca examinadora, que me lisonjearam ao aceitar o convite para avaliar esse trabalho. Espero não decepciona-los.

Sei que vai ficar chato não citar nominalmente todos os meus colegas Laprotoxianos e GBEnses, mas não posso fazer muito a esse respeito. Agradeço de verdade a cada um, mesmo que não tenham tido nenhum envolvimento no que eu fiz. Eu aprendi com vocês (mudar a cor do título de um PDF conta como aprendizado) e o convívio foi fantástico. Obrigado a todos por rirem comigo, chorarem comigo, me guiarem e me colocarem no meu lugar sempre que necessário.

Cortez-Lopes, Gottschald, Lemelle, Sachett: muito obrigado por tornarem meus dias mais alegres, as aulas mais divertidas, e as discussões mais científicas. Prefiro não pensar como teria sido fazer esse trabalho sem vocês por perto (OK, provavelmente ele nem teria existido).

(Parabéns! Você já leu uma página de agradecimentos!)

Sr. Carrer Andreis, meu IC favorito! Muito obrigado por embarcar em um projeto sem saber que ele era impossível (e por conclui-lo!). Tenho certeza de que ainda vais crescer muito, e queria deixar isso registrado aqui, só para poder dizer “eu já sabia” :)

Ringrazio il Dott. Stefano Ciurli, il Dott. Francesco Musiani, e la Dott.ssa Barbara Zambelli, per il appoggio e l'insegnamento, mentre ero nella Università di Bologna (nonostante il fatto che nessuna cosa che ho fatto lì è su questa tesi). Ringrazio anche a Val Broll per tutti le altre cose che abbiamo fatto in Italia.

Agradeço especialmente à Silvia Centeno e ao Luciano Saucedo, cujo trabalho excede absurdamente o que seria esperado de uma secretaria. Muito obrigado pelos lembretes, e-mails repassados, telefonemas, quebradas de galho e fofocas cotidianas!

Westphalen, Maletich, Paiva, Lindenau, Souza, Camargo, Pivetta: amizades de muito antes e durante o doutorado. Obrigado pelas experiências novas que vocês me propõem sempre que nos vemos, e por me lembrarem que existe vida fora da universidade.

À minha família – pai, mãe, irmã – muito obrigado! Por estarem sempre dispostos a me fazer ver a vida por um ângulo diferente. Por ficarem ao meu lado. Por existirem e me amarem, basicamente.

Agradeço finalmente ao Sandro. Os últimos quatro anos foram uma montanha-russa e tu estiveste *ready for the ride* o tempo todo, segurando a minha mão em cada momento de fraqueza, me inspirando a continuar mesmo quando tudo parecia perdido. Longe ou perto, chorando ou sorrindo, tu ficaste ao meu lado. Por isso e um bilhão de outras coisas, obrigado! *J'ai besoin de toi, seulement toi, et en plus je t'aime*. Tradução simultânea, para deixar registrado: te amo, Sandro. Obrigado por estar na minha vida.

"Interacting with the natural world, we are denied certainty. And always will be."

Michael Crichton

Este trabalho é dedicado a todas as minorias que foram (e são) vítimas de preconceito travestido de ciência.

SUMÁRIO

LISTA DE ABREVIATURAS

BI	inferência Bayesiana
EC	Enzyme Comission
ML	*maximum likelihood*, máxima verossimilhança
NCBI	National Center for Biotechnology Information
NMA	Análise de Modos Normais
PDB	Protein Data Bank
RMN	ressonância magnética nuclear
SAXS	espalhamento de raios-X a baixo ângulo
TIM	triosefosfato isomerase
UniProt	Universal Protein Resource
UPGMA	Unweighted Pair Group Method with Arithmetic Average

RESUMO

Ureases são enzimas níquel-dependentes que catalisam a hidrólise da ureia em amônia e dióxido de carbono. Além disso, apresentam diversas propriedades independentes da catálise, sendo consideradas proteínas *moonlighting*. São amplamente distribuídas na natureza, sendo encontradas em bactérias, arqueas, plantas e fungos, podendo se organizar em unidades funcionais compostas por uma, duas ou três subunidades. Sua ativação, que envolve a passagem da enzima de sua forma apo-urease a sua forma holo-urease, requer pelo menos três proteínas acessórias. Parte de suas propriedades não-catalíticas é associada à liberação de peptídeos internos da proteína nativa. Nesse contexto, a presente tese se dedicou ao estudo de diferentes aspectos da biologia estrutural de ureases. Ao elaborar uma narrativa filogenética, envolvendo varredura de bancos de dados em larga escala e diferentes algoritmos de reconstrução de árvores, foi possível propor uma rota evolutiva indicando a transição de três subunidades para uma única unidade funcional, sem passar por intermediários de duas cadeias. Quanto ao seu processo de ativação, por meio de múltiplos cálculos de atracamento baseados em dados experimentais prévios (especialmente SAXS), foram propostas estruturas para suas diferentes etapas, em resolução atomística. Finalmente, o comportamento dinâmico de diferentes peptídeos derivados de urease foram analisados computacionalmente por meio de simulações de longa duração (500 ns) e associados a outros dados obtidos *in vitro*, permitindo justificar efeitos diferenciais obtidos na aplicação destes peptídeos. De maneira geral, o trabalho empregou técnicas computacionais à análise de ureases, fornecendo bases para futuros estudos de suas propriedades, sejam catalíticas ou não, incluindo sua aplicação biotecnológica.

ABSTRACT

Ureases are nickel-dependent enzymes that catalyze the hydrolysis of urea into ammonia and carbon dioxide. They have many catalysis-independent properties, being considered moonlighting proteins. Ureases are found in bacteria, archaea, plants, and fungi, and may be organized in functional units composed by one, two, or three subunits. Their activation, involving the transition from apo to holourease, requires at least three accessory proteins. Some of their non-catalytic properties are related to the release of internal peptides from the native protein. In this context, this thesis was developed upon the study of different aspects of urease structural biology. By phylogenetical reconstruction, employing large-scale databank scans and different tree-building algorithms, we were able to propose an evolutionary route by which the transition from three to one functional subunits was possible, with no need for a two-chained intermediate. Regarding the activation mechanism, we have proposed structural models for the oligomeric intermediates of the process by multiple docking calculations, at atomistic resolution, based on previous experimental data (especially SAXS). Additionally, the dynamical behavior of different urease-derived peptides was analyzed by computational simulations at large time scales (500 ns) and correlated to *in vitro* results, allowing us to explain the variable effects observed after their application on test systems. In short, in this work we have employed computational techniques to the analysis of urease, providing working grounds for further studies of this enzyme and its properties (catalytical or not), including its biotechnological applicability.

ÍNDICE DE FIGURAS

1. INTRODUÇÃO

"I centrifuged some of the filtrate and looked at the precipitate under the microscope. It appeared to be crystalline. It had extremely high urease activity. I telephoned to my wife that I probably had obtained crystalline urease. That night I slept but little."

James B. Sumner

1.1 UREASES: ASPECTOS HISTÓRICOS

Ureases (ureia amidohidrolases, EC 3.5.1.5) são enzimas amplamente distribuídas na natureza, sendo encontradas em plantas, fungos e bactérias. Elas se caracterizam por uma mesma propriedade catalítica, a hidrólise da ureia a amônia e ácido carbâmico (Figura 1) (KRAJEWSKA, 2009A).

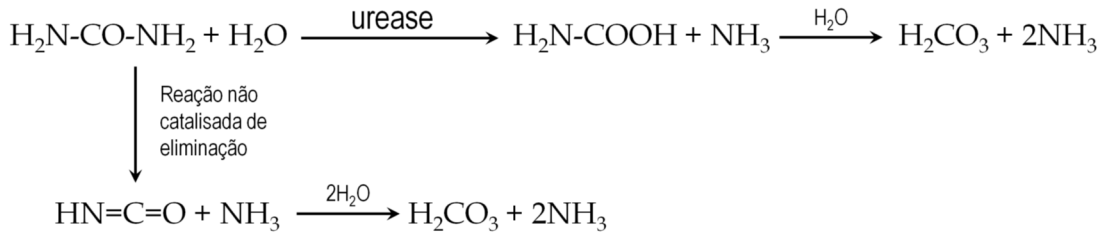


Figura 1. Esquema da reação catalisada pela urease: hidrólise da ureia produzindo carbamato e amônia, com o primeiro sendo hidrolisado espontaneamente a outra molécula de amônia e ácido carbônico (adaptado de KRAJEWSKA, 2009A).

Tanto a ureia quanto as ureases representam marcos da investigação científica (MOBLEY ET AL. 1995). Conforme levantamento realizado por FEARON (1923), a ureia foi descoberta em 1773, em urina humana, sendo posteriormente, o primeiro composto orgânico a ser sintetizado a partir de compostos inorgânicos, por WÖHLER (1828). Em 1798, foi proposto que a amônia da urina provinha da fermentação da ureia, mas o isolamento do primeiro microorganismo ureolítico (*Micrococcus ureae*), deu-se apenas em 1864. Em 1874, uma enzima

ureolítica foi obtida a partir de urina pútrida, sendo tal enzima chamada, a partir de 1890, de urease.

A urease passou a estar disponível em grandes quantidades a partir de 1909, quando Takeuchi e Inone descobriram que sementes de soja (*Glycine max*) eram ricas nesta enzima. Este trabalho foi, também, o primeiro a identificar a urease nos chamados vegetais superiores. Em 1916, Matter e Marshall descobriram que o feijão-de-porco, ou *jack bean* (*Canavalia ensiformis*), apresentava até quinze vezes mais urease que a soja, promovendo-o a material ideal para estudos dessa enzima (dados retirados de SUMNER, 1937).

Em 1926, James Sumner cristalizou a urease a partir de feijão-de-porco (SUMNER, 1926). Seu trabalho teve importância fundamental para a Bioquímica por demonstrar, pela primeira vez, que enzimas são proteínas e que podem ser cristalizadas. À época (e até meados da década de 1930), as enzimas eram consideradas compostos diferentes de todos os outros conhecidos até então (sendo descritas como colóides possivelmente associados a proteínas), o que dificultou a aceitação e a divulgação adequada deste trabalho pioneiro (SUMNER, 1937). Em 1946, Sumner recebeu o Prêmio Nobel pela “descoberta de que enzimas podem ser cristalizadas” (MANCHESTER, 2004).

Em 1975, foi demonstrado que a urease de feijão-de-porco possuía íons de níquel em seu sítio ativo e que estes eram essenciais para catálise (DIXON ET AL., 1975). Até então, o níquel era tratado como um metal sem relevância biológica (THAUER, 2001), mesmo quando consideradas as dificuldades analíticas que impediam uma análise adequada deste metal naquela época (ZERNER, 1991). Além da urease, somente outras oito enzimas contendo níquel foram descritas até o momento (RAGSDALE, 2009; SYDOR & ZAMBLE, 2013; BOER ET AL., 2013).

Desde a descoberta de sua presença em soja, em 1909, a urease tem sido alvo de intensas pesquisas, incluindo análises de sua ocorrência e papel na natureza; mecanismo e especificidade de ação; interação com compostos diversos; seqüenciamento; organização gênica e cristalização (QIN & CABRAL, 2002; KRAJEWSKA, 2009A). Um trabalho desenvolvido por nosso grupo descreveu uma neurotoxina oriunda do feijão-de-porco, identificada como uma isoforma da urease majoritária da planta (CARLINI & GUIMARÃES, 1981; FOLLMER ET AL., 2001). Suas propriedades neurotóxicas mostraram-se independentes de sua

capacidade ureolítica, o que deu início à busca de outras propriedades catálise-independentes em ureases (CARLINI & POLACCO, 2008).

Considera-se que o papel principal de ureases seja o de permitir que os organismos utilizem a ureia, endógena ou exógena, como fonte de nitrogênio (JABRI ET AL., 1995; QIN & CABRAL, 2002; KRAJEWSKA, 2009A). Além deste papel, todavia, as ureases apresentam grande versatilidade de aplicações, podendo atuar no transporte de nitrogênio, na defesa em plantas e na neutralização da acidez no ambiente gástrico de mamíferos para colonização bacteriana (CARLINI & POLACCO, 2008; FOLLMER, 2010).

Além das plantas, arqueas, bactérias e fungos sintetizam ureases. De maneira geral, considera-se que todos os animais perderam essa enzima em sua história evolutiva (FUJIWARA & NOGUCHI, 1995), apesar de algum autores proporem sua existência em moluscos e cordados basais (PEDROZO ET AL. (1996AB; XUE ET AL., 2006).

As ureases têm papel importante na patogênese de várias espécies bacterianas, incluindo *Proteus mirabilis*, *Staphylococcus saprophiticus*, *Yersinia enterocolitica* e *Ureaplasma urealiticum* (MOBLEY ET AL., 1995). O exemplo mais frequente na literatura é o da urease de *Helicobacter pylori*, devido à alta prevalência desse microrganismo como patógeno humano e pelo papel essencial da enzima em sua patogênese (EATON ET AL., 1991, SIRKO & BRODZIK, 2000, OLIVERA-SEVERO ET AL., 2006, FOLLMER, 2010).

A atividade ureásica foi encontrada em várias espécies de fungos, mas a descrição genética correspondente está disponível para apenas algumas delas (SIRKO & BRODZIK, 2000; KRAJEWSKA, 2009A). Exemplos incluem *Schizosaccharomyces pombe* (TANGE & NIWA, 1997) e os patógenos humanos *Coccidioides immitis* (YU ET AL., 1997); e *Cryptococcus neoformans* (COX ET AL., 2000; SINGH ET AL., 2013).

Em relação a ureases de plantas, os dados genéticos mais abrangentes estão disponíveis para soja (*G. max*) (POLACCO & HOLLAND, 1993, POLACCO & HOLLAND, 1994). Genes distintos, que codificam duas isoformas da urease, uma ubíqua e uma embrião-específica, assim como genes não ligados, codificadores de proteínas acessórias, foram identificados nessa planta (MEYER-BOTHLING & POLACCO, 1987). A urease embrião-específica é uma proteína abundante em sementes de várias espécies de plantas, incluindo soja, feijão-de-porco

(POLACCO & HOLLAND, 1994) e *Arabidopsis thaliana* (ZONIA ET AL., 1995), enquanto a forma ubíqua é encontrada em baixas quantidades nos tecidos vegetativos da maioria das plantas (SIRKO & BRODZIK, 2000). Recentemente, uma possível terceira isoforma foi proposta com base no genoma de soja (WITTE, 2011).

A urease das folhas da amoreira *M. alba* já foi purificada e caracterizada (HYRAIAMA ET AL., 2000B), bem como a urease das sementes do algodoeiro, *Gossypium hirsutum* (MENEGASSI ET AL., 2008). Entretanto, bioquimicamente, a urease vegetal melhor caracterizada é a isoforma majoritária da urease de *C. ensiformis*, JBURE-I (TAKISHIMA ET AL., 1988, RIDDLES ET AL., 1991, HIRAI ET AL., 1993, KARMALI & DOMINGOS, 1993). Sabe-se menos a respeito das isoformas minoritárias, sendo reconhecidas pelo menos duas: a canatoxina e a JBURE-II (CARLINI ET AL., 1997; FOLLMER ET AL., 2001, PIRES-ALVES ET AL., 2003, MULINARI ET AL., 2011).

1.2 ESTRUTURA DE UREASES

As ureases de plantas e fungos são proteínas homo-oligoméricas (isto é, formadas por subunidades idênticas), enquanto as ureases bacterianas são multímeros formados por complexos de duas ou três subunidades (MOBLEY ET AL., 1995, TANGE & NIWA, 1997). Observa-se uma similaridade significativa entre as sequências peptídicas de todas as ureases conhecidas (SIRKO & BRODZIK, 2000).

Resíduos amino-terminais dos monômeros das enzimas fúngicas e vegetais são similares aos das subunidades menores das enzimas bacterianas, enquanto as subunidades maiores de ureases bacterianas assemelham-se às porções carboxiterminais de ureases de plantas e fungos (Figura 2). A alta similaridade entre as sequências de ureases indica que são todas variantes de uma mesma enzima e que, provavelmente, possuem estruturas terciárias e mecanismos catalíticos similares (MOBLEY ET AL., 1995), o que permite que dados obtidos para ureases bacterianas sejam extrapolados para as enzimas vegetais ou fúngicas (FOLLMER, 2008).

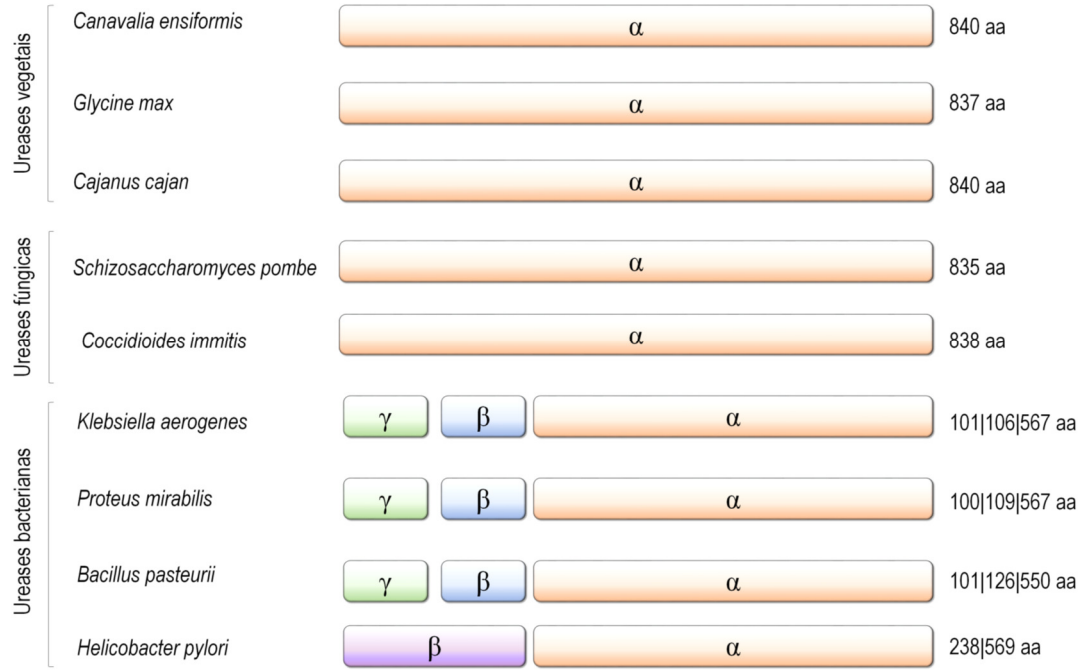


Figura 2. Representação esquemática das subunidades de urease (adaptado de SIRKO & BRODZIK, 2000; e KRAJEWSKA, 2009A).

As subunidades idênticas que compõem as ureases de plantas e fungos (com aproximadamente 90 kDa cada) são geralmente organizadas em trímeros α_3 ou hexâmeros α_6 (KRAJEWSKA, 2009A). Exemplos de ureases vegetais homo-hexaméricas incluem as enzimas de soja (*G. max*), feijão-de-porco (*C. ensiformis*), feijão-guandu (*Cajanus cajan*) e sementes de algodoeiro (*G. hirsutum*) (POLACCO & HAVIR, 1979, DAS ET AL., 2002, MENEGASSI ET AL., 2008). Organizações alternativas incluem α_2 para canatoxina, urease de amoreira (*M. alba*) e do fungo *S. pombe* e α_4 para urease do fungo *C. immitis* (LUBBERS ET AL., 1996, HYRAIAMA ET AL., 2000B, FOLLMER ET AL., 2001, MIRBOD ET AL., 2002).

As ureases bacterianas são compostas por três subunidades distintas, uma maior (α, 60-76 kDa) e duas menores (β, 8-21 kDa e γ, 6-14 kDa) (Figura 2), geralmente formando trímeros do tipo $(\alpha\beta\gamma)_3$ (SIRKO & BRODZIK, 2000, KRAJEWSKA, 2009A). Os exemplos típicos são as ureases de *Klebsiella aerogenes* (JABRI ET AL., 1995) e de *Bacillus pasteurii* (BENINI ET AL., 1999). Organizações alternativas incluem $(\alpha\beta\gamma)_4$ para a enzima de *Staphylococcus saprophyticus* (SCHÄFER & KALTWASSER, 1994) e $(\alpha\beta\gamma)_5$ para a de *Staphylococcus leei* (JIN ET AL., 2004). Em contraste à organização geral das ureases bacterianas, as ureases de *Helicobacter* spp. são compostas por duas subunidades, α (61-66

kDa) e β (26-31 kDa), que em *H. pylori* e *H. mustelae* se organizam em um complexo dodecamérico $((\alpha\beta)_3)_4$ (HA ET AL., 2001; CARTER ET AL., 2011).

Algumas ureases bacterianas e vegetais tiveram suas estruturas tridimensionais resolvidas por cristalografia de raios X. As estruturas das enzimas de *K. aerogenes*, *B. pasteurii*, e *H. pylori* foram as primeiras ser determinadas e analisadas (JABRI ET AL., 1995, BENINI ET AL., 1999, HA ET AL., 2001). Apesar de ter sido a primeira enzima cristalizada (SUMNER, 1926), apenas recentemente foi descrita a estrutura tridimensional da urease do feijão-de-porco, *C. ensiformis* (BALASUBRAMANIAN & PONNURAJ, 2010). A essa, seguiu-se a cristalização e descrição estrutural da urease de feijão-guandu, *C. cajan* (BALASUBRAMANIAN ET AL., 2013A). Uma listagem detalhada das estruturas tridimensionais de ureases e suas variantes, depositadas no Protein Data Bank, encontra-se no Apêndice A.

As ureases apresentam um enovelamento típico das amidohidrolases, caracterizado pela presença de um barril $(\alpha\beta)_8$ (barril TIM) ligeiramente distorcido, onde se localiza o sítio ativo com dois átomos de níquel, seguido por um conjunto de fitas β antiparalelas (NAGANO ET AL., 2002), ambos formando a subunidade α. No caso de ureases de plantas e fungos, essa porção corresponde ao domínio α, podendo ser separado em sub-domínios (α_1 correspondendo ao barril TIM e α_2 correspondendo ao conjunto de fitas) (BALASUBRAMANIAN & PONNURAJ, 2010; BALASUBRAMANIAN ET AL., 2013A). As subunidades β e γ apresentam predominantemente estruturas do tipo αβ (JABRI ET AL.,1995, BALASUBRAMANIAN ET AL., 2013A) (Figura 3).

Uma estrutura do tipo hélice-volta-hélice forma uma aba (*flap*) sobre o sítio ativo, sendo associada à alta especificidade das ureases pelo seu substrato, a ureia (JABRI ET AL., 1995). Esta aba encontrar-se-ia aberta enquanto a enzima não recebesse do meio o substrato, fechando-se de forma simultânea à entrada da ureia no sítio ativo (BENINI ET AL., 1999, BENINI ET AL., 2001, BALASUBRAMANIAN ET AL., 2013). A movimentação desta aba seria facilitada por regiões de alças, consideradas conformacionalmente mais flexíveis.

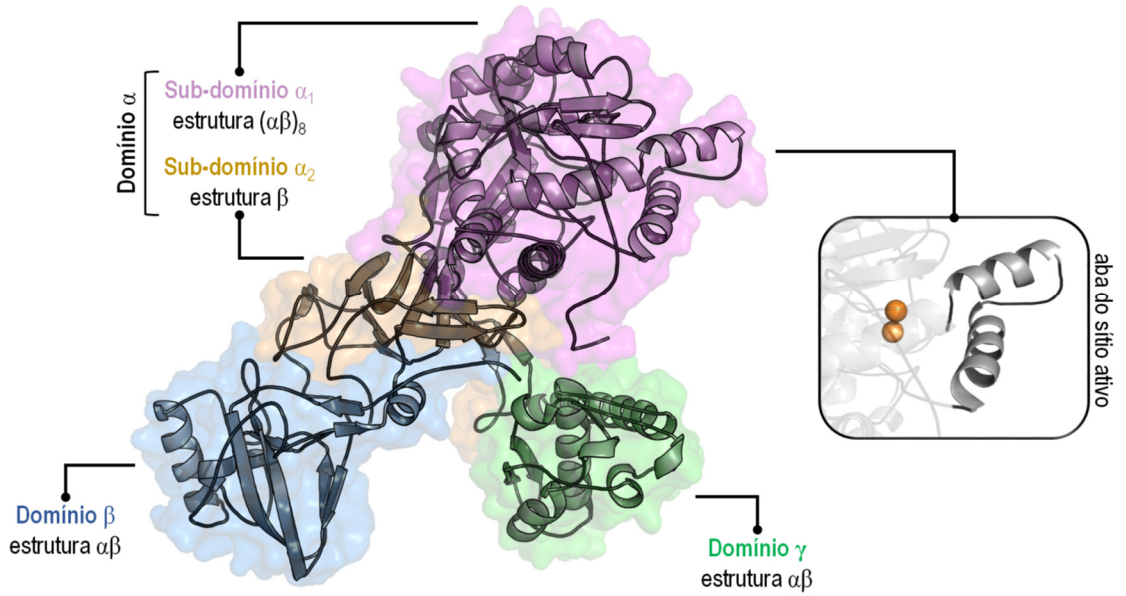


Figura 3. Organização estrutural de ureases, representada com base na estrutura cristalográfica da enzima de *C. cajan* (PDB ID 4G7E). Átomos de níquel representados como esferas em laranja.

1.3 Sítio catalítico e ativação de ureases

A resolução das estruturas cristalográficas de ureases bacterianas e vegetais permitiu uma análise detalhada do sítio ativo destas enzimas. Estes são praticamente sobreponíveis, indicando que um mesmo perfil de enovelamento deva ser comum a todas as ureases (Krajewska, 2009A; Balasubramanian et al., 2013A). O sítio ativo possui um centro binuclear com níquel. Casos onde há substituição natural deste por ferro ou zinco foram relatados, mas sempre com atividade muito reduzida em relação a enzimas equivalentes com níquel no sítio ativo (Carter et al., 2011B; Follmer et al., 2002). Os íons de níquel^{II} são ligados à cadeia lateral da lisina carbamilada por meio de seus átomos de oxigênio. O Ni(1), é também coordenado a duas histidinas, por meio de seus átomos de nitrogênio (N_ε em uma delas e N_δ em outra), enquanto o Ni(2), coordena a átomos de nitrogênio de duas histidinas (N_ε) e ao oxigênio do ácido aspártico (O_δ) (Figura 4). Além disso, os íons de Ni^{II} estão conectados a um íon hidróxido (WB, hidróxido-ponte) que, juntamente com duas moléculas de água, W1 no Ni(1) e W2 no Ni(2), e uma terceira molécula de água, W3, próxima à saída do sítio ativo, formam um grupamento tetraédrico de água. Esse grupamento (*cluster*) é unido por ligações de hidrogênio e preenche a cavidade

do sítio ativo, sendo substituído pela ureia quando da reação catalisada pela urease (BENINI ET AL., 1999).

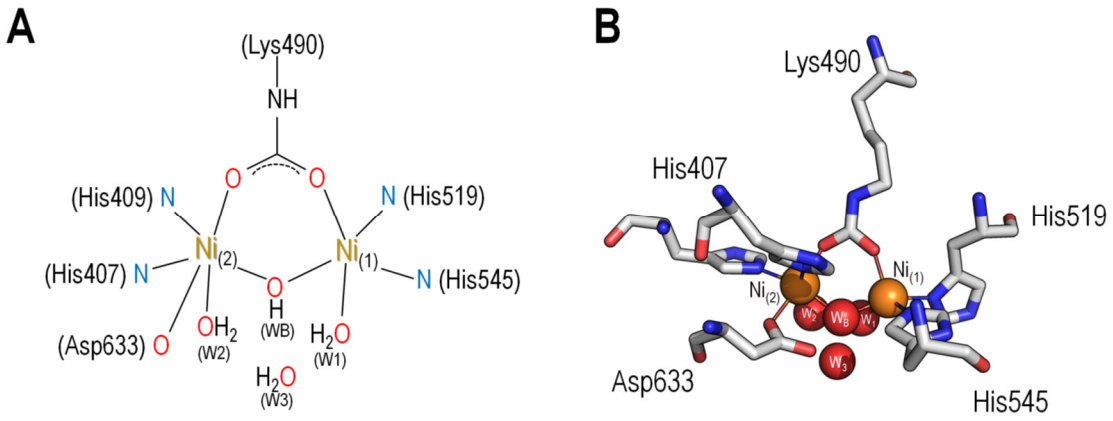


Figura 4. Sítio ativo de ureases: (A) Representação esquemática (adaptado de BENINI ET AL., 1999); (B) Representação espacial (PDB ID 2UBP). A numeração dos resíduos corresponde à urease de *C. ensiformis*.

Uma série de etapas é necessária para que a urease passe de sua forma inativa (apourease) para sua forma ativa (holourease), contendo níquel. A inserção desse metal no sítio ativo requer a interação da enzima com proteínas acessórias. O processo melhor descrito deriva de estudos em *K. aerogenes*, onde as proteínas UreD, UreF e UreG atuam como chaperonas da urease inativa (UreABC), permitindo mudanças conformacionais, carbamilação da lisina e hidrólise de GTP, enquanto a proteína UreE atua como metalochaperona, transportando Ni^{II} (ZAMBELLI ET AL., 2011). A proposta mais difundida é de que essas proteínas se liguem sequencialmente à apourease, mas a hipótese alternativa, de que um oligômero formado por elas atue diretamente sobre a enzima, não está descartada (Figura 5) (CARTER ET AL., 2009).

Devido a obstáculos em sua purificação, pouco se sabe a respeito da UreD, a primeira proteína a ligar na apourease. A UreF, que liga no complexo apourease-UreD, parece atuar como ativadora de GTPase, sendo correlacionada à atividade de GTPase da UreG, quando essa liga na apourease-UreDF. É proposto que a UreG catalise, na presença de CO_2, a formação de carboxifosfato, um agente reconhecido para carbamilação da lisina do sítio ativo. Enquanto o GTP é hidrolisado, a ligadora de níquel UreE entrega os íons

metálicos ao complexo apourease-UreDFG. (CARTER ET AL., 2009; ZAMBELLI ET AL., 2011).

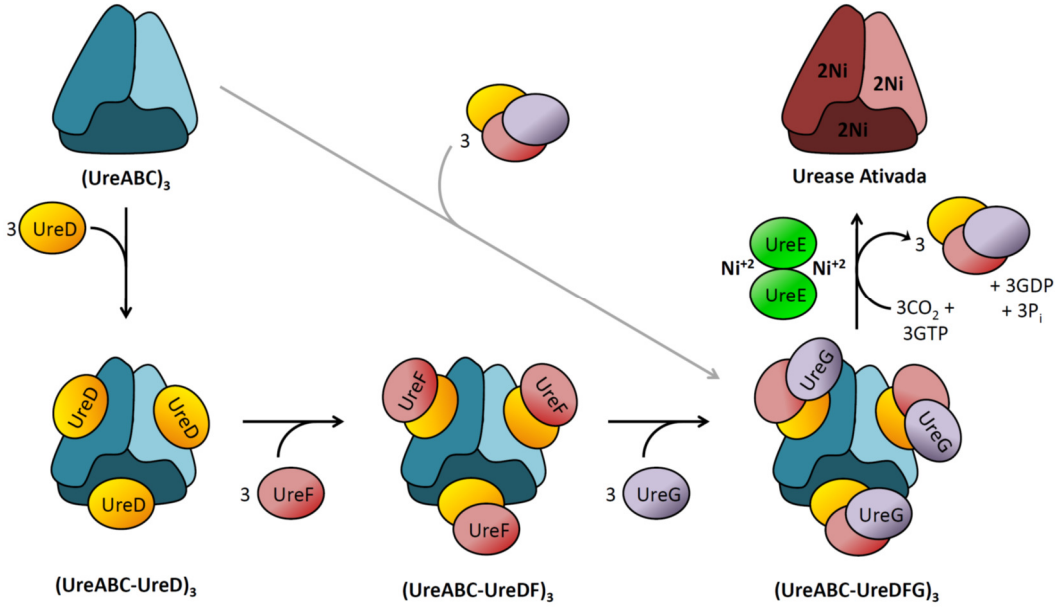


Figura 5. Via de ativação da urease de *K. aerogenes*. A rota alternativa é representada pela seta em cor cinza (baseado em CARTER ET AL., 2009).

Em plantas, o sistema de ativação é equivalente, apesar da ausência da metalochaperona UreE (WITTE, 2011). Nesses organismos, a atividade ligadora de níquel parece ter sido incorporada na UreG, que apresenta uma cauda amino-terminal rica em resíduos de histidina e aspartato, cuja capacidade de ligar metais já foi estudada (POLACCO ET AL., 2013). A mesma combinação de funções ocorre em fungos (SINGH ET AL., 2013).

1.4 PROPRIEDADES CATALÍTICAS E CATÁLISE-DEPENDENTES EM UREASES

A urease é considerada uma das enzimas mais proficientes conhecidas até o momento (KRAJEWSKA, 2009A). Devido principalmente ao fato de a hidrólise não catalisada da ureia nunca ter sido observada, é difícil estabelecer um valor inequívoco para a proficiência da enzima. Baseando-se em comparações com a reação não catalisada de eliminação, a proficiência da enzima foi estimada como sendo 10^{14} vezes superior a esta, enquanto estudos teóricos propuseram um valor de até 10^{32} vezes (ESTIU & MERZ, 2004).

Os mecanismos de hidrólise da ureia catalisados pela urease atualmente considerados são aqueles propostos por BENINI ET AL. (1999) (Figura 6) e por KARPLUS ET AL. (1997) (Figura 7). Baseados em trabalhos anteriores, eles assumem que, no sítio ativo da enzima, a ureia liga-se ao íon Ni(1), mais eletrofílico, e ao átomo de oxigênio do grupo carbonila, tornando o carbono da ureia mais eletrofílico e, portanto, mais susceptível a ataque nucleofílico. Após tomar o lugar das moléculas de água W1-W3, a ureia liga-se ao Ni(2), por meio do átomo de nitrogênio de um dos seus grupos amino, tornando sua ligação bidentada com a urease. Acredita-se que essa ligação facilite o ataque nucleofílico da água sobre o átomo carbono do grupo carbonila, resultando na formação de um intermediário tetraédrico, do qual NH_3 e carbamato são liberados.

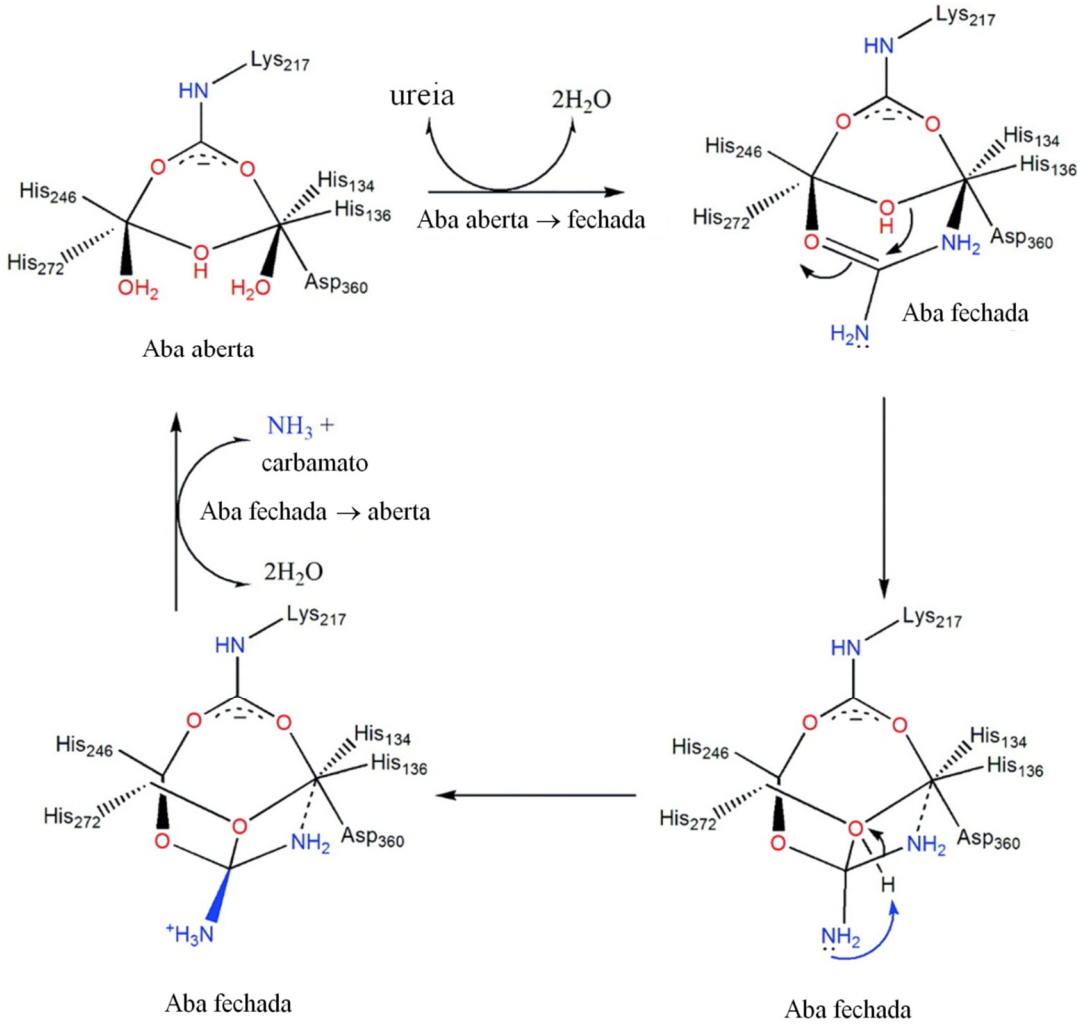


Figura 6. Mecanismo de ureólise proposto por BENINI ET AL., 1999. A reação envolve um intermediário tetraédrico que gera carbamato após a liberação de NH_3. O hidróxido-ponte age como nucleófilo para a carbonila, além de protonar o NH_2 livre. A numeração corresponde aos resíduos na urease de *K. aerogenes* (adaptado de ESTIU & MERZ, 2007).

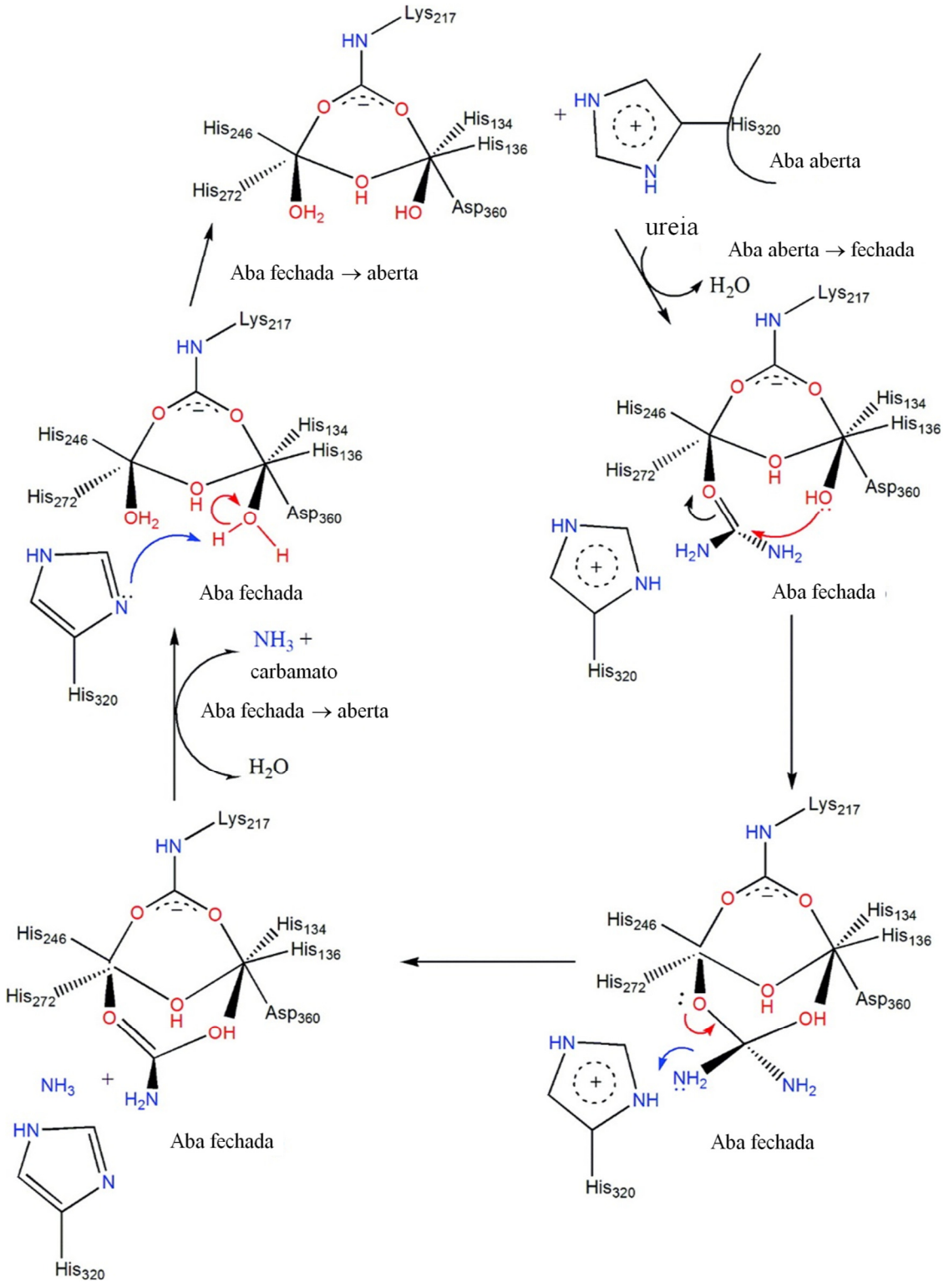


Figura 7. Mecanismo de ureólise proposto por KARPLUS ET AL., 1997. A reação envolve um intermediário tetraédrico que gera ácido carbâmico após a liberação de NH_3. O mecanismo se baseia nas suposições de que uma molécula de água fica retida após a coordenação da ureia e que a His320 está protonada. A numeração corresponde aos resíduos na urease de *K. aerogenes* (adaptado de ESTIU & MERZ, 2007).

Enquanto BENINI ET AL. (1999) propõem que o ataque nucleofílico é executado pelo hidróxido-ponte, que atua também como ácido geral, fornecendo prótons para o grupo NH_3, KARPLUS ET AL. (1997) defendem que é uma histidina, localizada na aba do sítio ativo, que age como ácido geral nessa protonação. Estes autores consideram como possibilidade alternativa a ligação monodentada da urease ao Ni(1), com o Ni(2) fornecendo a molécula de água como nucleófilo para o átomo de carbono do grupamento carbonila da ureia.

Além disso, ESTIU & MERZ (2007), baseado em modelos simplificados do sítio ativo da enzima, propõem que os mecanismos de hidrólise e eliminação poderiam ocorrer de maneira competitiva na urease, e que um mecanismo batizado por eles de “eliminação assistida por proteína” seria favorecido. Considera-se que os mecanismos catalíticos propostos para a urease apresentam várias controvérsias que permanecem não esclarecidas (KRAJEWSKA, 2009A), embora novas evidências, especialmente oriundas de estudos envolvendo inibidores (BENINI ET AL., 2013) tenham apontado para o mecanismo proposto por BENINI ET AL. (1999).

Independente da maneira como ocorre, a hidrólise da ureia possui papel importante em diversos organismos. Em plantas, a urease se junta à arginase e à glutamina sintetase no grupo das enzimas-chave no metabolismo de ureia. Nesses organismos, quantidades significativas de nitrogênio fluem na forma de ureia, derivada da arginina e, possivelmente, da degradação de purinas e ureídos (POLACCO & HOLLAND, 1994). O nitrogênio presente na ureia não está disponível para a planta até que seja hidrolisado pela urease. Após a hidrólise, a amônia gerada é incorporada em compostos orgânicos por meio da glutamina sintetase, principalmente (SIRKO & BRODZIK, 2000).

Ureases também possuem papel importante na germinação e no metabolismo de nitrogênio das plântulas, podendo atuar conjuntamente à arginase na utilização das reservas proteicas da semente enquanto germinam (POLACCO & HOLLAND, 1993). Enquanto a urease ubíqua (encontrada em virtualmente todos os tecidos vegetais) é responsável pela reciclagem da ureia oriunda do metabolismo da planta, a urease embrião-específica, originalmente descrita para soja, tem papel fisiológico desconhecido. É proposto que ela esteja envolvida na proteção contra predadores devido à toxicidade da amônia produzida (POLACCO & HOLLAND, 1994, FOLLMER, 2008), embora mecanismos

adicionais de toxicidade, independentes da atividade catalítica da enzima, venham sendo descritos (ver adiante).

Entre as bactérias ureolíticas, as patogênicas são as mais frequentemente estudadas, uma vez que a patogênese é descrita como relacionada à hidrólise da ureia, ocasionando aumento do pH e toxicidade pela amônia e seus derivados (BURNE & CHEN, 2000). A ureia é o principal produto de excreção de nitrogênio da maioria dos animais terrestres, sendo amplamente disponível para bactérias ureolíticas. Em humanos, os sistemas urinário e digestório são os locais mais comuns de infecção por essas bactérias (BURNE & CHEN, 2000; FOLLMER, 2010; KONIECZNA ET AL., 2012).

O aumento do pH da urina pode causar várias complicações em humanos, incluindo a precipitação de íons solúveis presentes na urina, causando a formação de cálculos renais e, em casos extremos, a necrose do tecido renal. Tais cálculos são compostos principalmente por estruvita e apatita e seu principal agente causador é a bactéria *Proteus mirabilis* (MOBLEY ET AL., 1995, BURNE & CHEN, 2000).

No trato digestório humano, a bactéria ureolítica *H. pylori* é o principal causador de infecção (BURNE & CHEN, 2000). Ela normalmente coloniza a mucosa estomacal, aumentando localmente o pH altamente acídico, o que permite que a bactéria sobreviva nesse ambiente hostil (HA ET AL., 2001). Simultaneamente, ocorre dano ao tecido do hospedeiro, causando gastrite e úlceras gastrointestinais. Este dano está principalmente relacionado à amônia, derivada da ureólise, e à monocloramina, derivada da reação oxidativa das células imunitárias (FOLLMER, 2010; KONIECZNA ET AL., 2012).

A atividade ureolítica que ocorre no solo é de grande importância para a agricultura. Apesar de microrganismos serem relevantes para esse processo, ele deriva principalmente das chamadas “ureases de solo” (BREMNER & KROGMEIER, 1989). Essas ureases são resíduos de plantas e células bacterianas mortas, e sua estabilidade no meio extracelular se deve à imobilização em argilas e substâncias húmicas do solo (GIANFREDA ET AL., 1995). A alta estabilidade destas ureases é o que permite que a ureia seja utilizada como fertilizante de maneira eficiente (KRAJEWSKA, 2009A).

Adicionalmente, as ureases podem promover a formação de carbonato de cálcio em sedimentos geológicos, solos, e fontes de água. Apesar de não estar

totalmente resolvido, o papel das ureases bacterianas nesse processo parece ser triplo, aumentando a alcalinidade de forma favorável à precipitação de $CaCO_3$, aumentando a concentração de carbono inorgânico dissolvido no meio e, por fim, servindo de sítio de nucleação para os cristais (CASTANIER ET AL., 1999, MÉTAYER-LEVREL ET AL., 1999).

1.5 PROPRIEDADES CATÁLISE-INDEPENDENTES EM UREASES

A capacidade de catalisar a hidrólise de ureia a amônia e ácido carbâmico é considerada a característica fundamental das ureases. Tem-se demonstrado, entretanto, que essas enzimas são proteínas multifuncionais, apresentando várias propriedades independentes da catálise.

Descoberta em 1981, a canatoxina é uma proteína altamente tóxica de *C. ensiformis* que induz convulsão e morte em ratos e camundongos quando injetada intraperitonealmente (CARLINI & GUIMARÃES, 1981). Posteriormente, a canatoxina foi identificada como uma isoforma menos abundante de urease (FOLLMER ET AL., 2001) e uma família de genes relacionados a ureases foi descoberta nessa planta (PIRES-ALVES ET AL., 2003).

Além de neurotóxica em mamíferos, essa isoforma de urease apresenta atividade inseticida contra besouros (Coleoptera) e percevejos (Hemiptera), propriedade compartilhada com a isoforma majoritária da enzima de *C. ensiformis* e com a urease ubíqua de soja (CARLINI ET AL., 1997, CARLINI & GROSSI-DE-SÁ, 2002, FOLLMER ET AL., 2004, STANISÇUASKI ET AL., 2005). Esses grupos de insetos possuem catepsinas como principais enzimas digestivas, contrastando com grupos que possuem tripsinas realizando tal função, que são imunes à toxicidade da canatoxina. Essa diferença evidencia a importância da chamada “ativação proteolítica” das ureases. Nesse processo, originalmente descrito para canatoxina, um peptídeo entomotóxico de 10 kDa é liberado pela digestão da enzima por catepsinas, sendo o principal responsável pela atividade inseticida da canatoxina (ver item 1.5.1) (CARLINI ET AL., 1997; FERREIRA-DASILVA ET AL., 2000). Adicionalmente, os efeitos inseticidas das ureases de soja e feijão-de-porco persistem após tratamento com inibidores irreversíveis de ureases, demonstrando que um domínio distinto do sítio ativo está envolvido nas propriedades inseticidas (FOLLMER ET AL., 2004, CARLINI & POLACCO, 2008).

O mecanismo de ação associado à atividade inseticida de ureases ou de seus peptídeos derivados aguarda elucidação. Entre os efeitos entomotóxicos, a diminuição da diurese em *Rhodnius prolixus* (Hemiptera) foi investigada, revelando que o peptídeo e a urease intacta modulam diferentes cascatas de sinalização em túbulos de Malpighi intactos: enquanto a urease ativa a via dos eicosanoides e transporte de cálcio, o peptídeo altera os níveis de cGMP e o potencial trans-epitelial (CARLINI & POLACCO, 2008, STANISÇUASKI ET AL., 2009).

Além de propriedades inseticidas, as ureases possuem outros efeitos que são independentes da catálise. As enzimas de soja, feijão-de-porco e, em menor grau, de *H. pylori*, inibem o crescimento vegetativo de diversos fungos filamentosos e leveduras, possivelmente induzindo plasmólise e danos à parede celular (BECKER-RITT ET AL., 2007; POSTAL ET AL., 2012). Além do "domínio inseticida", existem indicativos de que outras regiões da urease podem estar envolvidas no efeito sobre fungos, incluindo uma possível atividade de celulase (BALASUBRAMANIAN ET AL., 2013A).

Outras propriedades testadas para essas ureases incluem a ligação a glicoconjugados (FOLLMER ET AL., 2001) e a ativação de plaquetas (FOLLMER ET AL, 2004), que ocorre por meio da via das lipoxigenases (CARLINI ET AL. 1985; OLIVERA-SEVERO ET AL., 2006; WASSERMANN ET AL., 2010). Efeitos pró-inflamatórios também foram observados em estudos com a urease de *H. pylori*, envolvendo ativação de neutrófilos e inibição de sua apoptose (UBERTI ET AL., 2013). Algumas dessas propriedades, como a ativação de plaquetas, foram descritas recentemente para a urease de *P. mirabilis* (BROLL, 2013).

Uma propriedade proposta, mas ainda não testada, é a de que a urease extravasada de *B. pasteurii*, uma das "ureases de solo" presentes na rizosfera, possa atuar sobre células de raízes, induzindo secreção de compostos que podem ser consumidos, em última instância, pela bactéria (CARLINI & POLACCO, 2008). Recentemente foi demonstrado que a urease de *Bradyrhizobium japonicum*, uma bactéria fixadora de nitrogênio em plantas de soja, não está envolvida nos processos de nodulação ou fixação. O mesmo trabalho demonstrou, no entanto, que ureases de soja estão implicadas na eficiência da fixação biológica de nitrogênio e possuem papel quimiotático sobre a referida bactéria (MEDEIROS-SILVA, 2012).

Embora a maioria dos casos estudados retrate propriedades “belicosas” dessas enzimas, as ureases podem ter papel importante na interação positiva entre diferentes organismos. É o que ocorre em algumas associações liquênicas. Ureases com glicosilações específicas, aderidas à parede celular da alga, são ligantes específicos para lectinas secretadas pelo fungo, tendo papel fundamental no reconhecimento da compatibilidade interespecífica do líquen (SACRISTÁN ET AL., 2006).

Por apresentarem pelo menos duas funções não relacionadas, as ureases podem ser incluídas no grupo das chamadas *moonlighting proteins.* Originalmente chamado de “compartilhamento gênico” (PIATIGORSKY & WISTOW, 1989; PIATIGORSKY, 2007), esse fenômeno descreve cadeias polipeptídicas únicas com funções múltiplas, que não são efeito de fusão gênica, variação de *splicing* ou atividade enzimática promíscua (JEFFERY, 1999). Esse grupo de proteínas tem sido mais extensivamente estudado em leveduras e mamíferos (JEFFERY, 2009; WANG ET AL., 2013), tornando as ureases um exemplo interessante de multifuncionalidade em proteínas vegetais e microbianas.

1.5.1 Peptídeos derivados de ureases

Após as primeiras evidências de que a canatoxina poderia ser ativada proteoliticamente, vários estudos foram desenvolvidos para investigar essa hipótese (STANISÇUASKI & CARLINI, 2012). A digestão *in vitro* da proteína com enzimas oriundas das larvas de *Callosobruchus maculatus* (Coleoptera) originou peptídeos que foram fracionados e testados quanto a toxicidade em ninfas e adultos de *R. prolixus* (FERREIRA-DASILVA ET AL., 2000). Das diferentes frações produzidas, a que continha peptídeos de cerca de 10kDa mostrou-se tóxica para adultos por injeção, contrastando com a ausência de efeitos observada para a proteína intacta quando empregada a mesma rota. Frações com peptídeos menores foram ativas também contra ninfas, indicando que a atividade inseticida possa ser compartilhada por uma família de peptídeos, ou que o peptídeo principal possa sofrer etapas posteriores de digestão (STANISÇUASKI & CARLINI, 2012).

O principal peptídeo das frações de 10kDa foi sequenciado e denominado Pepcanatox (CARLINI ET AL., 2000). Estudos posteriores quanto a sua atividade e mecanismos de ação foram desenvolvidos com um peptídeo recombinante,

chamado Jaburetox-2Ec (MULINARI ET AL., 2007). Esse peptídeo mostrou-se tóxico a vários insetos (*Dysdercus peruvianus, R. prolixus, T. infestans*), incluindo espécies que não são afetadas pelas ureases nativas (como *Spodoptera frugiperda*) (conforme sumarizado por STANISÇUASKI & CARLINI, 2012).

Em relação aos peptídeos derivados de ureases, é importante destacar que o termo "peptídeo" é usado para definir um fragmento de proteína (independentemente de sua massa molecular), uma vez que o próprio Jaburetox-2Ec possui 93 aminoácidos, um tamanho comparável a pequenas proteínas (YANG ET AL., 2011). Antes da determinação da estrutura de uma urease de cadeia única por cristalografia de raios X (BALASUBRAMANIAN & PONNURAJ, 2010), uma estudo de modelagem *ab initio* do peptídeo Jaburetox-2Ec sugeriu a presença de motivos estruturais similares aos encontrados em peptídeos formadores de poros (Figura 8) (MULINARI ET AL., 2007; BARROS ET AL., 2009). Por simulação de dinâmica molecular foi demonstrado que o peptídeo se ancorava na interface polar:não-polar, enquanto experimentos com vesículas unilamelares demonstraram sua capacidade de desorganizar bicamadas lipídicas com características acídicas (BARROS ET AL., 2009).

Com a determinação da estrutura da urease de *C. ensiformis* (BALASUBRAMANIAN & PONNURAJ, 2010) e de *C. cajan* (BALASUBRAMANIAN ET AL., 2013A), foi confirmada a presença de um motivo β-*hairpin* na região correspondente ao peptídeo entomotóxico, conforme proposto pela modelagem *ab initio* (Figura 8). Adicionalmente, BALASUBRAMANIAN ET AL. (2013B), propuseram que a região do β-*hairpin* poderia se oligomerizar, formando um poro, o que poderia estar associado à sua atividade desorganizadora de membranas.

Em estudos visando identificar domínio(s) ou região(ões) relacionado(s) à atividade fungicida da urease de *C. ensiformis*, POSTAL ET AL. (2012) empregaram digestão enzimática para obter fragmentos da proteína e testa-los frente a diferentes fungos. Cinco dos peptídeos gerados por essa metodologia (menores que 10 kDa) foram identificados por espectrometria de massas e não apresentaram correspondência com nenhuma proteína antifúngica descrita. É ainda incerto se tais peptídeos estão realmente envolvidos na atividade antifúngica de ureases. Um dos peptídeos mostrou-se como parte da região N-

terminal do Pepcanatox, permitindo testes com sua forma recombinante, Jaburetox, frente a diferentes fungos. O peptídeo foi capaz de inibir o desenvolvimento micelial, mas em doses mais altas que as da urease completa, indicando que outras regiões podem estar envolvidas em sua atividade antifúngica (POSTAL ET AL., 2012).

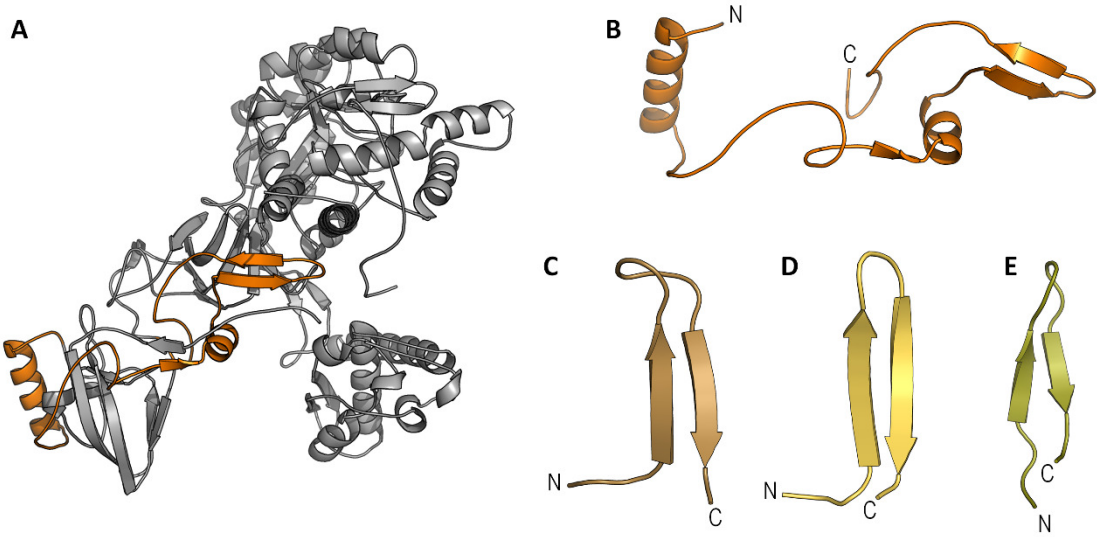


Figura 8. (A) Localização do peptídeo Pepcanatox na urease intacta; (B) peptídeo Pepcanatox isolado; peptídeos tóxicos com motivo β-*hairpin*: (C) polifemusina (PDB ID 1X7K), (D) protegrina (PDB ID 1PG1), (E) taquiplesina (PDB ID 1WO0). Os amino- e carboxiterminais dos peptídeos estão indicados. Figuras A e B baseados na estrutura da urease de *C. ensiformis* (PDB ID 3LA4).

1.6 APLICAÇÕES BIOTECNOLÓGICAS DE UREASES

As diferentes propriedades apresentadas pelas ureases permitem um amplo conjunto de aplicações biotecnológicas, tanto já testadas quanto hipotéticas. Diferentes métodos analíticos aplicados a ureases imobilizadas permitem medir a concentração de ureia em sangue, urina, água potável, esgoto e bebidas alcoólicas; além de permitir a avaliação de metais pesados em solos e reservatórios de água (QIN & CABRAL, 2002). Sistemas empregando urease como parte do método de detecção já foram propostos para análise dos níveis de arginina, creatinina e IgG no sangue (ALONSO ET AL., 1995; CULLEN ET AL., 1990; SANSUBRINO & MASCINI, 1994).

Também foi proposto o emprego de ureases na remoção da ureia em casos de falência renal (acoplado a um sistema de remoção da amônia)

(PRAKASH & CHANG, 1995); na produção de bebidas alcoólicas (KODAMA & YOTSUZUKA, 1996); e na recuperação de água contaminada, inclusive em sistemas de suporte a vida no espaço (KRAJEWSKA, 2009B). Ureases também já foram propostas como agentes de controle de pH em sistemas multi-enzimáticos industriais (QIN & CABRAL, 2002).

A precipitação de carbonatos induzida pela enzima tem sido empregada como meio de corrigir fissuras e outros pequenos danos em objetos artísticos e monumentos históricos (KRAJEWSKA, 2009B; RAUT ET AL., 2013). Além disso, sua aplicação na fixação de dunas e criação de ambientes projetados em desertos tem sido perseguida por arquitetos e bioquímicos (LARSSON, 2013).

A identificação de várias de suas propriedades *moonlighting* permitiu propor aplicações ainda mais abrangentes para ureases. Tais aplicações incluem o controle de fungos patogênicos para plantas e animais (POSTAL ET AL., 2012; BALASUBRAMANIAN ET AL., 2013A) e o desenvolvimento de plantas de importância agrícola resistentes a diferentes tipos de pragas (CARLINI & GROSSI-DE-SÁ, 2002; MULINARI ET AL., 2007; STANISÇUASKI & CARLINI, 2012).

1.7 MODELAGEM MOLECULAR COMPARATIVA

A modelagem molecular comparativa (ou “modelagem por homologia”) é uma técnica que permite a construção de modelos tridimensionais para uma proteína de estrutura desconhecida, baseando-se em estruturas de proteínas similares a ela em suas sequências de aminoácidos (MARTÍ-RENOM ET AL., 2000). Tais estruturas, determinadas por cristalografia de raios X ou ressonância magnética nuclear, são empregadas como moldes para obtenção da estrutura da proteína de interesse, chamada de proteína-alvo.

A modelagem comparativa consiste de quatro passos sequenciais principais: seleção do molde, alinhamento entre as sequências do molde e da proteína-alvo, construção do modelo e avaliação/validação do modelo. Se o modelo não for satisfatório, os três primeiros passos podem ser repetidos até que um modelo satisfatório seja obtido (SÁNCHEZ & ŠALI, 2000).

Uma vez alinhados molde e proteína-alvo, uma variedade de métodos pode ser utilizada para construir o modelo tridimensional de interesse. Entre eles, destaca-se a modelagem por satisfação de restrições espaciais, que utiliza distâncias interatômicas ou técnicas de otimização para satisfazer restrições

espaciais extraídas do alinhamento. Tais restrições são geralmente complementadas por restrições estereoquímicas de comprimentos de ligação, ângulos de ligação, ângulos diedrais e contatos entre átomos não ligados, obtidas do campo de força de mecânica molecular. O modelo é então derivado ao minimizarem-se as violações de todas as restrições obtidas (SCHWEDE ET AL., 2008).

Após sua construção, o modelo é avaliado e validado. As avaliações destes modelos podem ser de dois tipos: internas, quando avaliam sua autoconsistência, se ele satisfez ou não as restrições usadas para calculá-lo; ou externas, quando baseadas em informações que não foram utilizadas para calcular o modelo. Entre as variedades de avaliação externa tem-se o cálculo de perfis de energia do modelo e a avaliação estereoquímica (FORSTER, 2001). A validação do modelo envolve a comparação da estrutura obtida com dados experimentais pertinentes, como a conservação de sítios ativos ou de superfícies de contato com outras proteínas, por exemplo (SCHWEDE ET AL., 2008). A metodologia de modelagem comparativa é apresentada com mais detalhes no Apêndice C.

1.8 ATRACAMENTO PROTEÍNA-PROTEÍNA

O atracamento (ou *docking*) de proteínas é uma técnica que visa calcular a estrutura tridimensional de um complexo proteico, partindo das estruturas individuais das proteínas constituintes (RITCHIE, 2008). Ela requer a amostragem de uma grande quantidade de uma proteína em relação à outra, buscando identificar o complexo que apresente a menor energia livre, considerado como sendo o complexo proteico mais próximo do complexo nativo (GRAY, 2008).

Três elementos são essenciais para o atracamento proteína-proteína: uma representação adequada das estruturas proteicas e dos graus de liberdade que serão buscados; um algoritmo que explore o espaço conformacional o mais completamente possível; e uma função de pontuação (*scoring*) que classifique os complexos obtidos (SALADIN & PREVOST, 2008). Diferentes propostas e combinações são possíveis entre esses elementos, sendo implementadas em diferentes programas de atracamento proteico.

Apesar de avanços significativos nesse campo de pesquisa, ainda é frequente a obtenção de grandes quantidades de falsos-positivos, prejudicando

a identificação do complexo nativo (COMEAU ET AL., 2004). Uma das maneiras de contornar tal obstáculo envolve o emprego de dados bioquímicos e biofísicos para filtrar os resultados obtidos ou para guiar o processo de atracamento. Esses dados podem ser oriundos de experimentos de *cross-linking*, mutação sítio dirigida e co-purificação, por exemplo, e são usados para limitar as regiões de interação a serem analisadas pelo algoritmo de busca (MELQUIOND & BONVIN, 2008).

1.9 DINÂMICA MOLECULAR

As simulações de dinâmica molecular (DM), procedimentos baseados na computação do movimento dos átomos em uma molécula, remontam à década de 1950 (MAGINN & ELLIOT, 2010). Vêm sendo empregadas desde a década de 1970 no estudo de biomoléculas, quando se analisou um sistema protéico envolvendo um inibidor pancreático de tripsina bovina (MCCAMMON ET AL., 1977). Atualmente, tais simulações tornaram-se ferramentas amplamente difundidas. Seu emprego inclui a investigação da estrutura e dinâmica de biomoléculas em geral, abrangendo, por exemplo, desnaturação e enovelamento protéicos (PONDER & CASE, 2003), translocação de tRNAs no ribossomo (SANBONMATSU ET AL., 2005) e dissolução de capsídeos virais (LARSSON ET AL., 2012).

A integração da equação de movimento de Newton (abaixo) é a caractrística fundamental dos cálculos de DM que, quando realizada sucessivamente e sobre todos os átomos do sistema, gera uma seqüência de diferentes posições dos átomos em função do tempo, ou seja, uma trajetória de movimento das moléculas em estudo (LEACH, 2001).

$$\text{aceleração do átomo } i \leftarrow \frac{d^2 r_i(t)}{dt^2} = \frac{F_i}{m_i} \begin{array}{l} \rightarrow \text{força exercida sobre o átomo } i \\ \rightarrow \text{massa do átomo } i \end{array}$$

A integração é realizada de forma que uma força F_i causa uma aceleração sobre um determinado átomo i e, em consequência, acarreta mudança de sua posição num intervalo de tempo Δt relativo à aceleração. As equações de movimento, portanto, são integradas de maneira determinar a

trajetória de cada átomo, ou seja, sua posição e velocidade entre o tempo t e o tempo $t+\mathrm{d}t$. As trajetórias de todos os átomos, expressas como um número pré-definido de passos (*time-steps*), são então armazenadas para análises posteriores. Passos de cálculo de 1 fs são normalmente empregados em simulações de dinâmica de proteínas, de maneira que 1 ns de simulação requer 10^6 cálculos de força e energia para cada átomo (LEACH, 2001; SCHLICK, 2006; SERDYUK ET AL., 2007). A referida equação, entretanto, não é capaz de determinar a magnitude e a direção da força Fi sobre os átomos do sistema, nem sua relação com as características químicas de cada molécula em estudo (LEACH, 2001). Tais parâmetros são calculados em função de mudanças na energia potencial/cinética entre a posição atual e a posição seguinte (a que representará o próximo passo da simulação) sobre cada átomo separadamente. Esta superfície de energia potencial representa a energia de cada molécula, sendo descrita pelo denominado Campo de Força (SCHLICK, 2006).

O campo de força pode ser definido como um conjunto de funções e parametrizações usadas em cálculos de mecânica molecular (DE SANT'ANNA, 2002). Estas funções definem as energias de estiramento de ligação e de distorção de ângulo de ligação (tanto de valência quanto de diedro) de uma molécula quando comparadas com a sua conformação não tensionada (aquela caracterizada pelos valores-padrão de comprimentos e de ângulos de ligação) (Figura 9). Nesse contexto, os campos de força, juntamente com possíveis termos adicionais (e.g. de interação entre átomos não ligados, de efeitos eletrostáticos, de ligação de hidrogênio e de outros efeitos estruturais), expressam o somatório das funções de energia potencial de cada átomo e calculam a energia dos sistemas em função das posições dos núcleos dos átomos, que são, por sua vez, representados por esferas unidas por molas (LEACH, 2001; SCHLICK, 2006).

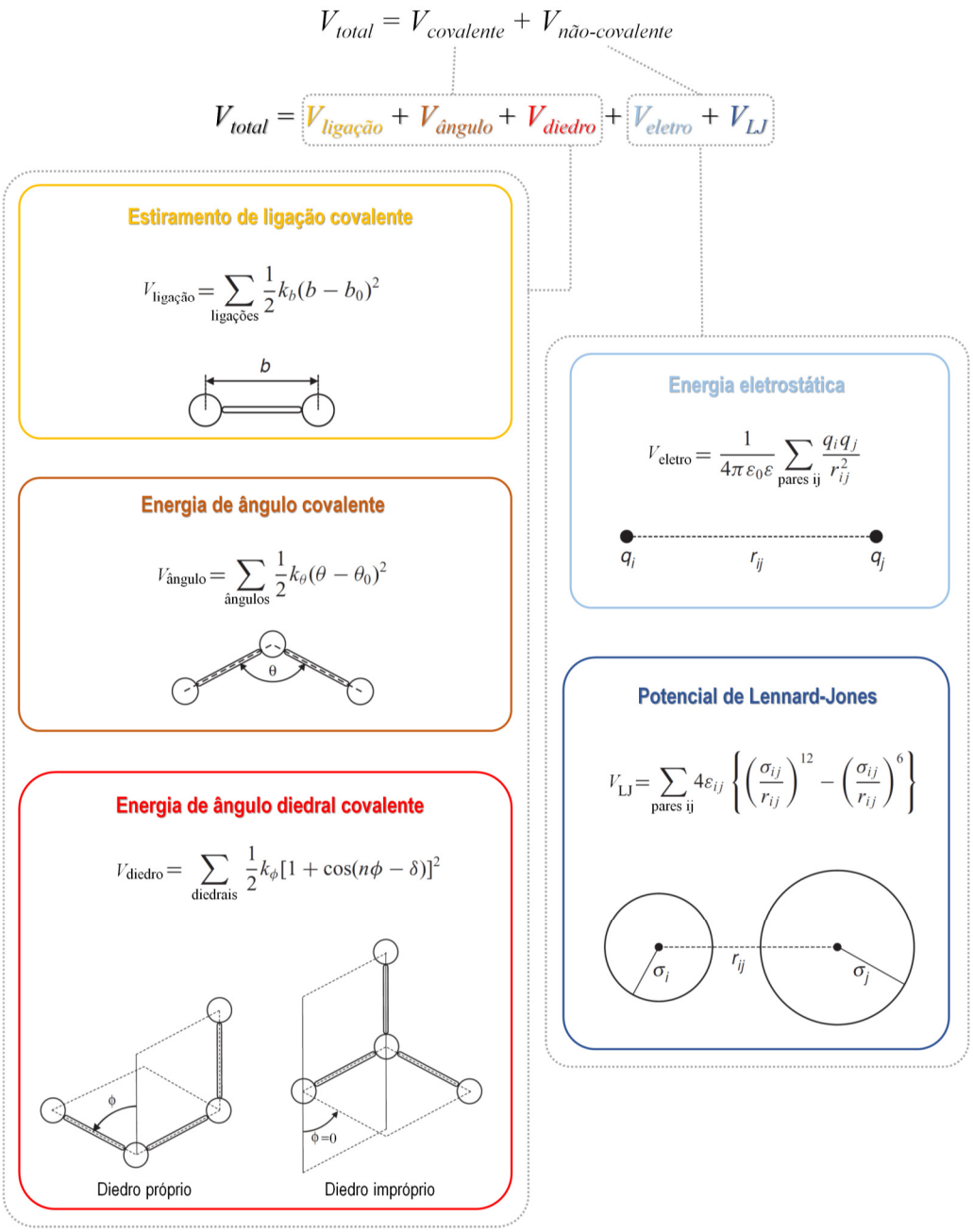


Figura 9. Funções de energia que compõem campos de força empregados em dinâmica molecular (adaptado de SERDYUK ET AL., 2007).

1.10 FILOGENÉTICA MOLECULAR

A filogenética molecular é o estudo de relações evolutivas com o emprego de dados moleculares, como sequências de proteínas e ácidos nucleicos (GRAUR & LI, 2000). Essas análises comparativas normalmente são representadas na forma de uma filogenia, ou árvore filogenética, que descreve as relações evolutivas entre as sequências (BALDAUF, 2003).

Uma filogenia é uma árvore formada por nós conectados por ramos. Cada ramo representa a persistência de uma linhagem ao longo do tempo, enquanto cada nó representa o surgimento de uma nova linhagem. Árvores filogenéticas não podem ser observadas diretamente, sendo inferidas a partir de sequências moleculares, por exemplo (GREGORY, 2008).

Para a reconstrução filogenética, parte-se normalmente de um grupo de sequências homólogas, com a possível adição de sequências que permitam enraizar a árvore. Essas sequências são agrupadas, por meio de um alinhamento múltiplo, que constitui a base para o processamento pelos diferentes algoritmos empregados em estudos de filogenia. Adicionalmente, um modelo de evolução de sequências deve ser escolhido. Os dados e o modelo são então usados para gerar uma grande quantidade de árvores que podem ser resumidas em uma árvore-consenso (representando ramificações que ocorrem na maioria das árvores amostradas) (HOLDER & LEWIS, 2003).

Os métodos de reconstrução filogenética podem ser baseados em distâncias ou caracteres. Nos métodos baseados em distância, a distância entre pares de sequências é calculada e a matriz resultante é empregada na reconstrução da árvore. Exemplos desse tipo de método incluem *neighbour joining* e UPGMA. Os métodos baseados em caracteres comparam simultaneamente todos as sequências do alinhamento, considerando um caractere (uma posição no alinhamento) de cada vez, para calcular a pontuação de cada árvore. Exemplos desses métodos incluem máxima parcimônia, máxima verossimilhança e inferência Bayesiana. (YANG & RANNALA, 2012)

Em teoria, a árvore com melhor pontuação é identificada por comparação de todas as árvores possíveis. Na prática, devido à enorme quantidade de árvores possíveis, uma busca exaustiva não é factível computacionalmente (exceto para conjuntos de dados muito pequenos). Por esse motivo, algoritmos heurísticos são empregados (HOLDER & LEWIS, 2003; YANG & RANNALA, 2012). Aspectos teóricos e detalhamento metodológico referentes à filogenética molecular são apresentados no Apêndice D.

2. OBJETIVOS

Considerando as diversas atividades apresentadas por ureases e suas possíveis aplicações biotecnológicas, o presente trabalho visa ampliar o conhecimento acerca dessas enzimas, aplicando técnicas computacionais ao estudo de diferentes aspectos de sua biologia estrutural. Três desses aspectos se destacam: a história evolutiva dessas enzimas, o processo de ativação das mesmas, e o comportamento conformacional dos peptídeos biologicamente ativos delas derivados.

Para o presente trabalho, portanto, as seguintes metas foram estabelecidas:

► analisar a história evolutiva de ureases, buscando compreender as relações de ancestralidade e derivação estabelecidas entre as enzimas de diferentes organizações estruturais;

► oferecer um arcabouço estrutural para as etapas de ativação de ureases, empregando os dados disponíveis na construção de modelos envolvendo a interação da urease com suas proteínas acessórias;

► compreender as diferenças de comportamento conformacional entre os peptídeos derivados da urease de *C. ensiformis*, sendo capaz de propor otimizações a sua aplicação biotecnológica.

Espera-se, a partir dos resultados alcançados aumentar a compreensão das relações estrutura-função em ureases, contribuindo em sua aplicação biotecnológica além de aprofundar o conhecimento de aspectos evolutivos a respeito dessas enzimas.

3. MÉTODOS

"Computers are useless. They can only give you answers."

Pablo Picasso

3.1 ANÁLISE FILOGENÉTICA

3.1.1 Obtenção e alinhamento de sequências

A prospecção por sequências de aminoácidos de ureases foi realizada por meio de busca por palavra-chave ("urease") no National Center for Biotechnology Information (SAYERS ET AL., 2012). As sequências obtidas foram manualmente inspecionadas. Sequências incompletas (com tamanho inferior às menores ureases confirmadas), erroneamente rotuladas (verificadas por comparação a outras ureases) ou relacionadas às proteínas acessórias de ureases foram removidas.

Sequências de bactérias e arqueas foram alinhadas (com base em gênero de origem), visando encontrar similaridades. Sequências com identidade igual ou superior a 95% foram consideradas idênticas e agrupadas, sendo uma das componentes do grupo eleita como sequência representativa. O algoritmo ClustalW (LARKIN ET AL., 2007) foi utilizado em todos os alinhamentos. Do conjunto original de 14221 sequências, 32 delas provinham de fontes eucarióticas. Sua menor quantidade e ausência de duplicidades dispensou filtragens manuais. Através dos alinhamentos por espécie e gênero, o número de sequências microbianas representativas foi reduzido a 162.

As sequências de ureases de bactérias e arqueas foram alinhadas às ureases de plantas e fungos, e suas subunidades foram unidas manualmente em uma só sequência com base nesses alinhamentos, formando uma só sequência γβα, ou equivalente. O número de sequências microbianas foi reduzido ainda mais, visando diminuir a carga computacional. Baseado na árvore-guia gerada pelo ClustalW, sequências altamente similares (80% de identidade ou superior) foram consideradas idênticas, e somente um representante foi escolhido. O conjunto microbiano final foi montado e alinhado

com as sequências eucarióticas, totalizando 124 sequências de ureases. Regiões com longos gaps foram removidas manualmente dos alinhamentos.

3.1.2 Análise de sequências e construção de árvores filogenéticas

Regiões altamente variáveis nas sequências alinhadas foram identificadas com emprego da ferramenta SimPlot (LOLE ET AL., 1999). Tal ferramenta compara o percentual de identidade entre grupos de sequências dada uma janela de cobertura específica. Para este procedimento, foi utilizada uma janela de 20 aminoácidos, deslocando-se resíduo a resíduo. Com base nesses resultados, o alinhamento inicial originou três subconjuntos distintos: sequências completas de urease (sem *gaps* longos), regiões altamente variáveis e regiões conservadas.

Os programas ProtTest (ABASCAL ET AL., 2005) e MEGA5 (TAMURA ET AL., 2011) foram usados para identificar o modelo de substituição de aminoácidos mais adequado para as sequências alinhadas. Nesse caso, foi sugerido o modelo de Whelan e Goldman (WAG) com distribuição Gamma discreta para contabilizar diferenças na taxa evolutiva entre sítios (+G), considerando alguns deles como evolutivamente invariáveis (+I) (WHELAN & GOLDMAN, 2001).

Árvores filogenéticas foram calculadas por máxima verossimilhança (*maximum likelihood*, ML) utilizando o MEGA5 com 1000 replicatas de *bootstrap*, e por inferência Bayesiana (*Bayesian inference*, BI) utilizando o programa MrBayes (HUELSENBECK & RONQUIST, 2001) por 7,5 x 10^6 gerações, amostradas a cada 10 gerações. Todas as árvores foram enraizadas no ramo contendo o maior número de sequências de ureases pertencentes a arquea. As árvores obtidas foram visualizadas e editadas por meio do programa FigTree (RAMBAUT, 2012).

3.2 FORMAÇÃO DE COMPLEXOS PROTEICOS

3.2.1 Estruturas proteicas

A estrutura da urease nativa de *K. aerogenes* foi obtida do Protein Data Bank, PDB ID 1FWJ (PEARSON ET AL., 1997). A sua organização trimérica (trímeros de monômeros funcionais, 3xUreABC) foi reconstruída a partir da

informação de simetria cristalográfica com o programa PyMol 1.3 (SCHRÖDINGER LLC).

A modelagem molecular comparativa foi empregada para obtenção das estruturas completas das proteínas acessórias UreD, UreF e UreG de *K. aerogenes*. As proteínas UreD (UniProt ID Q09063.1) e UreF (UniProt ID P18318.1) foram modeladas tendo o complexo UreH-UreF de *H. pylori* (PDB ID 3SF5) (FONG ET AL., 2011) como molde. A proteína UreG (UniProt ID P18319) foi modelada com base na HypB de *Methanocaldococcus jannaschii* (PDB ID 2HF9) (GASPER ET AL., 2006), seguindo uma sugestão prévia da literatura (CARTER ET AL., 2009). Todos os alinhamentos de sequência foram realizados com ClustalW (LARKIN ET AL., 2007) e a construção de modelos foi realizada com o programa Modeller 9v10 (SÁNCHEZ & ŠALI, 2000). Para cada caso, vinte modelos foram construídos. Esses modelos foram avaliados estereoquimicamente com Procheck (LASKOWSKI ET AL., 1993) e seus perfis tridimensionais foram comparados a perfis ideais com Verify3D (LÜTHY ET AL., 1992). O modelo de melhor pontuação para cada proteína foi então selecionado.

3.2.2 Atracamento proteína-proteína

O complexo de ativação da urease foi construído de maneira sequencial, conforme proposto por QUIROZ-VALENZUELA ET AL. (2008). A primeira etapa envolveu o acoplamento UreD-$(UreABC)_3$. A estrutura $(UreABC\text{-}UreD)_3$ resultante foi então acoplada com UreF, formando a estrutura $(UreABC\text{-}UreDF)_3$. Esta, por sua vez, foi acoplada a UreG, resultando no complexo $(UreABC\text{-}UreDFG)_3$. Adicionalmente, o dímero UreD-UreF, obtido por sobreposição dos modelos de UreD e UreF à estrutura cristalográfica de FONG ET AL. (2011), foi também acoplado a $(UreABC)_3$ para comparação. As poses de ligação das proteínas acessórias foram replicadas para obter a ligação tripla a $(UreABC)_3$.

Com exceção do último estágio (ligação de UreG), todos os cálculos de atracamento foram realizados sem restrição, ou seja, cada proteína acessória estava livre para varrer toda a superfície do oligômero em busca de seu sítio preferencial de ligação. Os sítios mais prováveis de ligação da UreG à estrutura $(UreABC\text{-}UreDF)_3$ foram obtidos do trabalho de BOER & HAUSINGER (2012) e empregados na restrição das poses obtidas. Cada estágio de atracamento foi realizado em três programas independentes: PatchDock (SCHNEIDMAN-DUHOVNY

ET AL., 2005), Hex (MACINDOE ET AL., 2010), e PIPER (KOZAKOV, ET., 2006) implementado no ClusPro 2.0 (COMEAU ET AL., 2004).

3.2.3 Comparação com dados de SAXS

O espalhamento de raios-X a baixo ângulo (SAXS) é um método estabelecido para a caracterização estrutural de biomoléculas em solução. A técnica fornece estruturas tridimensionais globais, empregando reconstruções *ab initio* e modelagem híbrida, permitindo caracterizar sistemas flexíveis (MERTENS & SVERGUN, 2010; PETOUKHOV & SVERGUN, 2013). O funcionamento da técnica é representado esquematicamente na Figura 10. De maneira geral, partindo-se do perfil obtido são construídos modelos estruturais *ab initio* formados por contas (*beads*), até que se obtenham estruturas que possam gerar o mesmo perfil. Muitas vezes, modelos comparativos são empregados para guiar a interpretação destes perfis (SVERGUN & KOCH, 2002).

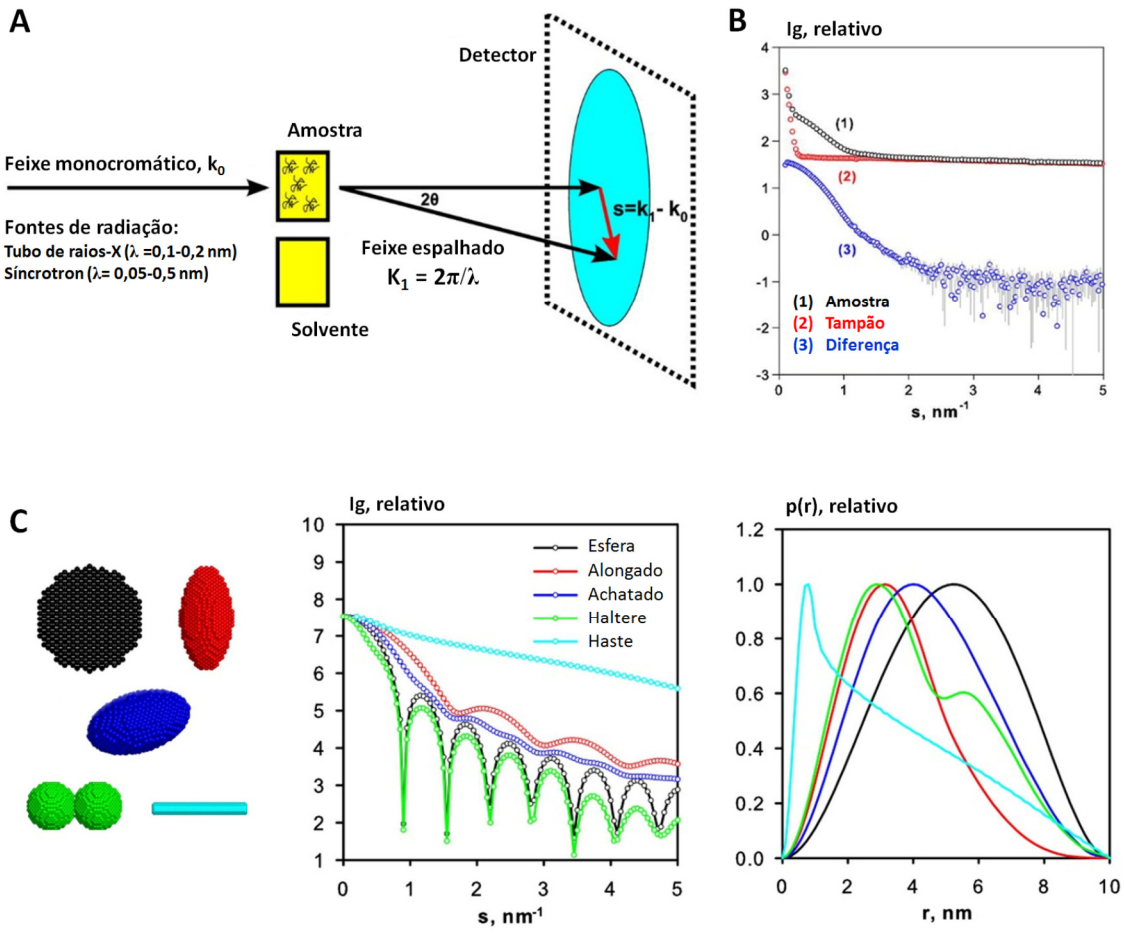


Figura 10. Etapas gerais da técnica de SAXS. (A) Representação esquemática do experimento típico de SAXS. (B) Padrões de espalhamento de raios-X de uma amostra proteica, do tampão, e a sua diferença,

contendo apenas a contribuição da proteína ajustada pela concentração de soluto. (C) Exemplos de sólidos geométricos e suas intensidades de espalhamento (Ig) e funções de distribuição de distância (p(r)); os modelos de *beads* são equivalentes aos empregados na modelagem de estruturas obtidas a partir da técnica de SAXS (adaptado de MERTENS & SVERGUN, 2010).

Dados estruturais obtidos a partir de perfis experimentais de SAXS podem ser comparados a estruturas proteicas, para as quais um perfil teórico de SAXS deve ser calculado. No presente trabalho, tal cálculo foi realizado no servidor FoXS (SCHNEIDMAN-DUHOVNY ET AL., 2010). Os perfis experimentais de SAXS para os complexos $(UreABC\text{-}UreD)_3$ e $(UreABC\text{-}UreDF)_3$ foram obtidos do trabalho de QUIROZ-VALENZUELA ET AL. (2008).

3.2.4 Avaliação de energia relativa de interação e modos normais

Levando-se em conta a ausência de dados de SAXS para validação das poses de ligação da UreG aos modelos de $(UreABC\text{-}UreDF)_3$, as soluções de atracamento foram agrupadas ("clusterizadas") com o MMTSB Tool Set (FEIG ET AL., 2004), pelo método de clusterização hierárquica baseada em RMSD mútuo, além de serem avaliadas em termos de energia relativa com FoldX (GUEROIS ET AL., 2002). Preferiu-se tal abordagem em vez da escolha da solução de melhor pontuação devido ao fato das conformações proteicas de baixa energia, próximas da conformação nativa, tenderem a se agrupar, fazendo com que *clusters* maiores indiquem com maior confiabilidade a solução real para o atracamento (conforme revisado por RITCHIE, 2008). Assim, é esperado que poses de ligação próximas da orientação nativa apresentem termos de energia mais baixos. Por serem derivados de atracamento rígido, os modelos tiveram suas interações de interface refinadas com ferramentas dedicadas para tal fim no FoldX antes de serem avaliados em termos energéticos.

Para inspecionar possíveis movimentos funcionais da estrutura obtida para o complexo $(UreABC–UreDFG)_3$, a apoproteína e o oligômero da urease foram submetidos à análise de modos normais (NMA). Essa análise é empregada no estudo de movimentos lentos de larga escala em macromoléculas (YANG ET AL., 2007). A análise classifica todas as deformações que uma proteína (modelada como um sistema oscilante harmônico) pode sofrer em torno de seu estado de equilíbrio. Os modos vibracionais de baixa frequência (ou baixa energia) correspondem aos movimentos coletivos, enquanto os movimentos de

alta frequência representam deformações locais. Os modos de baixa frequência estão associados a movimentos funcionalmente importantes para proteínas, e apenas um ou poucos deles são suficientes para descrever transições conformacionais nesses sistemas (SKJAERVEN ET AL., 2011). No presente trabalho, a análise de modos normais foi realizada com o servidor ElNémo (SUHRE & SANEJOUAND, 2004).

3.3 DINÂMICA DOS PEPTÍDEOS DERIVADOS DA UREASE

3.3.1 Estruturas peptídicas

A modelagem molecular comparativa foi empregada para obtenção das estrutura tridimensional do peptídeo Jaburetox e de seus mutantes (porções N- e C-terminais do peptídeo completo, correspondendo aos resíduos 1-44 e 45-93, respectivamente). A isoforma majoritária da urease de *C. ensiformis*, PDB ID 3LA4 (BALASUBRAMANIAN & PONNURAJ, 2010) foi empregada como molde. Todos os alinhamentos de sequência foram realizados com ClustalW (LARKIN ET AL., 2007) e a construção de modelos foi realizada com o programa Modeller 9v10 (SÁNCHEZ & ŠALI, 2000). Para cada caso, dez modelos foram construídos. Esses modelos foram avaliados estereoquimicamente com Procheck (LASKOWSKI ET AL., 1993) e seus perfis tridimensionais foram comparados a perfis ideais com Verify3D (LÜTHY ET AL., 1992). Os modelos de melhor pontuação para cada peptídeo foram então selecionados. A metionina aminoterminal e o segmento carboxiterminal LEHHHHHH, presentes nos três peptídeos e oriundos do processo de transformação, foram adicionados com SwissPDBviewer (GUEX & PEITSCH, 1993). Análises de hidropatia dos peptídeos foram realizadas via ProtParam (GASTEIGER ET AL., 2004), empregando a escala de KYTE & DOOLITTLE (1982).

3.3.2 Dinâmica molecular

O protocolo geral de simulação foi baseado em procedimentos previamente descritos (por exemplo, por DE GROOT & GRUBMÜLLER, 2001; SACHETT & VERLI, 2011; POL-FACHIN & VERLI, 2012). Foram realizadas três simulações: Jaburetox completo, mutante N-terminal e mutante C-terminal. Os cálculos de dinâmica molecular foram realizados com o pacote GROMACS 4.5

(HESS ET AL., 2008) empregado o campo de força GROMOS96 53a6 (OOSTENBRINK ET AL., 2004).

Os sistemas foram solvatados em caixas triclínicas, usando condições periódicas de contorno e o modelo de água SPC (BERENDSEN ET AL., 1987). Contra-íons (Na^+) foram adicionados para neutralização dos sistemas. O método Lincs (HESS ET AL., 1997) foi aplicado na restrição de ligações covalentes de forma a permitir um passo de integração de 2 fs, enquanto as interações eletrostáticas foram calculadas utilizando o método Particle-Mesh Ewald (DARDEN ET AL., 1993). A temperatura e a pressão do sistema foram mantidas constantes através do acoplamento dos peptídeos, íons e solvente a banhos externos de temperatura e pressão, utilizando constantes de acoplamento de, respectivamente, τ = 0,1 ps e τ = 0,5 ps (BERENDSEN ET AL., 1984). A constante dielétrica do meio foi tratada como ε = 1.

As simulações por dinâmica molecular geralmente iniciam com uma etapa de equilibração ou termalização. Nessa etapa ocorre o aquecimento gradativo do sistema, visando uniformizar as energias contidas na estrutura analisada (seja cristalográfica, derivada de RMN ou obtida por modelagem comparativa) para que sejam evitadas deformações nas moléculas em estudo. Para termalização, após 1 ps de restrição de posição, cada um dos sistemas foi aquecido lentamente de 50 K a 300 K de maneira que, em cada um dos seis passos de 5 ps, há o aumento da temperatura em 50 K. A simulação teve prosseguimento sem restrições na temperatura de equilíbrio de 300 K, por 500 ns, considerando um valor de referência de 3.5 Å entre átomos pesados para ligações de hidrogênio, e um ângulo de corte de 30° entre doadores e aceptores de hidrogênio (HESS ET AL., 2008).

4. RESULTADOS

"The section called 'results' consists of a stream of factual information in which it is considered extremely bad form to discuss the significance of the results you are getting."

Peter Medawar

Os resultados oriundos das metas propostas para este trabalho são apresentados na forma de artigos científicos. Cada uma das áreas é abordada conforme descrito a seguir.

A análise filogenética de ureases é apresentada no artigo "3-to-1: unraveling structural transitions in ureases", Ligabue-Braun R, Andreis FC, Verli H, Carlini CR. *Naturwissenschaften*. 100:459-467 (2013).

As propostas de modelos estruturais para os intermediários de ativação de ureases são apresentadas no artigo "Evidence-based docking of the urease activation complex", Ligabue-Braun R, Real-Guerra R, Carlini CR, Verli H. *Journal of Biomolecular Structure & Dynamics*. 31:854-861 (2013).

Por sua vez, a análise conformacional dos peptídeos bioativos derivados da urease de *C. ensiformis* foi somada a um conjunto de dados obtidos em bancada e é apresentada no artigo "Structure-function studies on jaburetox, a recombinant insecticidal and antifungal peptide derived from jack bean (*Canavalia ensiformis*) urease", Martinelli AH, Kappaun K, Ligabue-Braun R, Defferrari MS, Piovesan AR, Stanisçuaski F, Demartini DR, Dal Belo CA, Almeida CG, Follmer C, Verli H, Carlini CR, Pasquali G. *Biochimica et Biophysica Acta*. 1840:935-944 (2014)

Naturwissenschaften (2013) 100:459–467
DOI 10.1007/s00114-013-1045-2

ORIGINAL PAPER

3-to-1: unraveling structural transitions in ureases

Rodrigo Ligabue-Braun · Fábio Carrer Andreis · Hugo Verli · Célia Regina Carlini

Received: 5 February 2013 / Revised: 5 April 2013 / Accepted: 5 April 2013 / Published online: 26 April 2013

Abstract Ureases are nickel-dependent enzymes which catalyze the hydrolysis of urea to ammonia and carbamate. Despite the apparent wealth of data on ureases, many crucial aspects regarding these enzymes are still unknown, or constitute matter for ongoing debates. One of these is most certainly their structural organization: ureases from plants and fungi have a single unit, while bacterial and archaean ones have three-chained structures. However, the primitive state of these proteins — single- or three-chained — is yet unknown, despite many efforts in the field. Through phylogenetic inference using three different datasets and two different algorithms, we were able to observe chain number transitions displayed in a 3-to-1 fashion. Our results imply that the ancestral state for ureases is the three-chained organization, with single-chained ureases deriving from them. The two-chained variants are not evolutionary intermediates. A fusion process, different from those already studied, may explain this structural transition.

Keywords Urease · Phylogenetic tree · Structural transition · Evolution · Gene fusion · Gene disruption

Background

Ureases (urea amidohydrolases, EC 3.5.1.5) are found in plants, fungi and bacteria. These enzymes catalyze the hydrolysis of urea to ammonia and carbamate. The latter undergoes spontaneous hydrolysis to form carbonic acid and a second ammonia molecule (Mobley et al. 1995). Both urea and urease are hallmarks in the development of natural sciences (as reviewed by Krajewska 2009). After being discovered in human urine in

Communicated by: Sven Thatje

R. Ligabue-Braun and F.C. Andreis contributed equally to this work.

H. Verli and C.R. Carlini share senior authorship.

Electronic supplementary material The online version of this article (doi:10.1007/s00114-013-1045-2) contains supplementary material, which is available to authorized users.

R. Ligabue-Braun · H. Verli · C. R. Carlini
Graduate Program in Cellular and Molecular Biology, Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

R. Ligabue-Braun
e-mail: rodrigobraun@cbiot.ufrgs.br

F. C. Andreis
Biotechnology Undergraduate Program, Center of Biotechnology, UFRGS, Porto Alegre, RS, Brazil
e-mail: fabio.andreis@ufrgs.br

H. Verli
Center of Biotechnology and Faculty of Pharmacy, UFRGS, Porto Alegre, RS, Brazil

C. R. Carlini
Center of Biotechnology and Department of Biophysics-IB, UFRGS, Porto Alegre, RS, Brazil

H. Verli (✉)
Av. Bento Gonçalves, 9500, Campus do Vale, Caixa postal 15005, 91501-970 Porto Alegre, RS, Brazil
e-mail: hverli@cbiot.ufrgs.br

C. R. Carlini (✉)
Av. Bento Gonçalves, 9500, Campus do Vale, Prédio 43431, 91500-970 Porto Alegre, RS, Brazil
e-mail: ccarlini@ufrgs.br

1773, urea became the first organic compound synthesized from inorganic materials, in 1828. The notion that urine-derived ammonia originates from urea dates back to 1798, but only in 1890 the ureolytic enzyme was isolated. Jack bean (*Canavalia ensiformis*) urease was the first enzyme ever to be crystallized (Sumner 1926), proving that enzymes were proteins and that they could be crystallized. Of equal importance, this enzyme was the first shown to posses nickel ions in its active site, which are essential for its activity (Dixon et al. 1975). Despite the apparent wealth of data on ureases, many crucial aspects regarding these enzymes are still unknown, or constitute matter of ongoing debates. Such topics include the very nature of ureolysis (Karplus et al. 1997; Benini et al. 1999; Estiu and Merz 2007) and the mechanisms underlying urease activation (Carter et al. 2009; Zambelli et al. 2011).

One of the most striking aspects of ureases is their structural organization (Fig. 1). Microbial ureases are composed by three or two chains, while plant and fungal ureases are composed by a single subunit. The amino acid sequences of the smaller subunits of microbial ureases are very similar to the corresponding regions in the single units of eukaryotic ureases (Krajewska 2009).

This high similarity observed among ureases from different kingdoms suggest that they all derive from a common ancestral protein, and are likely to have similar tertiary structures and catalytic mechanisms (Jabri et al. 1995; Mobley et al. 1995; Sirko and Brodzik 2000; Carlini and Polacco 2008). These observations, however, do not address a central topic in urease structure, i.e., what was the primitive structural state of these enzymes? To this point, two possibilities arise, as pointed out previously by Hausinger (1993): "Did the gene encoding the single-subunit plant enzyme undergo disruption to yield the multiple genes encoding the two or three bacterial subunits? Or did the bacterial genes fuse to form the gene encoding the plant subunit?" Our work intended to help answering these long-held questions which, despite many efforts in the field, remained unanswered. By means of large-scale phylogenetic analyses, we were able to track the structural transition from three to one subunit in ureases, revealing also that the two-chained variants are not evolutionary intermediates between them. Hypotheses for the genetic fusion in these enzymes are also presented.

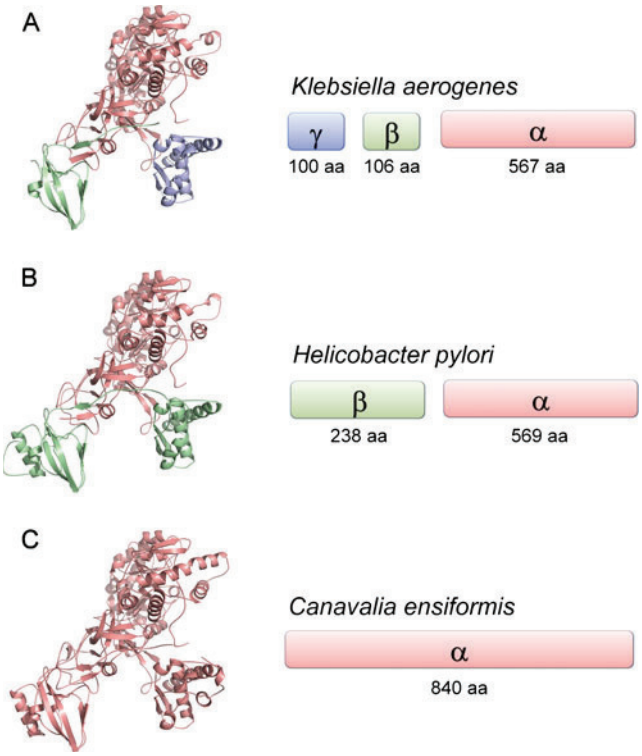


Fig. 1 Structural organization of ureases. Three-dimensional and schematic representations of the subunit organization of ureases. **a** Typical microbial urease composed by three chains (crystallographic structure from PDB ID 1FWJ). **b** Helicobacteraceae urease composed by two chains (crystallographic structure from PDB ID 1E9Z). **c** Typical eukaryotic urease composed by a single chain (crystallographic structure from PDB ID 3LA4)

Methods

All urease amino acid sequences (resulting from the "urease" keyword search) were retrieved from the National Center for Biotechnology Information (Sayers et al. 2012) on July 5, 2010. In order to obtain only complete urease sequences, manual filtering was carried out. Sequences that were incomplete, mistakenly labeled or related to urease accessory proteins were removed. All sequences related to Bacteria and Archaea in the resulting data were cross-compared to find similarities among them. The ClustalW algorithm (Larkin et al. 2007) was employed on all alignments, clustering together sequences with identity greater than 95 %. For practical reasons, these sequences were grouped by source organism genus. From the original 14,221 sequences data set, 32 sequences were from eukaryotic sources and did not require further filtering. Through sequence separation by genus, the number was reduced to 162 microbial representative sequences. These microbial ureases had their subunits' sequences aligned to ureases from plants and fungi, and these subunits were then joined together in a single sequence based on these alignments (forming a single $\gamma\beta\alpha$ sequence, or equivalent). The number of microbial sequences was further decreased to reduce the computational load. Based on the ClustalW alignment guide tree, highly similar sequences (80 % identical or higher) were considered as identical, and only one representative sequence was chosen. The final microbial data set was put together and aligned with the eukaryotic sequences, totaling 124 sequences in the final urease list (Table S1). Alignment sections with long gaps were removed.

To inspect the aligned urease sequences for highly variable regions, SimPlot software (Lole et al. 1999) was employed. To do so, the edited dataset was branched into three distinct subsets: complete urease sequences (without

long gaps), highly variable regions and conserved regions. Amino acid sequences evaluation by the ProtTest software (Abascal et al. 2005) and MEGA5 software (Tamura et al. 2011) suggested the Whelan and Goldman (WAG) model with discrete Gamma distribution to account for evolutionary rate differences among sites (four categories, +G), considering some sites to be evolutionarily invariable (+I) (Whelan and Goldman 2001). Maximum likelihood (ML) phylogenetic trees were calculated with MEGA5 with 1000 bootstrap replicates, and Bayesian inference (BI) trees with MrBayes software (Huelsenbeck and Ronquist 2001) for 7.5×10^6 generations, sampled every tenth generation. All trees were rooted in the branch containing the most archaean urease sequences. The obtained trees were visualized and edited with FigTree software (Rambaut 2012).

Results

Using ML and BI upon our alignments (Fig. S1), we were able to generate two phylogenetic trees for the complete urease sequences dataset (Figs. 2 and 3) and two for the conserved regions dataset (Figs. 4 and 5). For the highly variable regions dataset, one tree was generated through ML (Fig. S1), with very low bootstrap values. Convergence could not be reached throughout BI calculations, meaning that no considerable result could be obtained with these variable regions.

All trees displayed similar general convergence regarding their branching, with minor differences. We observed that, in the majority of cases, sequences belonging to organisms within the same phylum grouped together. All inferences displayed a similar distribution trend regarding ureases from five microbial groups composed of Euryarchaeota, Firmicutes, Actinobacteria, Proteobacteria and Cyanobacteria. Also, all phylogenetic inferences suggest that the number of urease chains evolved in a 3-to-1 fashion: three-chained ureases (those of most microbes) were of earlier existence, with a later structural unification originating single-chained ureases (those of plants and fungi). Two-chained ureases, belonging to Helicobacteraceae, are displayed as special situations among the three-chained enzymes, opposing the hypothesis considering them as intermediates between single- and three-chained ureases.

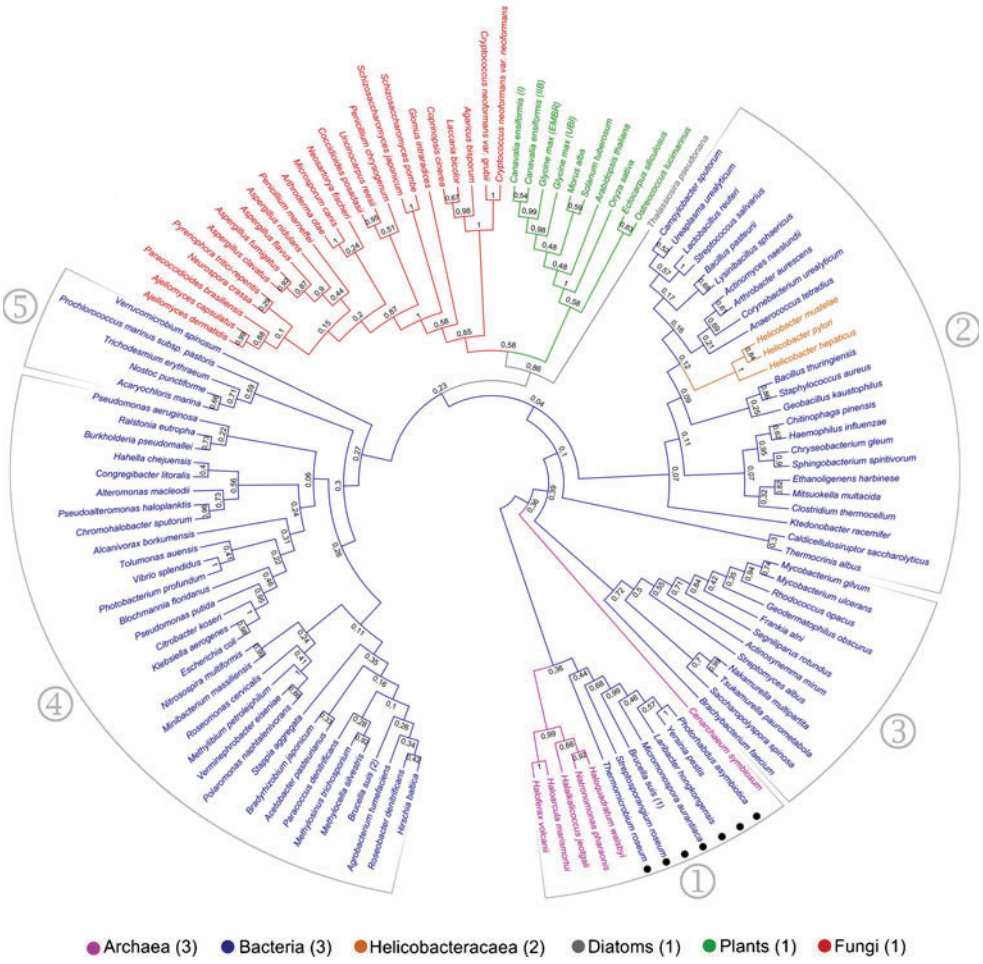


Fig. 2 Molecular phylogenetic analysis of complete urease sequences by maximum likelihood method. The evolutionary history was inferred by using the ML method based on the WAG+G+I model. The number of chains composing ureases from different groups is given in *brackets*. General microbial phyla separations are marked in *grey* (*1* Euryarchaeota, *2* Firmicutes, *3* Actinobacteria, *4* Proteobacteria, *5* Cyanobacteria). Grouping outliers are marked with *black dots*

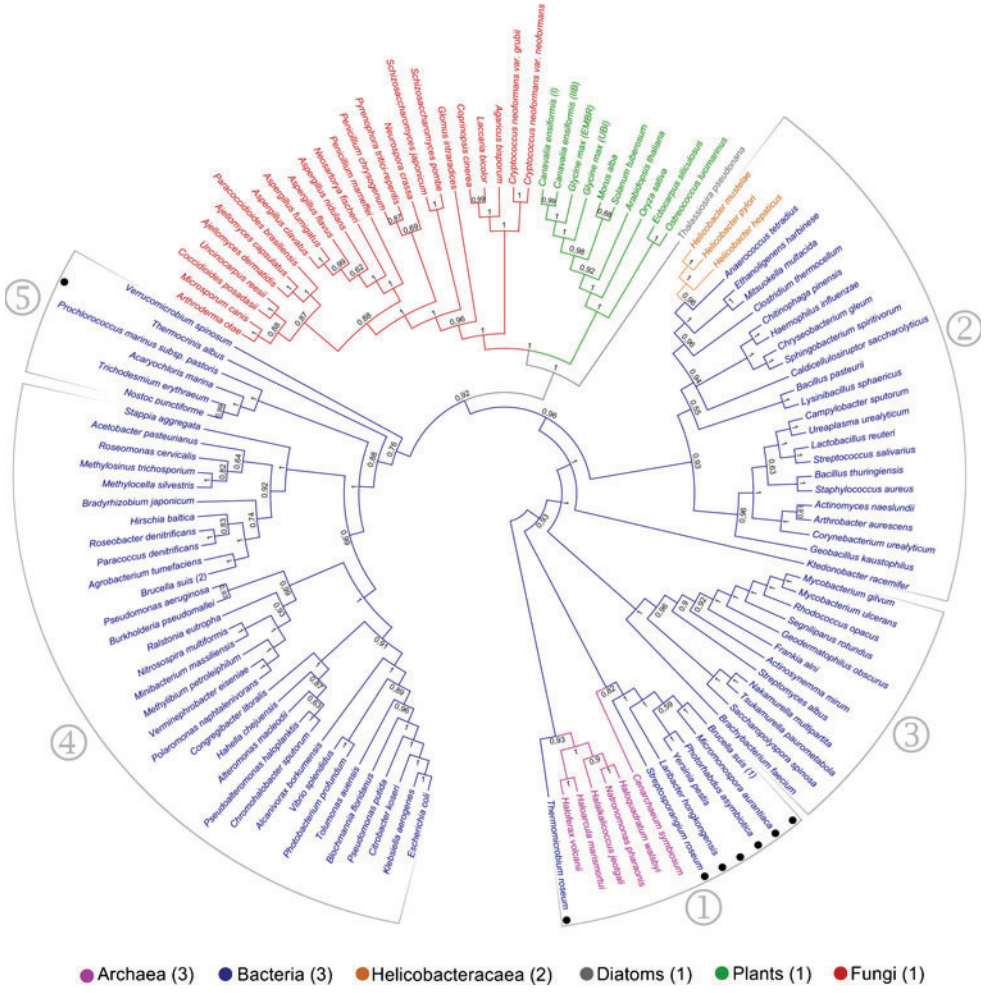


Fig. 3 Molecular phylogenetic analysis of complete urease sequences by Bayesian Inference method. The evolutionary history was inferred by using the Bayesian method based on the WAG+G+I model. The number of chains composing ureases from different groups is given in *brackets*. General microbial phyla separations are marked in *grey* (*1* Euryarchaeota, *2* Firmicutes, *3* Actinobacteria, *4* Proteobacteria, *5* Cyanobacteria). Grouping outliers are marked with *black dots*

Discussion

Urease has been considered an ancient enzyme, related to the putative primordial peptide cycle (Huber et al. 2003). This enzyme is found in the three domains of life, being synthesized in archaea, bacteria, fungi and plants (Carlini and Polacco 2008). There is uncertainty regarding urease presence in animals, and while some findings indicate its presence in some invertebrates (Pedrozo et al. 1996), others indicate that the enzyme comes from exterior sources (Hirayama et al. 2000) and that all animals lost urease in their evolutionary history (Fujiwara and Noguchi 1995). For this reason, ureases from putative animal sources were not included in our datasets.

Previous published attempts to establish the evolutionary history of ureases were based only in the catalytic subunit of microbial ureases (Contreras-Rodriguez et al. 2008), the full eukaryotic enzyme (Mulinari et al. 2011), or association of phylogenies at the organism level with enzyme characteristics (Navarathna et al. 2010). In this work we were able to reconstruct the possible evolutionary pathway followed by ureases in their structural transitions from multiple to single chains, supporting the "fusion hypothesis" (Hausinger 1993). It could be argued that this hypothesis was already supported by application of Ockham's razor (Gernert 2007), since the most abundant and primitive ureases appear to be three-chained, and therefore less likely to have all suffered fission. Only this observation, however, is too simplistic to grant validity to this assumption. The convergent results from four phylogenies, built based on two different methods upon two different datasets, nevertheless lend weight to the parsimony-derived conclusion. Additionally, these trees agree with previous analyses of the urease operon organization in fully sequenced microbial genomes (Zambelli et al. 2011). The clades as obtained in this work have *ure* operons with distinct structures: Clade 1 representatives (*H. marismortui*, *N. pharaonis*, *H. walsbyi*) are organized as *UreBCAGDEF*; Clade 2 (*C. thermocellum*, *H. influenzae*, *G. kaaustophilus*, *S. aureus*, *C. urealyticum*, *A. aurescens*, *L. sphaericus*, *U. urealyticum*) are organized as *UreABCEFGD* with the exception of *Helicobacter pylori* and *H. hepaticus* which are organized as *Ure(AB)C*-(unrelated gene)-*UreEFGD*; Clade 3 has only one analyzed representative (*M. gilvum*), organized as *UreABCGDEF*; Clade 4 has mixed

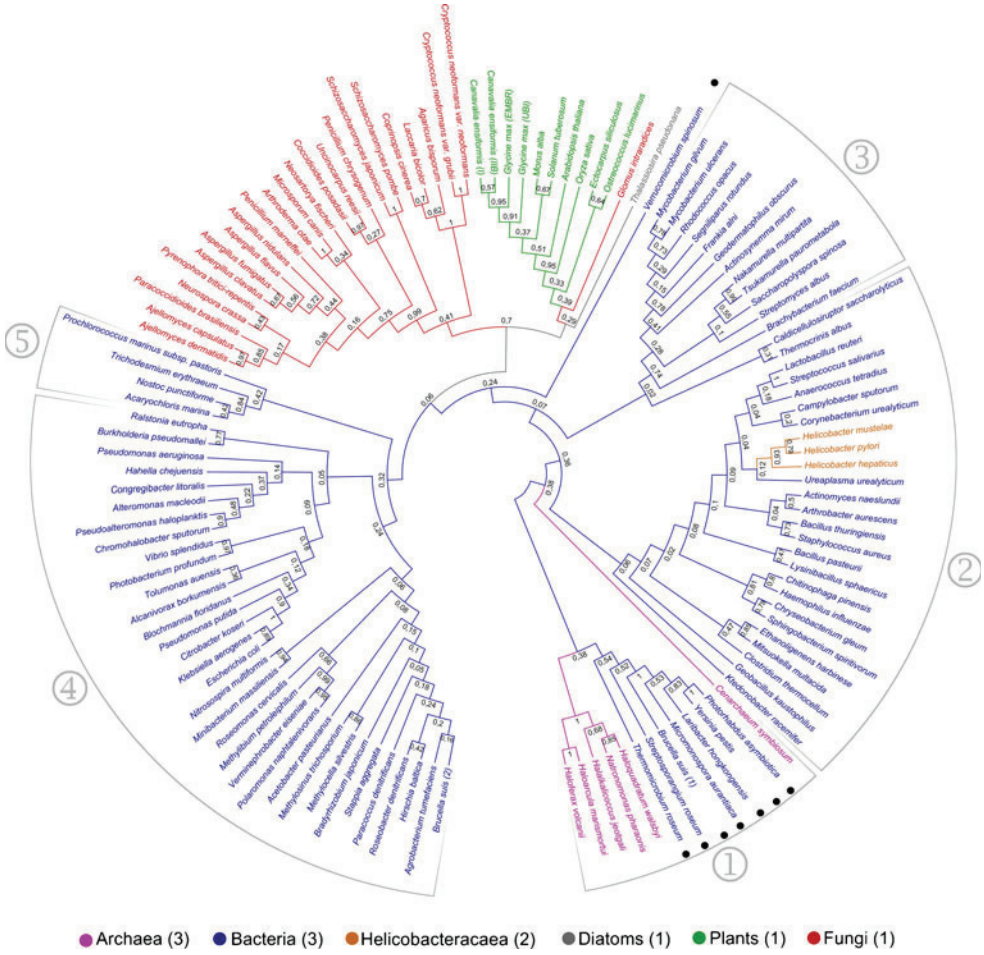


Fig. 4 Molecular phylogenetic analysis of conserved regions of urease sequences by Maximum Likelihood method. The evolutionary history was inferred by using the ML method based on the WAG+G+I model. The number of chains composing ureases from different groups is given in *brackets*. General microbial phyla separations are marked in *grey* (*1* Euryarchaeota, *2* Firmicutes, *3* Actinobacteria, *4* Proteobacteria, *5* Cyanobacteria). Grouping outliers are marked with *black dots*

assembles, including urease genes not organized in recognizable operons (*A. marina*); and Clade 5 (*P. aeruginosa*, *R. eutropha*, *B. pseudomallei*, *H. chejuensis*, *A. macleodii*, *P. haloplanktis*, *A. borkumensis*, *P. putida*, *E. coli*, *N. multiformis*, *P. naphtalivorans*, P. *denitrificans*, *M. silvestris*, *R. denitrificans*) is organized as either *UreDABCEFG* or *UreDA*-(unrelated gene)-*UreBCEFG*.

The enzyme dihydroorotase is considered the ancestral of all urease-related amidohydrolases (Holm and Sander 1997) and would be the ideal sequence to root the urease phylogenies. Dihydroorotase, however, is much shorter than urease and alignments including its sequence would interfere in the resulting trees. For this reason, a root had to be chosen among the clades in the urease trees. This rooting was performed by manually selecting the Archaean urease clade as a midpoint root, considering the intermediate phylogenetic position of Archaea in the tree of life (Woese et al. 1990), as a way of allowing comparisons among the different trees. When rooted by mathematical midpoint (Hess and Russo 2007), the trees showed no significant change in respect to those rooted in Archaean ureases (Supplementary Figs. S3, S4, S5 and S6). Inclusion in the alignments of regions that are highly variable among species was shown to be a source of "molecular noise" in the phylogenetic analyses. Attempts to obtain phylogenetic data from the highly variable regions were not statistically reliable, while removal of these regions had little impact on the obtained trees. These observations point to advantages in working with less variable regions of alignments, an approach employed for taxonomic phylogenies of fungi (Câmara et al. 2002), arthropods (Arango 2003; Hunt and Vogler 2008), cyanobacteria (Gaylarde et al. 2005), and viruses (Korber et al. 2001).

The occurrence of three-chained ureases in two of the three domains of life (i.e., Bacteria and Archaea) indicates that this structural organization is widespread and may be considered primitive in relation to one-chained ureases. The two-chained enzymes, which could have been taken as intermediates between single- and triple-chained ureases, seem to have arisen from bacterial triple chains in a process unrelated to the origin of single chained enzymes. Bacteria from the genus *Helicobacter* spp. are known to be subject to distinct selective pressures related to surviving in the gastric environment (Gueneau and Loiseaux-De Goër 2002) and the dodecameric macromolecular organization observed for *H. pylori* and *H. mustelae* ureases seem to be an adaptive response to such harsh conditions (Ha et al. 2001; Carter et

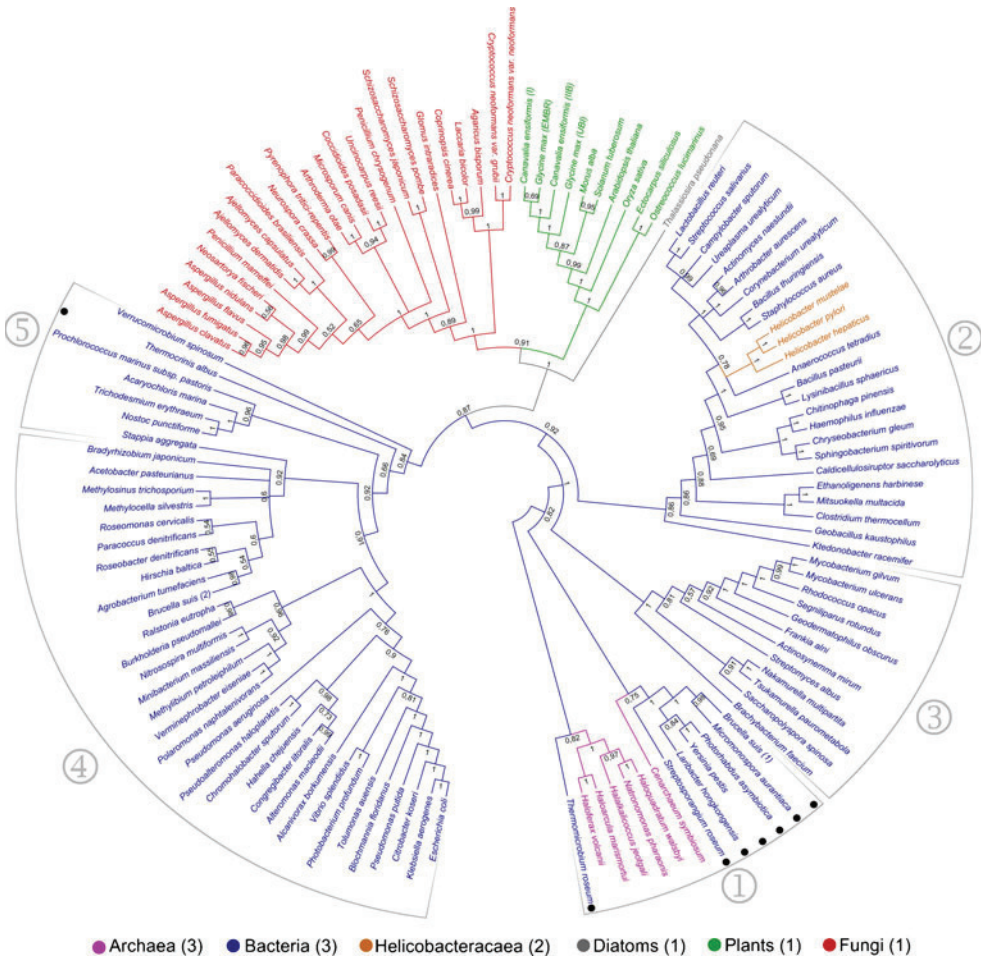


Fig. 5 Molecular phylogenetic analysis of conserved regions of urease sequences by Bayesian Inference method. The evolutionary history was inferred by using the Bayesian method based on the WAG+G+I model. The number of chains composing ureases from different groups is given in *brackets*. General microbial phyla separations are marked in *grey* (*1* Euryarchaeota, *2* Firmicutes, *3* Actinobacteria, *4* Proteobacteria, *5* Cyanobacteria). Grouping outliers are marked with *black dots*

al. 2011). The linking region (connecting the equivalents of subunits γ and β) of *Helicobacter* spp., which is highly divergent in relation to the same region in eukaryotes (Ha et al. 2001), could take part in this differential organization.

There are advantages associated with linking subunits. By artificial genetic fusion, it has been shown that some oligomeric viral proteins are benefited with enhanced folding rate and structural stability, and increased tolerance to insertions of other segments (Liang et al. 1993; Ma et al. 1993; Peabody 1997). When artificially fused, the genetic subunits of cytochrome ubiquinol oxidase from *E. coli* yielded an active enzyme, similar to the one from *Thermus thermophilus* (Ma et al. 1993). Differently from natural genetic fusions in *Drosophila* (Jones and Begun 2005), which resulted in new genes with new functions, the joining of urease subunits at the genetic level incorporating linking segments kept the original ureolytic activity. Ureases, however, are recognized as multifunctional (or moonlighting) proteins. They have many catalysis-independent effects, including neurotoxicity to mammals, insecticidal activities against Coleopterans and Hemipterans, fungistatic properties, pro-inflammatory roles, glycoconjugate binding properties, platelet activation ability, and inter-specific communication action (reviewed by Carlini and Polacco 2008; Stanisçuaski and Carlini 2012). Most of these properties have not yet been mapped to particular regions of these proteins, the exception being an "entomotoxic domain" containing the insectcidal peptide(s) released by *Canavalia ensiformis* ureases upon insect digestion (Ferreira-DaSilva et al. 2000; Piovesan et al. 2008; Defferrari et al. 2011). This "domain" is located in the β–α intersubunit region, and may be subject to faster divergence rates, since it is not involved in catalysis or subunit association (Mulinari et al. 2007). Also intriguing is the need for non-catalytic subunits/domains in urease, considering that only the TIM-barrel sub-domain from the α subunit is responsible for catalysis (Balasubramanian and Ponnuraj 2010). While the β subunit has been implicated in the urease activation process (Carter et al. 2011), no specific function was ascribed to the γ subunit.

Reviewing the literature we were not able to track other proteins that have undergone a similar process of natural subunit fusion. The events observed for ureases seem unparalleled, and not related to immunoglobulin genetic fusion (Tonegawa 1983) or exon shuffling (Kaessman 2010). The main difference is that for most of the studied microbial ureases, the γ, β, and α subunits are ordered and genetically

adjacent in the same operon (Zambelli et al. 2011), and not dispersed along the genome, as is the case for precursors of many merged genes. It is expected from the different mechanisms that may lead to chimeric genes that both inter and intragenic regions can be equally affected by segment insertions (Kaessman 2010). This does not seem to have occurred for ureases, since only intergenic regions were incorporated.

From the phylogenies, it is inferred that the transition from three subunits to one unit occurred as a single event. One mechanism that could be held responsible for such one-step result would be transcription or translation readthrough, which bypasses stop codons, incorporating intergenic regions as coding sequences. Translation bypass of stop codons has been well documented in yeast, where it takes part in complex regulatory mechanisms (von der Haar and Tuite 2007), while transcriptional readthrough has been implicated in human genetic disorders (Du et al. 2009), plant responses to stress (Hernández-Pinzón et al. 2009), and prostate adenocarcinoma, where many transcription-induced chimeras were found (Nacu et al. 2011). For ureases, however, readthrough events alone would not explain how the fused genes were finally incorporated into the genomes of their source organisms, requiring other subsequent process(s) at transcriptional and translational levels. We speculate that different genetic codes would be responsible for stop codons being unrecognized as such, allowing continuous transcription of the urease $\gamma\beta\alpha$ complex into a single chain. One such candidate would be the Chlorophyceae *Scenedesmus obliquus* mitochondrial code, which takes UCA as a stop codon instead of coding for serine, as occurs in the standard genetic code (Nedelcu et al. 2000). In this scenario, for intergenic codes to be translated it would be required that urease genes were transferred from the mitochondrial genome to the nuclear genome, where stop codons would not be recognized as such. There are, however, some difficulties with this hypothesis. No UCA codon is found in the terminal serine position of region γ from plants and fungi enzymes, and no clear serine position is found in the β chain terminus of the same ureases. Regarding the serine codons in the terminus of γ chain, it could be argued that serine would be beneficial in that position, thus allowing only conservative mutations of that originally misinterpreted codon. Regarding cellular location, urease is generally considered a cytoplasmic protein, but it has also been found in membrane fractions and cell wall from plants (Aguetoni Cambuí et al. 2009). Proteomic studies also indicate that *C. ensiformis* urease is either bound to mitochondria or spatially related to mitochondrial proteins (Demartini et al. 2011). Inter-specific horizontal gene transfer could also be responsible for the hypothetical readthrough. These transfers are now considered a major genome shaping tool when involving transposable elements (Schaack et al. 2010), and many instances of gene transfer from bacteria to eukaryotes have been documented, including organelle-to-nucleus transfers, such as the nuclear mitochondrial insert transferred to the *A. thaliana* chromosome 2 (Dunning Hotopp 2011). Little data is available on inter-specific urease genetic transfers and the only case reported so far is the second *ure* gene cluster from *Brucella suis* (Contreras-Rodriguez et al. 2008).

The genetic code readthrough hypothesis would also require intergenic regions of sufficient length to account for the incorporated segments. Urease operons from some bacteria related to Clade 2, such as *Staphylococcus saprophyticus*, *Streptococcus thermophilus*, and *Corynebacterium glutamicum* (Gene IDs 3616069, 3167116, and 1021080, respectively), have intercistronic regions that would satisfy this requirement. In other cases, exemplified by the *Proteus mirabilis* urease, there are not enough codons in these regions. On the contrary, there is even superposition of the last codon of the β subunit with the start codon of the α subunit (Jones and Mobley 1989). When eukaryotic gene organization is taken into account, the picture becomes even more complex. A preliminary inspection of annotated genomes at the Ensembl database (Flicek et al. 2012), revealed that all urease genes from plants deposited in there, i.e., *Arabidopsis thaliana*, *Brassica rapa*, *Glycine max* (both isoforms), *Oryza sativa*, *Physcomitrella patens*, *Populus trichocarpa*, *Sorghum bicolor*, and *Vitis vinifera*, have their coding sequences arranged in 18 exons. The same number of exons is observed for *Solanum tuberosum*, the first (and so far, only) case of alternative splicing in ureases (Witte et al. 2005). For fungal urease genes deposited at Ensembl, the number of exons is variable. It ranges from a single coding sequence with no introns (*Magnaporthe poae*, *Ustilago maydis*, *Schizosaccharomyces pombe*), to a variable number of exons: three (*Fusarium oxysporum*, *Gaeumannomyces graminis*, *Gibberella moniliformis*, *Magnaporthe oryzae*), four (*Aspergillus fumigatus*, *Neurospora crassa*), five (*A. fumigatus*, *A. niger*, *Nectria haematococca*, *Gibberella zeae*), 6 (*A. flavus*, *A. nidulans*), 7 (*Neosartorya fischeri*, *Phaeosphaeria nodorum*), 8 (*Aspergillus terreus*), 9 (*Fusarium oxysporum*), 13 (*Puccinia graministritici*), 14 (*Gibberella zeae*) or 15 (*Puccinia graministritici*). The difference in exon number between plants and fungi may be a reflection of strict structural conservation in plants. Until more genomes are deciphered, allowing a more in-depth analysis of their urease-coding segments, this interpretation remains speculative.

Conclusions

From the phylogenies presented in this work, we conclude that ureases were originally composed by three chains, and their transition to single-chained enzymes did not involve two-chained intermediates. We also speculate that the 3-to-1 transition took place as a single event, and hypothesize on a mechanism that would

result in the fused urease. Nonetheless, many questions remain unanswered. It is as if the unraveling of urease evolutionary paths (the "what") begins to be established, while the mechanisms underlying the urease structural transitions (the "how") still await further investigation. We expect that the large datasets and multiple approaches employed in this work contribute to enhance the comprehension of the unique case of urease subunits fusion, encouraging further research on the subject.

Acknowledgments The authors thank Charley C. Staats, Cláudia L. Fernandes, Dennis M. Junqueira, and Marilene H. Vainstein for many insightful suggestions and friendly help with the methodology. This work was supported by the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), MCT; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), MEC; Financiadora de Estudos e Projetos (FINEP); and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Aguetoni Cambuí C, Gaspar M, Mercier H (2009) Detection of urease in the cell wall and membranes from leaf tissues of bromeliad species. Physiol Plant 136:86–93

Arango CP (2003) Molecular approach to the phylogenetics of sea spiders (Arthropoda: Pycnogonida) using partial sequences of nuclear ribosomal DNA. Mol Phylogenet Evol 28:588–600

Balasubramanian A, Ponnuraj K (2010) Crystal structure of the first plant urease from jack bean: 83 years of journey from its first crystal to molecular structure. J Mol Biol 400:274–283

Benini S, Rypniewski WR, Wilson KS, Miletti S, Ciurli S, Mangani S (1999) A new proposal for urease mechanism based on the crystal structures of the native and inhibited enzyme from *Bacillus pasteurii*: why urea hydrolysis costs two nickels. Structure 7:205–216

Câmara MP, Palm ME, van Berkum P, O'Neill NR (2002) Molecular phylogeny of Leptosphaeria and Phaeosphaeria. Mycologia 94:630–640

Carlini CR, Polacco JC (2008) Toxic properties of ureases. Crop Sci 48:1665–1672

Carter EL, Flugga N, Boer JL, Mulrooney SB, Hausinger RP (2009) Interplay of metal ions and urease. Metallomics 1:207–221

Carter EL, Boer JL, Farrugia MA, Flugga N, Towns CL, Hausinger RP (2011) Function of UreB in *Klebsiella aerogenes* urease. Biochemistry 50:9296–9308

Contreras-Rodriguez A, Quiroz-Limon J, Martins AM, Peralta H, Avila-Calderon E, Sriranganathan N, Boyle SM, Lopez-Merino A (2008) Enzymatic, immunological and phylogenetic characterization of *Brucella suis* urease. BMC Microbiol 8:121

Defferrari MS, Demartini DR, Marcelino TB, Pinto PM, Carlini CR (2011) Insecticidal effect of *Canavalia ensiformis* major urease on nymphs of the milkweed bug *Oncopeltus fasciatus* and characterization of digestive peptidases. Insect Biochem Mol Biol 41:388–399

Demartini DR, Carlini CR, Thelen JJ (2011) Global and targeted proteomics in developing jack bean (*Canavalia ensiformis*) seedlings: an investigation of urease isoforms mobilization in early stages of development. Plant Mol Biol 75:53–65

Dixon NE, Gazzola C, Blakeley RL, Zerner B (1975) Jack bean urease (EC 3.5.1.5). A metalloenzyme. A simple biological role for nickel? J Am Chem Soc 97:4131–4133

Du L, Damoiseaux R, Nahas S, Gao K, Hu H, Pollard JM, Goldstine J, Jung ME, Henning SM, Bertoni C, Gatti RA (2009) Nonaminoglycoside compounds induce readthrough of nonsense mutations. J Exp Med 206:2285–2297

Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. Trends Genet 27:157–163

Estiu G, Merz KM Jr (2007) Competitive hydrolytic and elimination mechanisms in the urease catalyzed decomposition of urea. J Phys Chem B 111:10263–10274

Ferreira-DaSilva CT, Gombarovits ME, Masuda H, Oliveira CM, Carlini CR (2000) Proteolytic activation of canatoxin, a plant toxic protein, by insect cathepsin-like enzymes. Arch Insect Biochem Physiol 44:162–171

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS et al (2012) Ensembl 2012. Nucleic Acids Res 40:D84–D90

Fujiwara S, Noguchi T (1995) Degradation of purines: only ureidoglycollate lyase out of four allantoin-degrading enzymes is present in mammals. Biochem J 312:315–318

Gaylarde PM, Crispim CA, Neilan BA, Gaylarde CC (2005) Cyanobacteria from Brazilian building walls are distant relatives of aquatic genera. Omics 9:30–42

Gernert D (2007) Ockham's razor and its improper use. J Sci Explor 21:135–140

Gueneau P, Loiseaux-De Goër S (2002) Helicobacter: molecular phylogeny and the origin of gastric colonization in the genus. Infect. Genet Evol 1:215–223

Ha NC, Oh ST, Sung JY, Cha KA, Lee MH, Oh BH (2001) Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. Nat Struct Biol 8:505–509

Hausinger RP (1993) Urease. In: Hausinger RP (ed) Biochemistry of nickel. Plenum, New York, pp 23–57

Hernández-Pinzón I, de Jesús E, Santiago N, Casacuberta JM (2009) The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. J Mol Evol 68:269–278

Hess PN, Russo CAM (2007) An empirical test of the midpoint rooting method. Biol J Linn Soc 92:669–674

Hirayama C, Sugimura M, Saito H, Nakamura M (2000) Host plant urease in the hemolymph of the silkworm, *Bombyx mori*. J Insect Physiol 46:1415–1421

Holm L, Sander C (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 28:72–82

Huber C, Eisenreich W, Hecht S, Wächtershäuser G (2003) A possible primordial peptide cycle. Science 301:938–940

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hunt T, Vogler AP (2008) A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. Mol Phylogenet Evol 47:289–301

Jabri E, Carr MB, Hausinger RP, Karplus PA (1995) The crystal structure of urease from *Klebsiella aerogenes*. Science 268:998–1004

Jones CD, Begun DJ (2005) Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci USA 102:11373–11378

Jones BD, Mobley HL (1989) *Proteus mirabilis* urease: nucleotide sequence determination and comparison with jack bean urease. J Bacteriol 171:6414–6422

Kaessman H (2010) Origins, evolution, and phylogenetic impact of new genes. Genome Res 20:1313–1326

Karplus PA, Pearson MA, Hausinger RP (1997) 70 years of crystalline urease: what have we learned? Acc Chem Res 30:330–337

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58:19–42

Krajewska B (2009) Ureases I. Functional, catalytic and kinetic properties: a review. J Mol Catal B Enzym 59:9–21

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) ClustalW and ClustalX version 2. Bioinformatics 23:2947–2948

Liang H, Sandberg WS, Terwilliger TC (1993) Genetic fusion of subunits of a dimeric protein substantially enhances its stability and rate of folding. Proc Natl Acad Sci USA 90:7010–7014

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73:152–160

Ma J, Lemieux L, Gennis RB (1993) Genetic fusion of subunits I, II, and III of the cytochrome bo ubiquinol oxidase from *Escherichia coli* results in a fully assembled and active enzyme. Biochemistry 32:7692–7697

Mobley HL, Island MD, Hausinger RP (1995) Molecular biology of microbial ureases. Microbiol Rev 59:451–480

Mulinari F, Stanisçuaski F, Bertholdo-Vargas LR, Postal M, Oliveira-Neto OB, Rigden DJ, Grossi-de-Sá MF, Carlini CR (2007) Jaburetox-2Ec: an insecticidal peptide derived from an isoform of urease from the plant *Canavalia ensiformis*. Peptides 28:2042–2050

Mulinari F, Becker-Ritt AB, Demartini DR, Ligabue-Braun R, Stanisçuaski F, Verli H, Fragoso RR, Schroeder EK, Carlini CR, Grossi-de-Sá MF (2011) Characterization of JBURE-IIb isoform of *Canavalia ensiformis* (L.) DC urease. Biochim Biophys Acta 1814:1758–1768

Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, Peters BA, Modrusan Z, Jung K, Seshagiri S, Wu TD (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics 4:11

Navarathna DH, Harris SD, Roberts DD, Nickerson KW (2010) Evolutionary aspects of urea utilization by fungi. FEMS Yeast Res 10:209–213

Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G (2000) The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. Genome Res 10:819–831

Peabody DS (1997) Subunit fusion confers tolerance to peptide insertions in a virus coat protein. Arch Biochem Biophys 347:85–92

Pedrozo HA, Schwartz Z, Luther M, Dean DD, Boyan BD, Wiederhold ML (1996) A mechanism of adaptation to hypergravity in the statocyst of *Aplysia californica*. Hear Res 102:51–62

Piovesan AR, Stanisçuaski F, Marco-Salvadori J, Real-Guerra R, Defferrari MS, Carlini CR (2008) Stage-specific gut proteinases of the cotton stainer bug *Dysdercus peruvianus*: role in the release of entomotoxic peptides from *Canavalia ensiformis* urease. Insect Biochem Mol Biol 38:1023–1032

Rambaut A (2012) FigTree [http://tree.bio.ed.ac.uk/software/figtree/]

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD et al (2012) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 40:D13–D25

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546

Sirko A, Brodzik R (2000) Plant ureases: roles and regulation. Acta Bioch Pol 4:1189–1195

Stanisçuaski F, Carlini CR (2012) Plant ureases and related peptides: understanding their entomotoxic properties. Toxins 4:55–67

Sumner JB (1926) The isolation and crystallization of the enzyme urease. J Biol Chem 69:435–441

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581

von der Haar T, Tuite MF (2007) Regulated translational bypass of stop codons in yeast. Trends Microbiol 15:78–86

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Witte CP, Tiller S, Isidore E, Davies HV, Taylor MA (2005) Analysis of two alleles of the urease gene from potato: polymorphisms, expression, and extensive alternative splicing of the corresponding mRNA. J Exp Bot 56:91–99

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Zambelli B, Musiani F, Benini S, Ciurli S (2011) Chemistry of Ni^{2+} in urease: sensing, trafficking, and catalysis. Acc Chem Res 44:520–530

Table S1: GenInfo Identifiers for the urease sequences used in this work.

Source organism	GI		
	Alpha	**Beta**	**Gamma**
Acaryochloris marina	158338215	158338212	158338210
Acetobacter pasteurianus	258543403	258543404	258543405
Actinomyces naeslundii	4249601	4249599	326774128
Actinosynnema mirum	256374659	256374660	256374661
Agaricus bisporus	108859293		
Agrobacterium tumefaciens	15889672	159185200	159185201
Ajellomyces capsulatus	154281373		
Ajellomyces dermatitidis	261194902		
Alcanivorax borkumensis	110835580	110835581	110835582
Alteromonas macleodii	332142250	332142251	332142252
Anaerococcus tetradius	227500923	257066223	227484813
Arabidopsis thaliana	15220459		
Arthrobacter aurescens	119962058	119961176	119961081
Arthroderma otae	296818321		
Aspergillus clavatus	121702879		
Aspergillus flavus	238486420		
Aspergillus fumigatus	70990710		
Aspergillus nidulans	259489323		
Bacillus pasteurii	4557957	4557956	4557955
Bacillus thuringiensis	229061206	228909453	42782714
Blochmannia floridanus	33519971	33519972	33519973
Brachybacterium faecium	257067541	257067542	257067543
Bradyrhizobium japonicum	27376568	27376566	27376565
Brucella suis	225627811	237815759	225627809
Brucella suis	23501177	17987936	62289265
Burkholderia pseudomallei	53726283	53720268	53720267
Caldicellulosiruptor saccharolyticus	146297464	146297465	146297467
Campylobacter sputorum	260162332	260162331	260162330
Canavalia ensiformis	465008		
Canavalia ensiformis	219391588		
Cenarchaeum symbiosum	118575651	118575650	118575649
Chitinophaga pinensis	256420519	256420518	256420517
Chromohalobacter salexigens	92114426	92114427	92114428
Chryseobacterium gleum	300776841	300776840	300776839
Citrobacter koseri	157148630	157148629	237729986
Clostridium thermocellum	256003460	125974321	125974322
Coccidioides posadasii	303322831		
Congregibacter litoralis	88703327	88703326	88703325
Coprinopsis cinerea	299750141		
Corynebacterium urealyticum	172041452	172041451	172041450
Cryptococcus neoformans var. grubii	23822289		
Cryptococcus neoformans var. neoformans	58270418		
Ectocarpus siliculosus	299469707		

Escherichia coli	187775871	15800665	15800664
Ethanoligenens harbinense	317133580	317133581	317133582
Frankia alni	111220900	111220901	111220902
Geobacillus kaustophilus	56420465	56420466	56420467
Geodermatophilus obscurus	284988814	284988813	284988812
Glomus intraradices	297185890		
Glycine max	351722261		
Glycine max	351724331		
Haemophilus influenzae	260582757	16272484	16272485
Hahella chejuensis	83647213	83647212	83647211
Halalkalicoccus jeotgali	300710018	300710019	300710017
Haloarcula marismortui	55376690	55376689	55376691
Haloferax volcanii	292654327	292654326	292654328
Haloquadratum walsbyi	110669477	110669476	110669478
Helicobacter hepaticus	32265907	32265906	
Helicobacter mustelae	291277506	291276538	
Helicobacter pylori	208434038	15644703	
Hirschia baltica	254295077	254295078	254295080
Klebsiella aerogenes	10835900	3212375	152971982
Ktedonobacter racemifer	298249513	298249512	298249511
Laccaria bicolor	170098735		
Lactobacillus reuteri	194467060	194467061	194467062
Laribacter hongkongensis	226939965	226939964	226939963
Lysinibacillus sphaericus	169828213	169828212	169828211
Methylibium petroleiphilum	124265860	124265861	124265862
Methylocella silvestris	217978168	217978169	217978170
Methylosinus trichosporium	296447068	296447069	296447070
Micromonospora aurantiaca	315504551	302868979	302868980
Microsporum canis	238839950		
Minibacterium massiliensis	152980166	152980572	152980409
Mitsuokella multacida	255658020	260881031	255658018
Morus alba	222143560		
Mycobacterium gilvum	145223938	145223939	145223940
Mycobacterium ulcerans	183982730	118618421	183982728
Nakamurella multipartita	258650533	258650532	258650531
Natronomonas pharaonis	76801650	76801649	76801651
Neosartorya fischeri	119477773		
Neurospora crassa	85116050		
Nitrosospira multiformis	32966209	82702369	82702368
Nostoc punctiforme	186681311	186681310	186681309
Oryza sativa	17402589		
Ostreococcus lucimarinus	145343758		
Paracoccidioides brasiliensis	295673098		
Paracoccus denitrificans	119383953	119383954	119383956
Penicillium chrysogenum	255945993		
Penicillium marneffei	212540166		

Photobacterium profundum	90414537	90414538	90414539
Photorhabdus asymbiotica	253989859	253989860	253989861
Polaromonas naphthalenivorans	121603883	121603884	121603886
Prochlorococcus marinus subsp. pastoris	33861519	33861520	33861521
Pseudoalteromonas haloplanktis	77360699	77360698	77360697
Pseudomonas aeruginosa	15600061	15600060	15600058
Pseudomonas putida	167033934	167033935	170721596
Pyrenophora tritici-repentis	189197501		
Ralstonia eutropha	113867104	113867103	73540695
Rhodococcus opacus	226365149	226365148	226365147
Roseobacter denitrificans	339502215	339502218	339502219
Roseomonas cervicalis	296536422	296536421	296536420
Saccharopolyspora spinosa	348172145	348172144	41350155
Schizosaccharomyces japonicus	213406373		
Schizosaccharomyces pombe	19115725		
Segniliparus rotundus	296394185	296394186	296394187
Solanum tuberosum	14599413		
Sphingobacterium spiritivorum	300772489	227537815	227537814
Staphylococcus aureus	15925280	257423771	15925278
Stappia aggregata	118591406	118591403	118591402
Streptococcus salivarius	2507522	2501628	2501633
Streptomyces albus	291450084	291450085	291450086
Streptosporangium roseum	271964227	271964226	271964224
Thalassiosira pseudonana	224014054		
Thermocrinis albus	289548494	289548493	289548481
Thermomicrobium roseum	221632556	221632557	221632558
Tolumonas auensis	237808376	237808377	237808378
Trichodesmium erythraeum	113474593	113474588	113474587
Tsukamurella paurometabola	296141089	296141090	296141091
Uncinocarpus reesii	258565811		
Ureaplasma urealyticum	171920953	185178985	167972857
Verminephrobacter eiseniae	121608691	121608692	121608694
Verrucomicrobium spinosum	171911815	171911816	171911817
Vibrio splendidus	84387642	84387641	84387640
Yersinia pestis	45442252	108808387	22125138

ALPHA
BETA
GAMMA

Figure S1 (previous page). Alignment of urease sequences employed in this work. Regions marked in red were removed for containing gaps and regions marked in yellow were removed after SimPlot variability analyses. The regions corresponding to subunits γ, β, and α are marked accordingly.

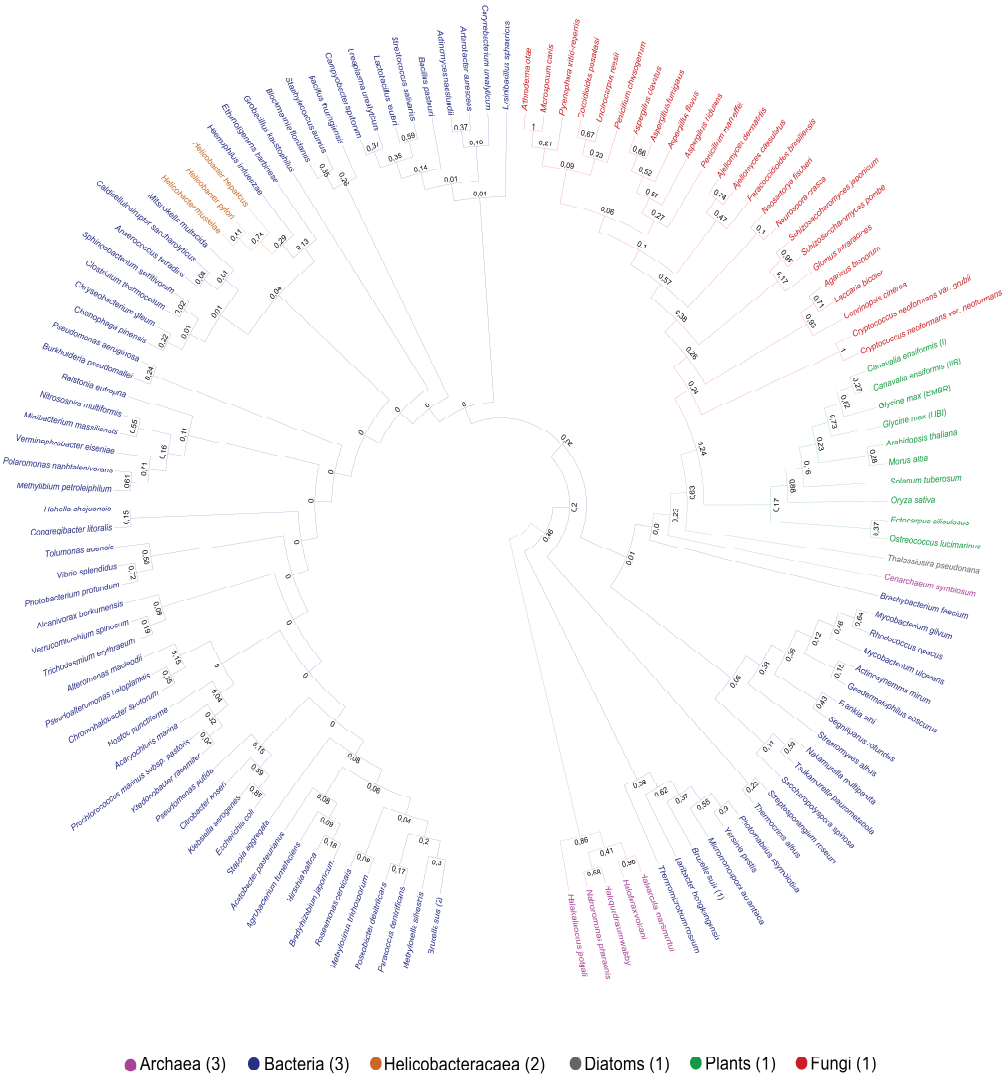


Figure S2. Molecular phylogenetic analysis of highly variable regions of urease sequences by Maximum Likelihood method. The evolutionary history was inferred by using the ML method based on the WAG+G+I model. The number of chains composing ureases from different groups is given in brackets.

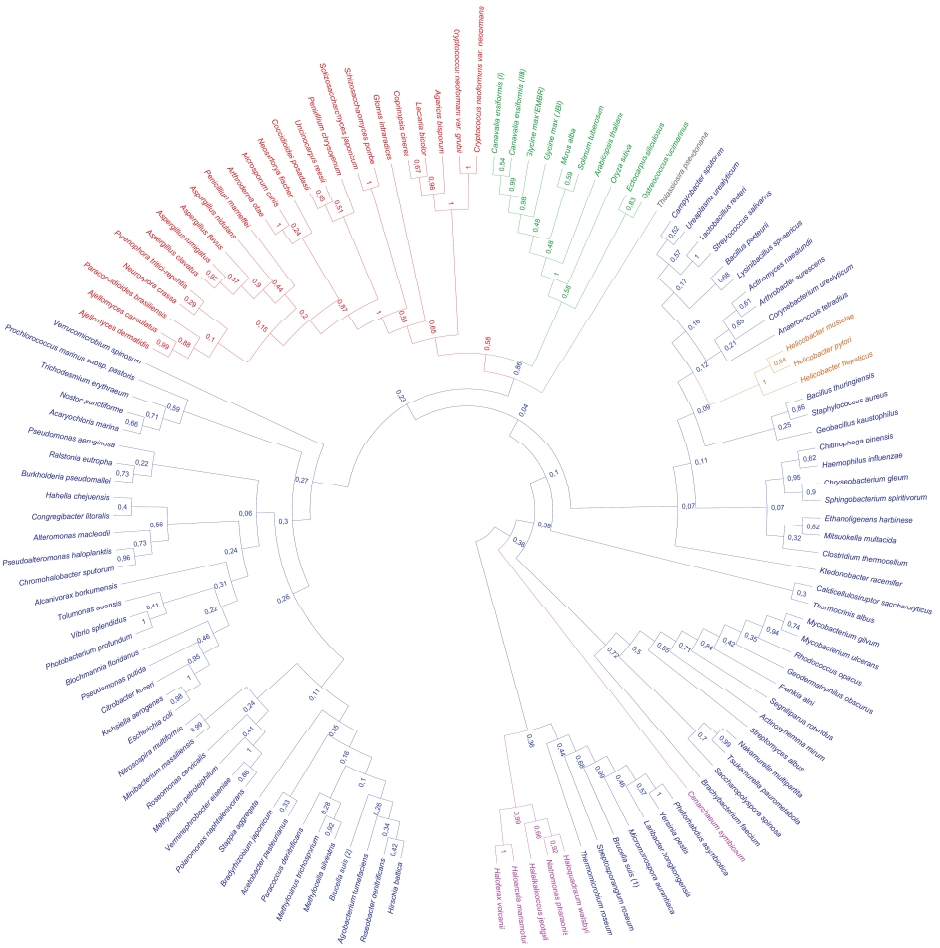


Figure S3. Mathematical midpoint rooting of the complete urease sequences tree by Maximum Likelihood method. The evolutionary history was inferred by using the ML method based on the WAG+G+I model. Mathematical midpoint was calculated with FigTree. For color references, please refer to Figure S2.

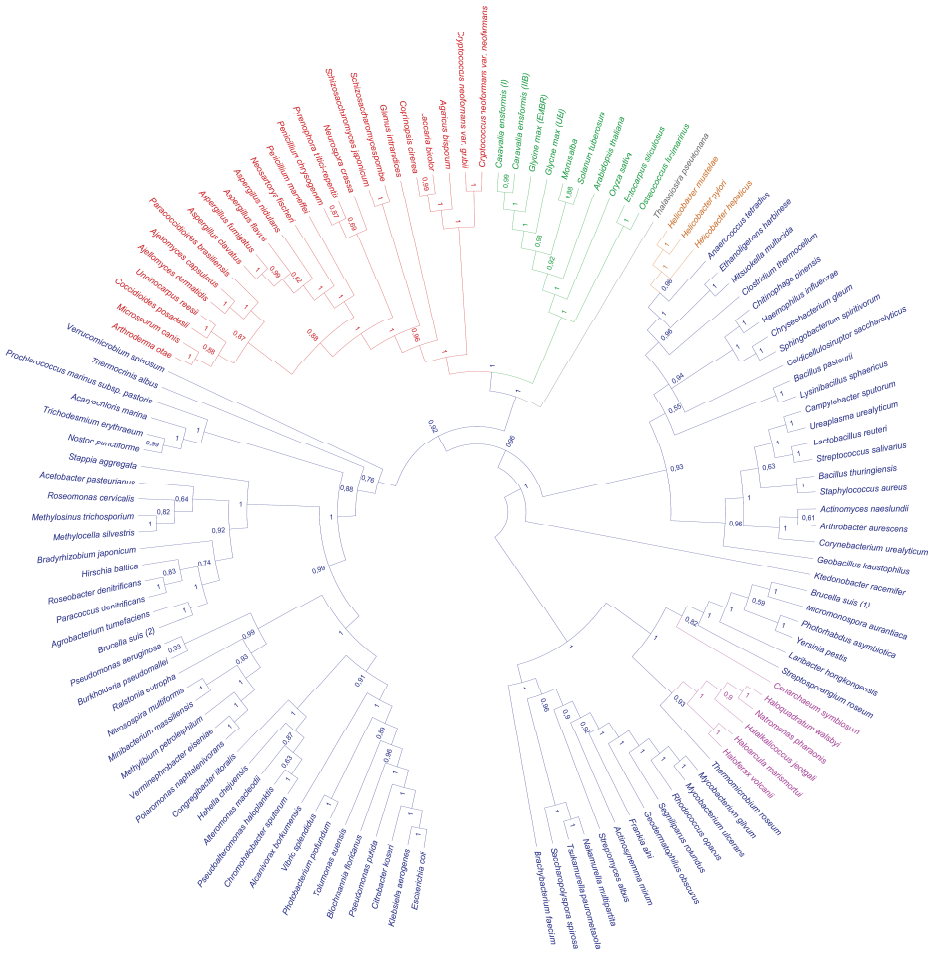


Figure S4. Mathematical midpoint rooting of the complete urease sequences tree by Bayesian Inference method. The evolutionary history was inferred by using the Bayesian method based on the WAG+G+I model. Mathematical midpoint was calculated with FigTree. For color references, please refer to Figure S2.

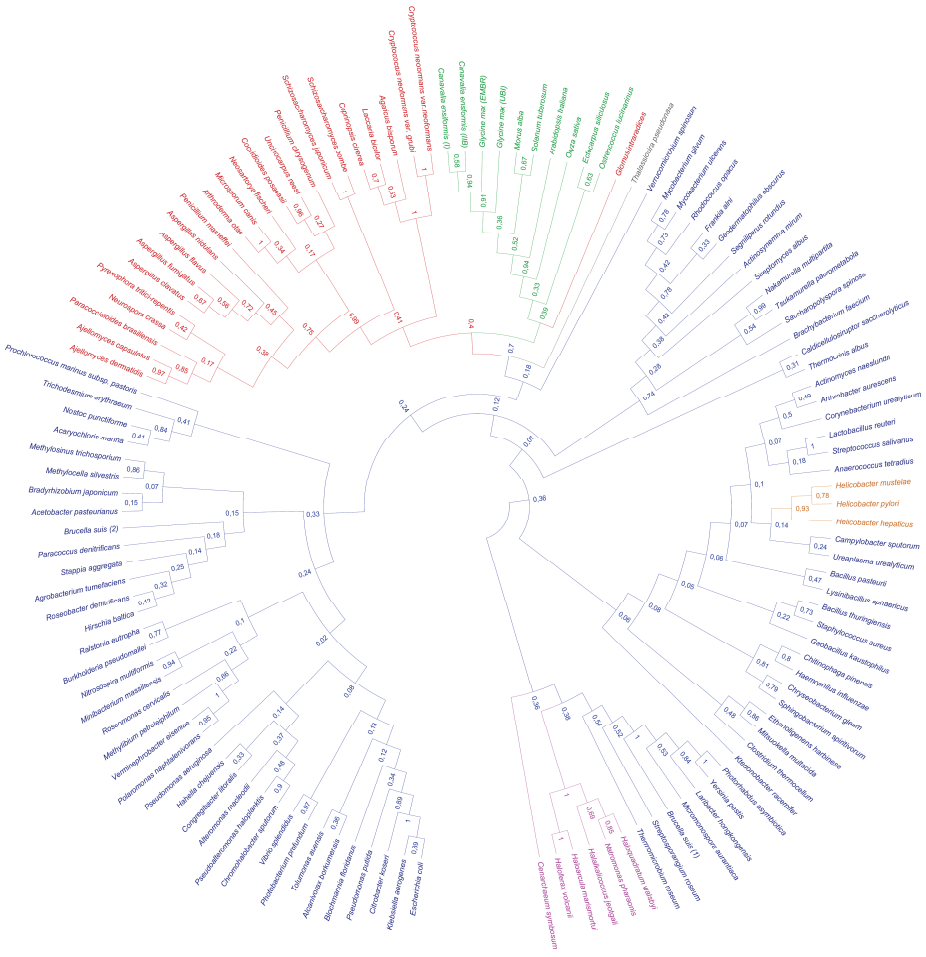


Figure S5. Mathematical midpoint rooting of the urease conserved regions tree by Maximum Likelihood method. The evolutionary history was inferred by using the ML method based on the WAG+G+I model. Mathematical midpoint was calculated with FigTree. For color references, please refer to Figure S2.

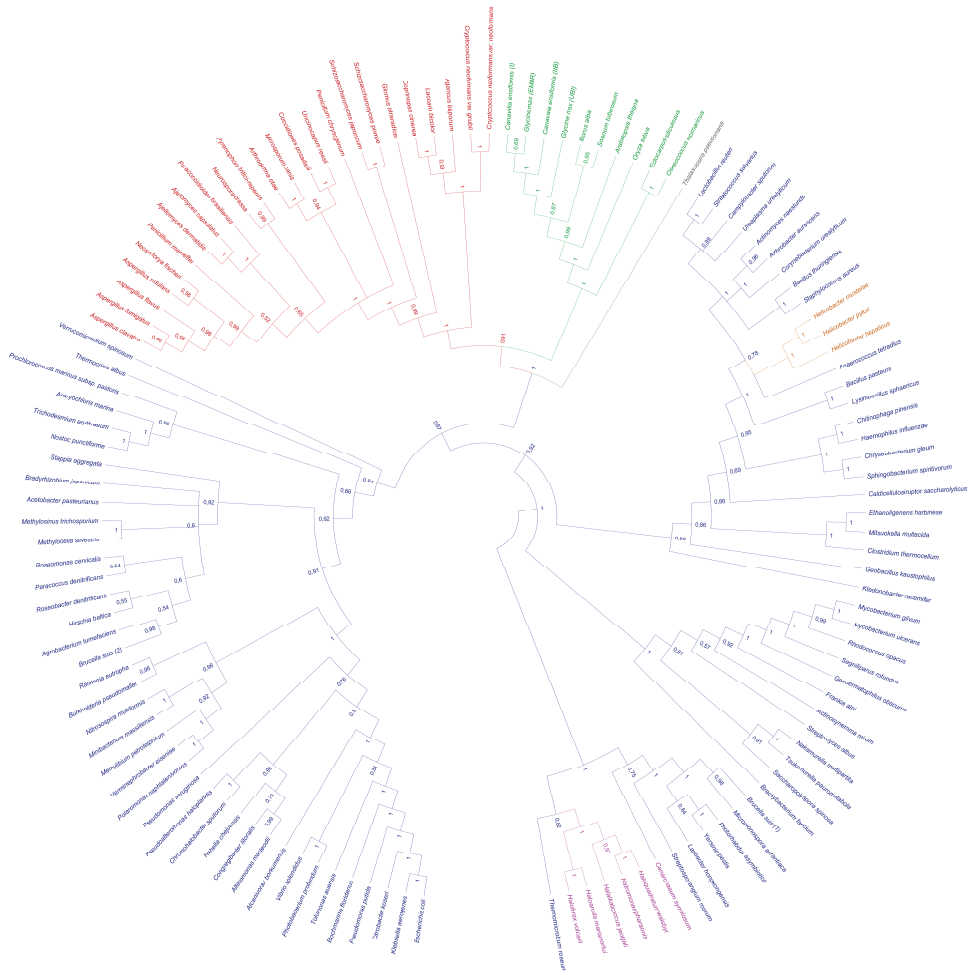


Figure S6. Mathematical midpoint rooting of the urease conserved regions tree by Bayesian Inference method. The evolutionary history was inferred by using the Bayesian method based on the WAG+G+I model. Mathematical midpoint was calculated with FigTree. For color references, please refer to Figure S2.

Journal of Biomolecular Structure and Dynamics, 2013
Vol. 31, No. 8, 854–861, http://dx.doi.org/10.1080/07391102.2012.713782

Taylor & Francis
Taylor & Francis Group

Evidence-based docking of the urease activation complex

Rodrigo Ligabue-Braun[a], Rafael Real-Guerra[a], Célia Regina Carlini[a,b,1] and Hugo Verli[a,c,1]*

[a]*Graduate Program in Cellular and Molecular Biology, Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil;* [b]*Department of Biophysics-IB, UFRGS, Porto Alegre, RS, Brazil;* [c]*Faculty of Pharmacy, UFRGS, Porto Alegre, RS, Brazil*

Communicated by Ramaswamy H. Sarma

(*Received 28 January 2012; final version received 13 July 2012*)

Ureases require accessory proteins for their activation and proper function. In *Klebsiella aerogenes*, UreD, UreF, UreG, and UreE are sequentially complexed to UreABC as required for its activation. Until now, only low-resolution structures are available for this activation complex. To circumvent such limitation, our work intends to provide an atomic-level model for the $(UreABC–UreDFG)_3$ complex from *K. aerogenes*, by employing comparative modeling associated to sequential macromolecular dockings, validated through small-angle X-ray scattering profiles and comparison with results from cross-linking, mutagenesis, and pull-down experiments. Additionally, normal mode analyses of the obtained complex supported the characterization of the elevated flexibility of both UreD–UreF dimer and $(UreABC–UreDFG)_3$ oligomer, explaining the previously observed diffuse binding of UreD to the apoenzyme. The model shown here is the first atomic-level depiction of this complex, a required step for the unraveling of the urease activation process.

An animated Interactive 3D Complement (I3DC) is available in Proteopedia at http://proteopedia.org/w/Journal:JBSD:6

Keywords: oligomer; macromolecular docking; UreD; UreF; UreG; UreABC

Introduction

Ureases (urea amidohydrolases, EC 3.5.1.5) are metalloenzymes that catalyze the hydrolysis of urea to carbon dioxide and ammonia (Krajewska, 2009). The vast majority of ureases have nickel in their active sites, with exceptions having iron or zinc (Carter, Tronrud, Taber, Karplus, & Hausinger, 2011; Follmer et al., 2001). X-ray crystallography studies revealed that, despite having different quaternary structures, plant and bacterial ureases share a common basic trimeric tertiary structure. The active site is composed by two nickel atoms coordinated by four histidine residues, an aspartate, a carbamylated lysine, and four water molecules (Balasubramanian & Ponnuraj, 2010; Benini et al., 1999; Ha et al., 2001; Jabri, Carr, Hausinger, & Karplus, 1995; Pearson, Michel, Hausinger, & Karplus, 1997; Sheridan, Wilmot, Cromie, van der Logt, & Phillips, 2002). Genetic and biochemical studies carried out in prokaryotic and eukaryotic systems have shown that most of these enzymes require accessory proteins for the correct assembly of their metallocenters, as reviewed by Carter, Flugga, Boer, Mulrooney, and Hausinger (2009) and Zambelli, Musiani, Benini, and Ciurli (2011).

The current model for urease metallocenter assembly (Figure 1) derives mostly from studies based on *Escherichia coli* expression of the *Klebsiella aerogenes* urease gene cluster (Mulrooney & Hausinger, 2003). In *K. aerogenes*, genes for the accessory proteins UreD, UreE, UreF, and UreG are grouped with the urease genes UreA, UreB, and UreC in the *ureDABCEFG* cluster. Knockout and complementation studies of each accessory protein revealed that all of them (with the exception of UreE) are essential for the production of a functional urease, being generally considered as urease-specific chaperones (Carter et al., 2009; Lee, Mulrooney, Renner, Markowicz, & Hausinger, 1992; Mulrooney, Ward, & Hausinger, 2005). Little is known about UreD, the first protein to bind the $(UreABC)_3$ oligomer, facilitating activation. UreF, which

*Corresponding author. Email: hverli@cbiot.ufrgs.br
[1]Both authors share senior authorship.

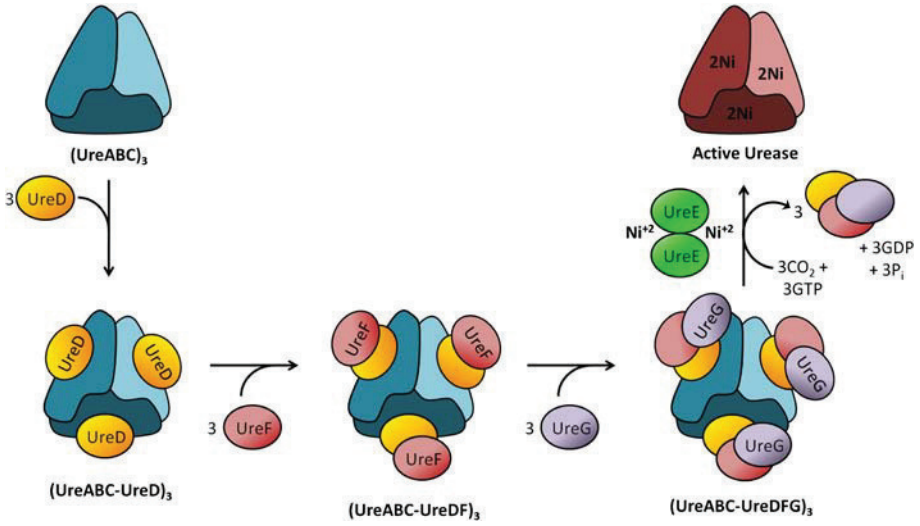


Figure 1. Urease activation pathway in *K. aerogenes* based on Quiroz-Valenzuela et al. (2008). The $(UreABC)_3$ apoprotein is formed by UreA, UreB, and UreC. The trimeric representation considers UreABC as a functional unit. The accessory proteins UreD, UreF, and UreG sequentially bind to form the $(UreABC–UreD)_3$, $(UreABC–UreDF)_3$, and $(UreABC–UreDFG)_3$ activation complexes. The dimeric UreE metallochaperone delivers Ni^{+2} ions to $(UreABC–UreDFG)_3$ (requiring GTP hydrolysis). UreE and $(UreDFG)_3$ are then released from the activated enzyme.

binds $(UreABC–UreD)_3$, may act as a GTPase-activating protein, as it correlates to the UreG GTPase activity when the latter binds $(UreABC–UreDF)_3$. As GTP is hydrolyzed, the nickel-binding chaperone UreE delivers the metal ions to the $(UreABC–UreDFG)_3$ oligomer (Carter et al., 2009; Zambelli et al., 2011).

Despite the apparent wealth of information on these proteins, many aspects of the urease metallocenter assembly remain obscure. One of such aspects is the structural behavior of these proteins when acting together towards urease activation. Major advances were made in this field when different oligomeric stages in the urease activation were studied by small-angle X-ray scattering (SAXS) (Quiroz-Valenzuela, Sukuru, Hausinger, Kuhn, & Heller, 2008). These analyses allowed the description of the *K. aerogenes* oligomers $(UreABC–UreD)_3$, and $(UreABC–UreDF)_3$ at low resolution.

In the present work, we submitted structural models of the urease accessory proteins (based on very recently solved structures) to macromolecular docking calculations, comparing the results with the previous SAXS-derived data. The obtained model includes a putative orientation for UreG at the $(UreABC–UreDFG)_3$ oligomer, for which there is no structural depiction to date. Our results offer a model of the complex of proteins required for urease activation.

Materials and methods

Protein structures

The structure of native *K. aerogenes* urease was taken from PDB ID 1FWJ (Pearson et al., 1997). Its trimeric organization was reconstructed from the crystal symmetry data with PyMol 1.3 (Schrödinger LLC). Comparative molecular modeling was used to obtain *K. aerogenes* UreD, UreF, and UreG full-length structures. UreD (UniProt ID Q09063.1) and UreF (UniProt ID P18318.1) were modeled taking the very recent *Helicobacter pylori* UreH–UreF complex (PDB ID 3SF5) (Fong et al., 2011) as template. UreG (UniProt ID P18319) was modeled based on HypB from *Methanocaldococcus jannaschii* (PDB ID 2HF9) (Gasper, Scrima, & Wittinghofer, 2006), as previously suggested by Carter et al. (2009). All sequence alignments were made with ClustalW (Larkin et al., 2007) (Supplementary Figures 1–3) and the modeling was carried out with MODELLER 9v10 (Sánchez & Šali, 2000). Twenty models were built for each modeling case. These models were stereochemically evaluated and theoretically validated for their three-dimensional profiles with PROCHECK (Laskowski, Macarthur, Moss, & Thornton, 1993) and Verify3D (Lüthy, Bowie, & Eisenberg, 1992), respectively. The best scored model for each protein was then selected. Despite uncertainties regarding the oligomeric state of some of the accessory proteins (Carter et al., 2009; Zambelli et al., 2011), their monomeric form was considered in the present work, as based on SAXS results (Quiroz-Valenzuela et al., 2008).

Molecular docking

The urease activation complex was built in a stepwise approach, as proposed by Quiroz-Valenzuela et al. (2008). The first step was the docking of UreD–$(UreABC)_3$. The resulting $(UreABC–UreD)_3$ structure was then docked

with UreF, forming the $(UreABC–UreDF)_3$ structure that was further docked with $UreG_3$ to give $(UreABC–UreDFG)_3$. Additionally, the UreD–UreF dimer, obtained by superposition of the UreD and UreF models to the crystallographic data from Fong et al. (2011), was also docked to $(UreABC)_3$ for comparison. Each accessory protein binding pose was replicated to produce its threefold binding to $(UreABC)_3$. With the exception of the last stage (binding of UreG), all docking calculations were performed without restrictions, that is, each accessory protein was free to search the entire oligomer surface for its preferential binding site. The most likely binding sites of UreG in the $(UreABC–UreDF)_3$ structure were taken from the work by Boer and Hausinger (2012) and used to restrict the resulting docked poses. Every docking stage was performed by three independent macromolecular docking programs: PatchDock (Schneidman-Duhovny, Inbar, Nussinov, & Wolfson, 2005), Hex (Macindoe, Mavridis, Venkatraman, Devignes, & Ritchie, 2010), and PIPER (Kozakov, Brenke, Comeau, & Vajda, 2006) via ClusPro 2.0 (Comeau, Gatchell, Vajda, & Camacho, 2004).

Comparison with SAXS results

Structural data from experimental SAXS profiles can be compared to atomic structures, for which a theoretical SAXS profile must be calculated. In the present work, this process was carried out in the FoXS server (Schneidman-Duhovny, Hammel, & Šali, 2010). The experimental SAXS profiles for the $(UreABC–UreD)_3$ and $(UreABC–UreDF)_3$ complexes were taken from the work by Quiroz-Valenzuela et al. (2008).

Relative energy evaluation

Considering the lack of SAXS data to validate the binding poses of UreG to the $(UreABC–UreDF)_3$ models, the docking solutions were clustered with MMTSB Tool Set (Feig, Karanicolas, & Brooks, 2004) by hierarchical clustering based on mutual RMSD and evaluated in terms of relative energy with FoldX (Guerois, Nielsen, & Serrano, 2002). This approach was chosen instead of selecting the best scoring solution, since near native, low-energy conformations of proteins tend to cluster together and larger clusters may point to the real docking solution, as reviewed by Ritchie (2008). Binding poses close to the native orientation are therefore expected to present lower energy terms. Since the obtained models were derived from rigid docking, refinement of the interface interactions in the complex were carried out with dedicated tools from FoldX prior to energy evaluation.

Normal mode analysis

To inspect potential functional motions of the obtained $(UreABC–UreDFG)_3$ structure, the urease apoprotein and oligomer were subjected to normal mode analyses (NMA) employing the ElNémo server (Suhre & Sanejouand, 2004).

Results and discussion

The models of UreD, UreF, and UreG were obtained from comparative modeling and further validated (sequence alignments and similarity information are presented in the Figures 1–3). UreH from *H. pylori* was chosen as template for UreD from *K. aerogenes* considering that the novel fold presented by UreH may likely be shared by UreD (Fong et al., 2011). Previous attempts of comparative modeling of this protein yielded only partial structures (Musiani, Bellucci, & Ciurli, 2011), thus supporting a model based on the most similar structure to date, that is, UreH from *H. pylori*.

Three docking softwares were employed in this work, as an attempt to sample diverse docking solutions (considering the different docking strategies from each software) and reach convergence among results from different softwares. The models were clustered first by the docking programs and again by the MMTSB tools. For the docking of $(UreABC–UreD)_3$ and $(UreABC–UreDF)_3$, the main filter for model evaluation was agreement with SAXS profiles, with all results being evaluated despite cluster ranking. For each docking step, 90 models were evaluated (the top 30 results from each software). Clustered results from each software were then clustered again with those from the other softwares. Three clusters were obtained, with the chosen cluster corresponding to approximately 45% of the analyzed structures. There were no other clusters with similar fit to SAXS profile. In addition, docking restricted by experimental SAXS profiles was attempted using FoXSDock (Schneidman-Duhovny, Hammel, & Šali, 2011). This approach resulted in a worse fit of theoretical and experimental curves when compared to the convergent models derived from sequential rigid docking (Supplementary Figure 4), which may be related to the symmetric organization of the complex.

The obtained $(UreABC–UreD)_3$ and $(UreABC–UreDF)_3$ complexes were initially validated by comparison to experimental SAXS profiles, and further reinforced by comparison to cross-linking and mutagenesis data (see below). The $(UreABC–UreD)_3$ model that agrees best with SAXS data had its theoretical curve fit to the experimental data with $\chi = .59$ (Figure 2(A)), while the best $(UreABC–UreDF)_3$ model had $\chi = .44$ (Figure 2 (B)). Alternatively, the UreD–UreF dimer was docked to the $(UreABC)_3$ structure, giving a similar result ($\chi = .47$) when compared to the $(UreABC–UreDF)_3$ model obtained from docking of the separated accessory proteins (Figure 2(C)). This small difference between the docking results of separate or bound UreD–UreF may reflect the intrinsic flexibility of these accessory proteins

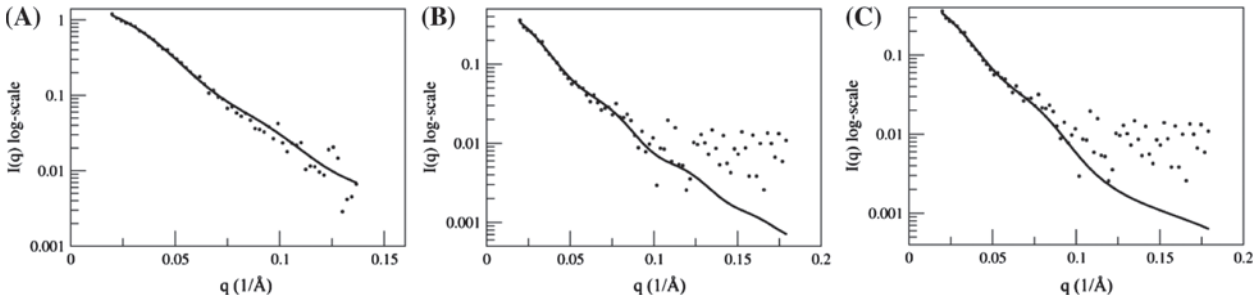


Figure 2. Comparison of theoretical (lines) and experimental (dots) SAXS profiles. Best curve fits for (A) the $(UreABC–UreD)_3$ model, (B) the $(UreABC–UreDF)_3$ model, and (C) $(UreABC–UreDF)_3$ derived from UreDF dimer docking.

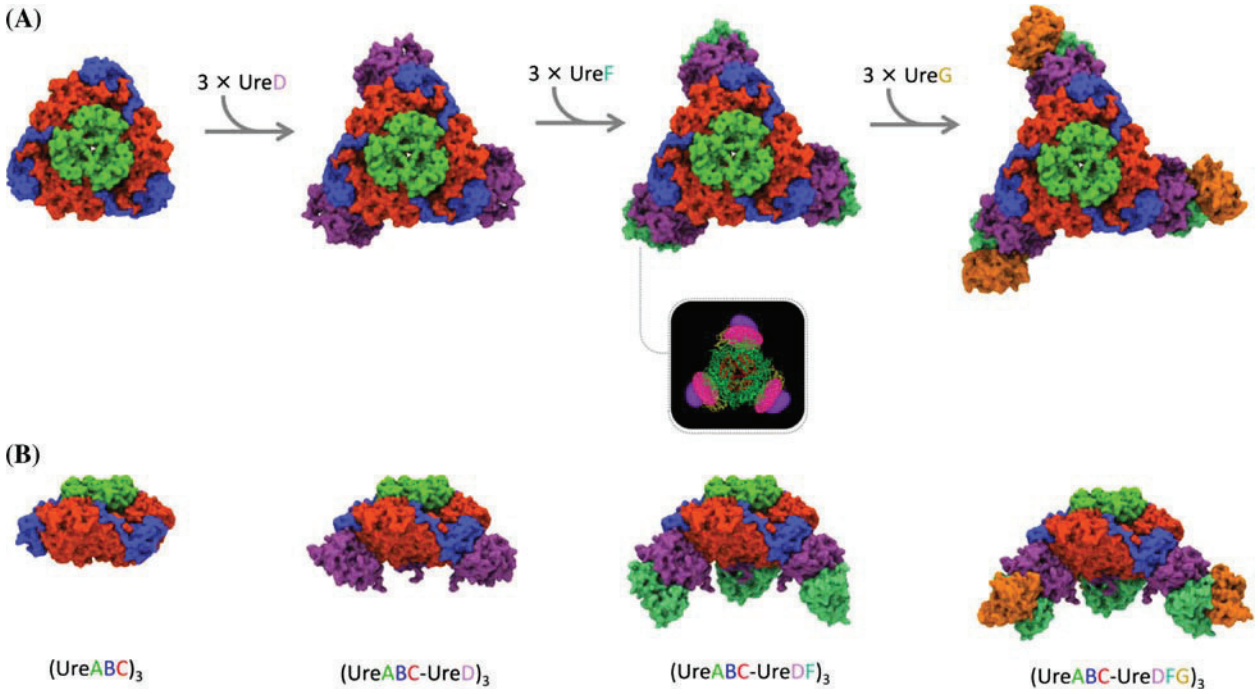


Figure 3. Putative model for the $(UreABC–UreDFG)_3$ complex in the urease activation pathway. The proteins are colored by chain. The structures in (B) are rotated by 90° in respect to those in (A). The inset represents the SAXS-derived model for $(UreABC–UreDF)_3$ from Quiroz-Valenzuela et al. (2008) reproduced with kind permission from Elsevier.

when forming a dimer, as observed from NMA (Supplementary Figure 5). Nonetheless, the available SAXS resolution for the complex does not allow unambiguous separation between these two results, so the stepwise docking result was chosen for the subsequent dockings based on the Quiroz-Valenzuela et al. (2008) model. The theoretical radii of gyration for the obtained complexes, 47.0 Å for $(UreABC–UreD)_3$ and 55.7 Å for $(UreABC–UreDF)_3$, agree with experimental observations (44.9 ± 0.7 and 53.7 ± 1.4 Å, respectively) (Quiroz-Valenzuela et al., 2008). For comparison, the theoretical radius of gyration of *K. aerogenes* crystallographic $(UreABC)_3$ as calculated in this work is 37.5 Å, while its SAXS-derived radius of gyration is 35.7 ± 0.8 Å (Quiroz-Valenzuela et al., 2008).

UreG was then docked to the $(UreABC–UreDF)_3$ models in better agreement with SAXS, derived from both the sequential docking of UreD and UreF, and the crystallography-based UreDF dimer. The preferable binding sites for UreG in the $(UreABC–UreDF)_3$ were derived from mutagenesis studies of the *K. aerogenes* activation complex (Boer & Hausinger, 2012), involving UreF mutated at residues Pro19, Tyr23, Glu30, Glu94, His214, Arg220, Leu221, Phe222, and Ser224 from UreF. Clustering of results at this stage revealed a single cluster with two sub-clusters, indicating high similarity

of the results. As shown by FoldX, the most prevalent pose of UreG interacted with (UreABC–UreDF)$_3$ with an increased strength of two order magnitude than the least prevalent one, thus it was selected as the most likely orientation for this protein on the putative model of the activation complex.

From the position of UreD in the docked complex (Figure 3 and Supplementary PDB file), the hinge region of UreB (residues 1–19) (Figure 4(A)), which is essential for proper urease activation (Carter et al., 2011), could not act as a binding site for accessory proteins. In fact, the obtained model suggests that this region could be required for proper 'gating' of the active site for activation, its absence leading to improper binding of UreB to the UreAC complex, since deletion of the hinge region abolishes urease activity (Carter et al., 2011). Likewise, the model indicates that the region itself does not bind directly to any accessory protein.

The putative complex also offers insights into the observed inability of UreB lacking the hinge region to bind soluble UreD fused to maltose binding protein (Boer, Quiroz-Valenzuela, Anderson, & Hausinger, 2010). Accordingly, UreD binds both UreB and UreC in the proposed structure, forming five hydrogen bonds and 143 nonbonded contacts with the former, and six hydrogen bonds and 81 nonbonded contacts with the latter. Thus, UreD may be unable to bind UreB alone, without the UreC complement (connected to UreB via its hinge region).

It is also noteworthy that no significant skew towards putative metal-binding residues is observed at the surface of the accessory proteins. Based on this observation, a clear pathway for nickel traffic towards the active site could not be proposed in the present work.

The putative structure for the activation complex agrees with previous pull-down (Boer et al., 2010) and

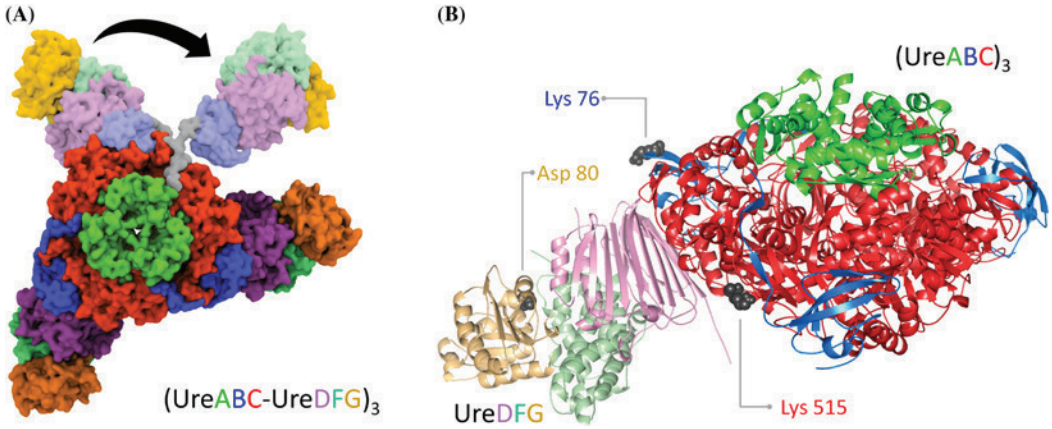


Figure 4. (A) One UreB–UreDFG complex is highlighted to indicate putative UreB repositioning (arrow), according to Quiroz-Valenzuela et al. (2008), the hinge region of the highlighted UreB is colored in grey and (B) Agreement with cross-linking observations. Proximity of Lys76 of UreB and Lys515 of UreC to UreD (in the UreDFG oligomer) and location of Asp 80 of UreG in the UreDFG interface. For clarity, only one of three UreDFG oligomers is shown.

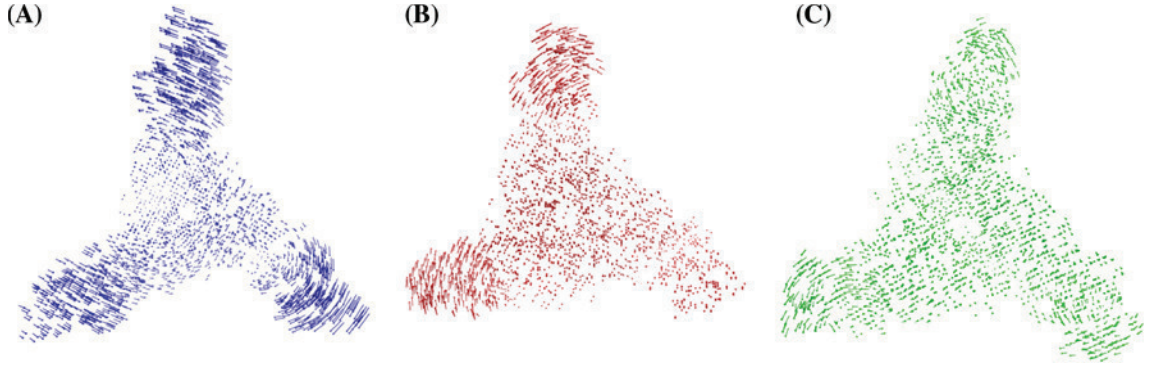


Figure 5. Intrinsic flexibility analyses for (UreABC–UreDFG)$_3$ complex. Three major movements are depicted (color-headed arrows). (A) 'pinching,' (B) 'twists,' and (C) 'torsions' (different torsion intensities are shown). The activation complex is depicted as looking from below, with the accessory proteins closer to the viewer.

cross-linking (Chang, Kuchar, & Hausinger, 2004) assays. The residue Asp80 from UreG, essential for stabilizing the interaction of this protein with (UreABC–UreDF) (Boer et al., 2010), is positioned at the interface between UreG and the UreABC–UreDF complex (Figure 4(B)) as expected. Similarly, the modeled complex has UreD close to Lys76 of UreB (9.4 Å) and Lys515 of UreC (11.5 Å) (Figure 4(B)), as previously proposed by studies employing different cross-linkers (Chang et al., 2004), such as DMS, DMA, and BS^3. Taking into account the length of these cross-linkers (up to 13 Å for BS^3), the observed distances are within an acceptable range to be considered 'close'.

The final $(UreABC–UreDFG)_3$ structure shows significant flexibility. While the urease apoenzyme $(UreABC)_3$ forms a less-flexible core, the accessory (UreDFG) proteins form dynamic 'flaps' as indicated by the conformational diversity observed for this oligomer when subjected to NMA. Besides subtle movements, these analyses reveal three classes of conformational changes, namely, 'pinching,' 'torsions,' and 'twists' (Figure 5). These movements, mainly brought about by the accessory proteins, might indicate a role of $(UreDFG)_3$ in performing, or assisting, in the displacement of UreB (Quiroz-Valenzuela et al., 2008) for proper nickel incorporation and urease activation. Considering that SAXS profiles are obtained from samples containing an ensemble of structures in solution, this flexibility may explain the deviation of the experimental SAXS curve when compared to the static $(UreABC–UreDF)_3$ structure derived from docking. Additionally, such flexibility may also be related to functional roles, since the UreDF dimer has been considered a possible intermediate in urease activation (Carter et al., 2009).

The recent solving of UreH–UreF 'dimer of heterodimers' from *H. pylori* (Fong et al., 2011) certainly is a milestone on the unraveling of the urease activation complex. However, the quaternary disposition of ureases from the *Helicobacter* genus is significantly different from those of other bacteria. While *H. pylori* urease adopts a sphere-like dodecameric form, *K. aerogenes*, *Bacillus pasteurii*, and possibly many other bacterial ureases adopt a pyramidal trimeric organization (Benini et al., 1999; Ha et al., 2001; Jabri et al., 1995; Krajewska, 2009; Pearson et al., 1997). Furthermore, despite being a UreD ortholog, UreH is very different from this protein sharing only 17% identity. Therefore, *H. pylori* is more of an exception than a general case when it comes to ureases and such difference is reflected in the distinct structural organization observed for its activation complex. The *K. aerogenes* $(UreABC–UreDFG)_3$ may thus serve as a general structural model for bacterial urease activation (not being applicable to *Helicobacter* spp.). It is also worth mentioning that plants and bacteria seem to share most of the activation mechanism, and that plant ureases can form symmetric hexamers by combining two trimers (as those found in bacteria) (Witte, 2011). It is thus tempting to hypothesize that plant trimeric ureases may be activated prior to the hexamer formation, since their symmetry would hinder binding of accessory proteins as observed for *K. aerogenes*.

In conclusion, the current work presents a putative model for the urease activation complex of *K. aerogenes*. It is not only the first 3D depiction of this oligomer, but also the first structural model to include UreG. Despite of the urease activation process being far more complex, including the additional UreE, nickel transference, and other poorly understood phenomena, our results are likely to expand the current knowledge on this essential step for proper ureolytic activity, aiding further high resolution studies of this macromolecular assembly.

Acknowledgments

We are grateful to Robert P. Hausinger for sharing the experimental SAXS profiles; Kam-Bo Wong for sharing the UreH–UreF structure; and Stefano Ciurli and Francesco Musiani for insightful discussions. This work was supported by the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), MCT; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), MEC; Financiadora de Estudos e Projetos (FINEP); and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.[10.1080/07391102.2012.713782].

References

Balasubramanian, A., & Ponnuraj, K. (2010). Crystal structure of the first plant urease from jack bean: 83 years of journey from its first crystal to molecular structure. *Journal of Molecular Biology, 400*, 274–283. doi: 10.1016/j.jmb.2010.05.009

Benini, S., Rypniewski, W.R., Wilson, K.S., Miletti, S., Ciurli, S., & Mangani, S. (1999). A new proposal for urease mechanism based on the crystal structures of the native and inhibited enzyme from *Bacillus pasteurii*: Why urea hydrolysis costs two nickels. *Structure, 7*, 205–216. doi: 10.1016/S0969-2126(99)80026-4

Boer, J.L., & Hausinger, R.P. (2012). *Klebsiella aerogenes* UreF: Identification of the UreG binding site and role in enhancing the fidelity of urease activation. *Biochemistry, 51*, 2298–2308. doi: 10.1021/bi3000897

Boer, J.L., Quiroz-Valenzuela, S., Anderson, K.L., & Hausinger, R.P. (2010). Mutagenesis of *Klebsiella aerogenes* UreG to probe nickel binding and interactions with other urease-related proteins. *Biochemistry, 49*, 5859–5869. doi: 10.1021/bi1004987

Carter, E.L., Boer, J.L., Farrugia, M.A., Flugga, N., Towns, C. L., & Hausinger, R.P. (2011). Function of UreB in *Klebsiella aerogenes* urease. *Biochemistry, 50*, 9296–9308. doi: 10.1021/bi2011064

Carter, E.L., Flugga, N., Boer, J.L., Mulrooney, S.B., & Hausinger, R.P. (2009). Interplay of metal ions and urease. *Metallomics, 1*, 207–221. doi: 10.1039/B903311D

Carter, E.L., Tronrud, D.E., Taber, S.R., Karplus, P.A., & Hausinger, R.P. (2011). Iron-containing urease in a pathogenic bacterium. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 13095–13099. doi: 10.1073/pnas.1106915108

Chang, Z., Kuchar, J., & Hausinger, R.P. (2004). Chemical cross-linking and mass spectrometric identification of sites of interaction for UreD, UreF, and urease. *The Journal of Biological Chemistry, 279*, 15305–15313. doi: 10.1074/jbc.M312979200

Comeau, S.R., Gatchell, D.W., Vajda, S., & Camacho, C.J. (2004). ClusPro: A fully automated algorithm for protein-protein docking. *Nucleic Acids Research, 32*, W96–W99. doi: 10.1093/nar/gkh354

Feig, M., Karanicolas, J., & Brooks, C.L. III (2004). MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling, 22*, 377–395. doi: 10.1016/j.jmgm.2003.12.005

Follmer, C., Barcellos, G.B., Zingali, R.B., Machado, O.L., Alves, E.W., Barja-Fidalgo, C., … Carlini, C.R. (2001). Canatoxin, a toxic protein from jack beans (*Canavalia ensiformis*), is a variant form of urease (EC 3.5.1.5): Biological effects of urease independent of its ureolytic activity. *Biochemical Journal, 360*, 217–224. doi: 10.1042/0264-6021:3600217

Fong, Y.H., Wong, H.C., Chuck, C.P., Chen, Y.W., Sun, H., & Wong, K.B. (2011). Assembly of preactivation complex for urease maturation in *Helicobacter pylori*: Crystal structure of UreF-UreH protein complex. *The Journal of Biological Chemistry, 286*, 43241–43249. doi: 10.1074/jbc.M111.296830

Gasper, R., Scrima, A., & Wittinghofer, A. (2006). Structural insights into HypB, a GTP-binding protein that regulates metal binding. *The Journal of Biological Chemistry, 281*, 27492–27502. doi: 10.1074/jbc.M600809200

Guerois, R., Nielsen, J.E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology, 320*, 369–387. doi: 10.1016/S0022-2836(02)00442-4

Ha, N.C., Oh, S.T., Sung, J.Y., Cha, K.A., Lee, M.H., & Oh, B.H. (2001). Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. *Nature Structural & Molecular Biology, 8*, 505–509. doi: 10.1038/88563

Jabri, E., Carr, M.B., Hausinger, R.P., & Karplus, P.A. (1995). The crystal structure of urease from *Klebsiella aerogenes*. *Science, 268*, 998–1004. doi: 10.1126/science.7754395

Kozakov, D., Brenke, R., Comeau, S.R., & Vajda, S. (2006). PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure. Function, and Bioinformatics, 65*, 392–406. doi: 10.1002/prot.21117

Krajewska, B. (2009). Ureases I. Functional, catalytic and kinetic properties: A review. *Journal of Molecular Catalysis B: Enzymatic, 59*, 9–21. doi: 10.1016/j.molcatb.2009.01.003

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., … Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics, 23*, 2947–2948. doi: 10.1093/bioinformatics/btm404

Laskowski, R.A., Macarthur, M.W., Moss, D.S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography, 26*, 283–291. doi: 10.1107/S0021889892009944

Lee, M.H., Mulrooney, S.B., Renner, M.J., Markowicz, Y., & Hausinger, R.P. (1992). *Klebsiella aerogenes* urease gene cluster: Sequence of ureD and demonstration that four accessory genes (ureD, ureE, ureF, and ureG) are involved in nickel metallocenter biosynthesis. *Journal of Bacteriology, 174*, 4324–4330. Retrieved from http://jb.asm.org/content/174/13/4324

Lüthy, R., Bowie, J.U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature, 356*, 83–85. doi: 10.1038/356083a0

Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.-D., & Ritchie, D.W. (2010). HexServer: An FFT-based protein docking server powered by graphics processors. *Nucleic Acids Research, 38*, W445–W449. doi: 10.1093/nar/gkq311

Mulrooney, S.B., & Hausinger, R.P. (2003). Nickel uptake and utilization by microorganisms. *FEMS Microbiology Reviews, 27*, 239–261. doi: 10.1016/S0168-6445(03)00042-1

Mulrooney, S.B., Ward, S.K., & Hausinger, R.P. (2005). Purification and properties of the *Klebsiella aerogenes* UreE metal-binding domain, a functional metallochaperone of urease. *Journal of Bacteriology, 187*, 3581–3585. doi: 10.1128/JB.187.10.3581–3585.2005

Musiani, F., Bellucci, M., & Ciurli, S. (2011). Model structures of *Helicobacter pylori* UreD(H) domains: A putative molecular recognition platform. *Journal of Chemical Information and Modeling, 51*, 1513–1520. doi: 10.1021/ci200183n

Pearson, M.A., Michel, L.O., Hausinger, R.P., & Karplus, P.A. (1997). Structures of Cys319 variants and acetohydroxamate-inhibited *Klebsiella aerogenes* urease. *Biochemistry, 36*, 8164–8172. doi: 10.1021/bi970514j

Quiroz-Valenzuela, S., Sukuru, S.C., Hausinger, R.P., Kuhn, L. A., & Heller, W.T. (2008). The structure of urease activation complexes examined by flexibility analysis, mutagenesis, and small-angle X-ray scattering. *Archives of Biochemistry and Biophysics, 480*, 51–57. doi: 10.1016/j.abb.2008.09.004

Ritchie, D.W. (2008). Recent progress and future directions in protein-protein docking. *Current Protein & Peptide Science, 9*, 1–15. Retrieved from http://www.benthamdirect.org/pages/content.php?CPPS/2008/00000009/00000001/0001K.SGM.

Sánchez, R., & Šali, A. (2000). Comparative protein structure modeling. Introduction and practical examples with Modeller. *Methods in Molecular Biology, 143*, 97–129. doi: 10.1385/1-59259-368-2:97

Schneidman-Duhovny, D., Hammel, M., & Šali, A. (2010). FoXS: A web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Research, 38*, W540–W544. doi: 10.1093/nar/gkq461

Schneidman-Duhovny, D., Hammel, M., & Šali, S. (2011). Macromolecular docking restrained by a small angle X-ray scattering profile. *Journal of Structural Biology, 173*, 461–471. doi: 10.1016/j.jsb.2010.09.023

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H.J. (2005). PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research, 33*, W363–W367. doi: 10.1093/nar/gki481

Sheridan, L., Wilmot, C.M., Cromie, K.D., van der Logt, P., & Phillips, S.E. (2002). Crystallization and preliminary X-ray structure determination of jack bean urease with a bound antibody fragment. *Acta Crystallographica Section D Biological Crystallography, 58*, 374–376. doi: 10.1107/S0907444901021503

Suhre, K., & Sanejouand, Y.H. (2004). ElNemo: A normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Research, 32*, W610–W614. doi: 10.1093/nar/gkh368

Witte, C.-P. (2011). Urea metabolism in plants. *Plant Science, 180*, 431–438. doi: 10.1016/j.plantsci.2010.11.010

Zambelli, B., Musiani, F., Benini, S., & Ciurli, S. (2011). Chemistry of Ni^{2+} in urease: Sensing, trafficking, and catalysis. *Accounts of Chemical Research, 44*, 520–530. doi: 10.1021/ar200041k

```
                    10        20        30        40        50        60
                     |         |         |         |         |         |
HPU_UreH      ------AQESKLRLKTKIGADGRCVIEDNFFTPPFKLMAPFYPKDDLAEIMLLAVSPGMM
KAU_UreD      MLPPLKKGWQATLDLRFHQAGGKTVLASAQHVGPLTVQRPFYPEEETCHLYLLHPPGGIV
                     .          *.*: *: .  .. *:.:  ****::: ..: **  . *::
Prim.cons.    MLPPLK2222222222222A2G22V22222222P22222PFYP22222222LL2222G22

                    70        80        90       100       110       120
                     |         |         |         |         |         |
HPU_UreH      RGDAQDVQLNIGPNCKLRITSQSFEKIHNTEDGFASRDMHIVVGENAFLDFAPFPLIPFE
KAU_UreD      GGDELTISAHLAPGCHTLITMPGASKFYRSSGAQALVRQQLTLAPQATLEWLPQDAIFFP
               **   :. ::.*.*:  **  . .*::.:... *    ::.:. :* *:: *   * *
Prim.cons.    2GD222222222P2C222IT22222K22222222A22222222222A2L222P222I2F2

                   130       140       150       160       170       180
                     |         |         |         |         |         |
HPU_UreH      NAHFKGNTTISLRSSSQLLYSEIIVAGRVARNELFKFNRLHTKISILQDEKPIYYDNTIL
KAU_UreD      GANARLFTTFHLCASSRLLAWDLLCLGRPVIGETFSHGTLSNRLEVWVDNEPLLVERLHL
              .*: :  **: * :**:**  :::  ** . .* *... * .::.:  *::*:  :.  *
Prim.cons.    2A22222TT22L22SS2LL2222222GR2222E2F2222L22222222D22P2222222L

                   190       200       210       220       230       240
                     |         |         |         |         |         |
HPU_UreH      DPKTTDLNNMCMFDG-----YTHYLNLVLVNCPIELSGVRECIEESEGVDGAVSETASSH
KAU_UreD      QEGELSSIAERPWVGTLLCYPATDALLDGVRDALAPLGLYAGASLTDRLLTVRFLSDDNL
              :    .      : *      :    *  *. .:   *:    . :: :  .   : ..
Prim.cons.    22222222222222GTLLCY222222L22V2222222G22222222222222222222222

                   250       260       270
                     |         |         |
HPU_UreH      LCVKALAKGSEPLLHLREKIARLVTQTTTQKV
KAU_UreD      ICQRVMRDVWQFLRPHLTGKSPVLPRIWLT--
              :* :.: .  : *       : ::.:
Prim.cons.    2C2222222222L22222222222222222KV
```

Alignment data :
Alignment length : 272
Identity (*) : 46 is 16.91 %
Strongly similar (:) : 55 is 20.22 %
Weakly similar (.) : 39 is 14.34 %
Different : 132 is 48.53 %
Sequence 0001 : UreHxx0 (261 residues).
Sequence 0002 : UreDxx1 (270 residues).

CLUSTALW options used :
endgaps=1
gapdist=8
gapext=0.2
gapopen=10.0
hgapresidues=GPSNDQERK
matrix=blosum
maxdiv=30
outorder=aligned
pwgapext=0.1
pwgapopen=10.0
pwmatrix=blosum
type=PROTEIN

Supplementary Figure 1. Sequence alignment of *K. aerogenes* UreD (target sequence) and *H. pylori* UreH (template).

```
                    10        20        30        40        50        60
                     |         |         |         |         |         |
HPU_UreF      NAHVDNEFLILQVNDAVFPIGSYTHSFGLETYIQQKKVTNKESALEYLKANLSSQFLYTE
KAU_UreF      MSTAEQRLRLMQLASSNLPVGGYSWSQGLEWAVEAGWVLDVAAFERWQRRQMTEGFFTVD
               : .::.: ::*: .: :*:*.*: * ***  ::   * :  :  .: : :::. *: .:
Prim.cons.    2222222222Q222222P2G2Y22S2GLE2222222V22222222222222222F2222

                    70        80        90       100       110       120
                     |         |         |         |         |         |
HPU_UreF      MLSLKLTYESALQQDLKKILGVEEVIMLSTSPMELRLANQKLGNRFIKTLQAMNELDMGE
KAU_UreF      LPLFARLYRACEQGDIAAAQRWTAYLLACRETRELREEERNRGAAFARLLSDWQPDCPPP
              :  :   *.:. * *:          :: . .. ***  ::: *  * : *.  :
Prim.cons.    2222222Y2222Q2D22222222222222222ELR222222G22F222L2222222222

                   130       140       150       160       170       180
                     |         |         |         |         |         |
HPU_UreF      FFNAYAQKTKDPTHATSYGVFAASLGIELKKALRHYLYAQTSNMVINCVKSVPLSQNDGQ
KAU_UreF      WRSLCQQ--SQ---LAGMAWLGVRWRIALPEMALSLGYSWIESAVMAGVKLVPFGQQAAQ
              : .   *  .:    :. . :..   * * :      *:  .. *:  ** **:.*: .*
Prim.cons.    222222QKT22PTH22222222222I2L22222222Y222222V222VK2VP22Q222Q

                   190       200       210       220
                     |         |         |         |
HPU_UreF      KILLSLQSPFNQLIEKTLELDESHLCTASVQNDIKAMQHESLYSRLYMS
KAU_UreF      QLILRLCDHYAAEMPRALAAPDGDIGSATPLAAIASARHETQYSRLFRS
              :::* * . :   : ::*   :..: :*:    * : :**: ****: *
Prim.cons.    222L2L22222222222L222222222A22221I2222HE22YSRL22S
```

Alignment data :
Alignment length : 229
Identity (*) : 42 is 18.34 %
Strongly similar (:) : 57 is 24.89 %
Weakly similar (.) : 26 is 11.35 %
Different : 104 is 45.41 %
Sequence 0001 : HPU_UreF (229 residues).
Sequence 0002 : KAU_UreF (224 residues).

CLUSTALW options used :
endgaps=1
gapdist=8
gapext=0.2
gapopen=10.0
hgapresidues=GPSNDQERK
matrix=blosum
maxdiv=30
outorder=aligned
pwgapext=0.1
pwgapopen=10.0
pwmatrix=blosum
type=PROTEIN

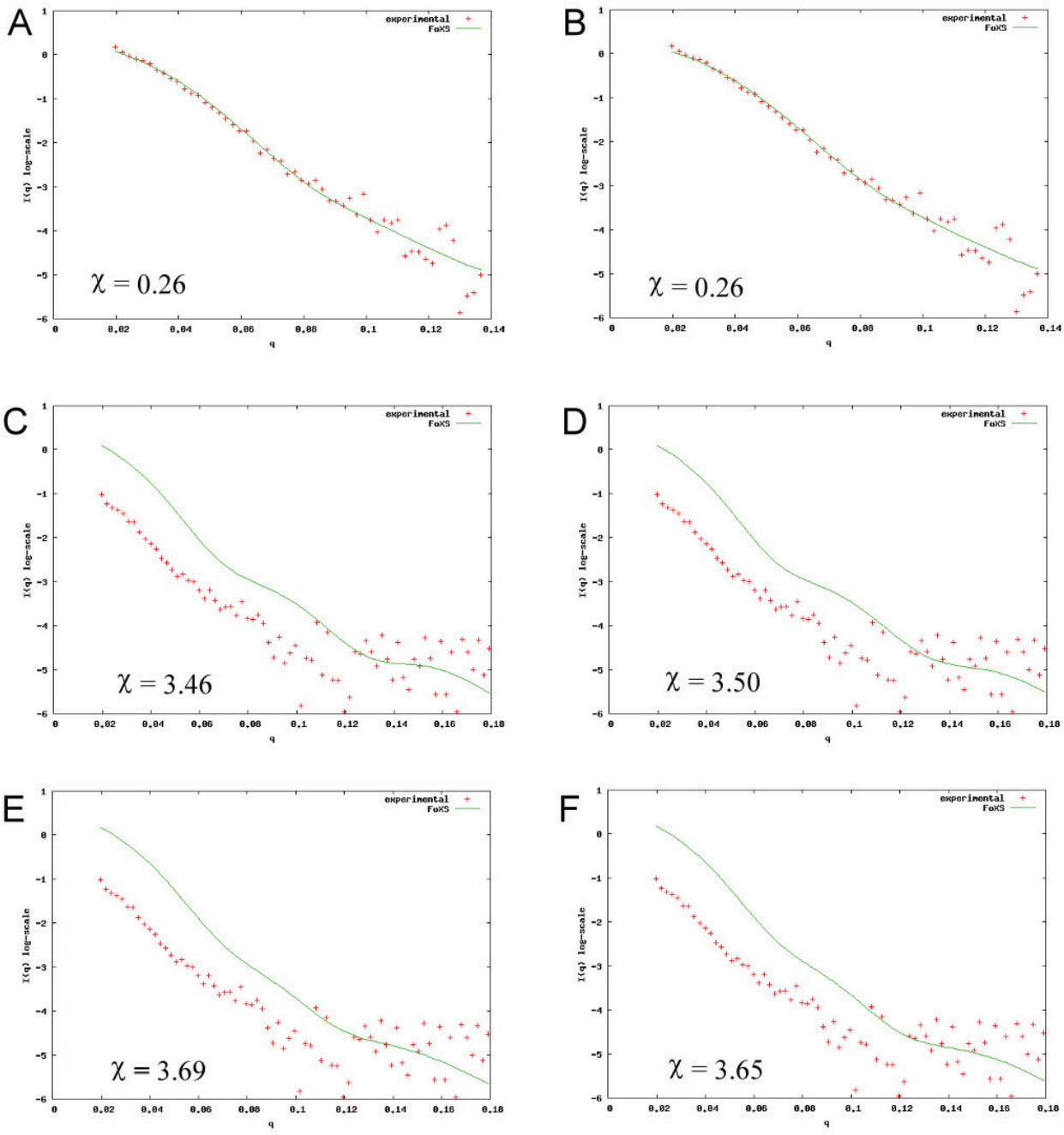
Supplementary Figure 2. Sequence alignment of *K. aerogenes* UreF (target sequence) and *H. pylori* UreF (template).

```
                    10        20        30        40        50        60
                     |         |         |         |         |         |
HypB_2HF9     KDILKANKRLADKNRKLLNKHGVVAFDFMGAIGSGKTLLIEKLIDNLKDKYKIACIAGDV
KAU_UreG      SEQXENCEXREGXREGMNSYKHPLRVGVGGPVGSGKTALLEALCKAMRDTWQLAVVTNDI
              .:  :  :   . .. : . :  : ... *.:***** *:* * . ::*.:::* ::.*:
Prim.cons.    22222222222222222222222222222G22GSGKT2L2E2L22222D2222A2222D2

                    70        80        90       100       110       120
                     |         |         |         |         |         |
HypB_2HF9     IAKFDAERMEKHGAKVVPLNTGKECHLDAHLVGHALEDLN----------LDEIDLLFIE
KAU_UreG      YTKEDQRILTEAGALAPERIVGVETGGCPHTAIREDASMNLAAVEALSEKFGNLDLIFVE
               :* * . : : ** .    .* *    .* . :   .:*          :.::**:*:*
Prim.cons.    22K2D2222222GA2222222G2E22222H222222222NLAAVEALSEK2222DL2F2E

                   130       140       150       160       170       180
                     |         |         |         |         |         |
HypB_2HF9     NVGNLICPADFDLGTHKRIVVISTTEGDDTIEKHPGIMKTADLIVINKIDLADAVGADIK
KAU_UreG      SGGDNLSATFSPELADLTIYVIDVAEGEKIPRKGGPGITKSDFLVINKTDLAPYVGASLE
              . *: :..:     :.  * **..:**:.  .*    :..:*::**** ***  ***.::
Prim.cons.    22G22222222222222I2VI222EG22222K22222222D22VINK2DLA22VGA222

                   190       200       210       220
                     |         |         |         |
HypB_2HF9     KMENDAKRINPDAEVVLLSLKTMEGFDKVLEFIEKSVKEVK
KAU_UreG      VMASDTQRMRGDRPWTFTNLKQGDGLSTIIAFLEDKGMLGK
               * .*::*:. *   .: .**  :*:..:: *:*..    *
Prim.cons.    2M22D22R222D2222222LK222G222222F2E222222K
```

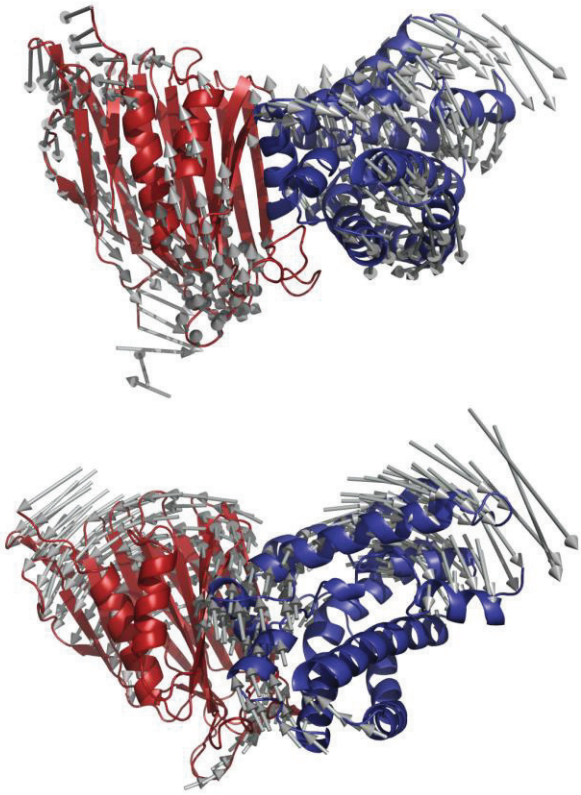
Alignment data :
Alignment length : 221
Identity (*) : 52 is 23.53 %
Strongly similar (:) : 47 is 21.27 %
Weakly similar (.) : 38 is 17.19 %
Different : 84 is 38.01 %
Sequence 0001 : HypB_2HF9 (211 residues).
Sequence 0002 : KAU_UreG (221 residues).

CLUSTALW options used :
endgaps=1
gapdist=8
gapext=0.2
gapopen=10.0
hgapresidues=GPSNDQERK
matrix=blosum
maxdiv=30
outorder=aligned
pwgapext=0.1
pwgapopen=10.0
pwmatrix=blosum
type=PROTEIN

Supplementary Figure 3. Sequence alignment of *K. aerogenes* UreG (target sequence) and HypB from *M. jannaschii* (template).



Supplementary Figure 4. Comparison of theoretical (lines) and experimental (crosses) SAXS profiles from dockings using FoXSDock. Despite the similar fit for $(UreABC\text{-}UreD)_3$ models, the comparison of $(UreABC\text{-}UreDF)_3$ models obtained from iterative accessory-protein docking from FoXDock with those from sequential rigid docking revealed that the latter had geater agreement with experimental data. Depicted are curve fits for (A) the highest-scoring $(UreABC\text{-}UreD)_3$ model, (B) the most energetically favorable $(UreABC\text{-}UreD)_3$ model, (C) highest-scoring $(UreABC\text{-}UreDF)_3$ model derived from the highest-scoring $(UreABC\text{-}UreD)_3$ model, (D) the most energetically favorable $(UreABC\text{-}UreDF)_3$ model derived from the highest-scoring $(UreABC\text{-}UreD)_3$ model, (E) highest-scoring $(UreABC\text{-}UreDF)_3$ model derived from the most energetically favorable $(UreABC\text{-}UreD)_3$ model, and (F) the most energetically favorable $(UreABC\text{-}UreDF)_3$ model derived from the most energetically favorable $(UreABC\text{-}UreD)_3$ model. The units of q are 1/Å.



Supplementary Figure 5. Intrinsic flexibility analyses for UreDF dimer. Two major movements are depicted (grey arrows). UreD is depicted in red and UreF in depicted in blue.

Biochimica et Biophysica Acta 1840 (2014) 935–944

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Structure–function studies on jaburetox, a recombinant insecticidal peptide derived from jack bean (*Canavalia ensiformis*) urease

Anne H.S. Martinelli [a,b], Karine Kappaun [a], Rodrigo Ligabue-Braun [a], Marina S. Defferrari [a], Angela R. Piovesan [a], Fernanda Stanisçuaski [a,c], Diogo R. Demartini [a], Chariston A. Dal Belo [d], Carlos G.M. Almeida [d], Cristian Follmer [e], Hugo Verli [a,c], Celia R. Carlini [a,b,f,1,*], Giancarlo Pasquali [a,c,1,**]

[a] Graduate Program in Cellular and Molecular Biology, Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
[b] Department of Biophysics, Institute of Biosciences, UFRGS, Brazil
[c] Department of Molecular Biology and Biotechnology, UFRGS, Porto Alegre, RS, Brazil
[d] Interdisciplinary Centre of Biotechnological Research, Universidade Federal do Pampa, UNIPAMPA, São Gabriel, RS, Brazil
[e] Department Physical Chemistry, Institute of Chemistry, Universidade Federal do Rio de Janeiro, RJ, Brazil
[f] Instituto do Cérebro (InsCer), Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 12 May 2013
Received in revised form 2 November 2013
Accepted 6 November 2013
Available online 14 November 2013

Keywords:
Site-directed mutagenesis
β-hairpin
Urease-derived peptide
Molecular modeling
Membrane-disturbing
Insect

ABSTRACT

Background: Ureases are metalloenzymes involved in defense mechanisms in plants. The insecticidal activity of *Canavalia ensiformis* (jack bean) ureases relies partially on an internal 10 kDa peptide generated by enzymatic hydrolysis of the protein within susceptible insects. A recombinant version of this peptide, jaburetox, exhibits insecticidal, antifungal and membrane-disruptive properties. Molecular modeling of jaburetox revealed a prominent β-hairpin motif consistent with either neurotoxicity or pore formation.
Methods: Aiming to identify structural motifs involved in its effects, mutated versions of jaburetox were built: 1) a peptide lacking the β-hairpin motif (residues 61–74), JbtxΔ-β; 2) a peptide corresponding the N-terminal half (residues 1–44), Jbtx N-ter, and 3) a peptide corresponding the C-terminal half (residues 45–93), Jbtx C-ter.
Results: 1) JbtxΔ-β disrupts liposomes, and exhibited entomotoxic effects similar to the whole peptide, suggesting that the β-hairpin motif is not a determinant of these biological activities; 2) both Jbtx C-ter and Jbtx N-ter disrupted liposomes, the C-terminal peptide being the most active; and 3) while Jbtx N-ter persisted to be biologically active, Jbtx C-ter was less active when tested on different insect preparations. Molecular modeling and dynamics were applied to the urease-derived peptides to complement the structure–function analysis.
Major conclusions: The N-terminal portion of the Jbtx carries the most important entomotoxic domain which is fully active in the absence of the β-hairpin motif. Although the β-hairpin contributes to some extent, probably by interaction with insect membranes, it is not essential for the entomotoxic properties of Jbtx.
General significance: Jbtx represents a new type of insecticidal and membrane-active peptide.

1. Introduction

Ureases (EC 3.5.1.5, urea amidohydrolase), are nickel dependent enzymes that catalyze urea hydrolysis into ammonia and carbon dioxide. Evolutionarily conserved [1], these proteins have been isolated from a wide variety of organisms including plants, fungi and bacteria. In plants, ureases contribute to the bioavailability of nitrogen and in defense mechanisms [2,3]. Ureases represent an unexplored group of plant proteins with potential use for insect control [3,4] and as antifungal agents [5]. Studies have shown that ureases from *Canavalia ensiformis* (jack bean) and *Glycine max* (soybean) display insecticidal activity (reviewed in [6]) and antifungal properties, inhibiting growth and affecting membrane integrity of filamentous fungi [7] as well as of yeasts [8] in the 10^{-7} M range. The urease from pigeon pea (*Cajanus cajan*) was recently described to exhibit insecticidal and antifungal properties at similar dose ranges [9].

The molecular basis of the insecticidal mechanism of action of plant ureases is not yet completely understood [6]. It has been demonstrated that the entomotoxic effect of canatoxin [10], an isoform of *C. ensiformis* (jack bean) urease [11], is partially due to an internal 10 kDa peptide

Abbreviations: Jbtx, jaburetox; JbtxΔ-β, β-hairpin deleted version of Jbtx; Jbtx N-ter, N-terminal domain of Jbtx; Jbtx C-ter, C-terminal domain of Jbtx; Jbtx-2Ec, a version of Jbtx containing a V5 epitope; LUV, large unilamellar vesicle; MD, molecular dynamics; RMSD, root mean square deviation; CD, circular dichroism

* Correspondence to: C. R. Carlini, Center of Biotechnology and Department of Biophysics, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, Predio 43.431, Porto Alegre, RS, CEP 91501-970, Brazil. Tel.: +55 51 3308 7606.
** Correspondence to: G. Pasquali, Department of Molecular Biology and Biotechnology, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, Predio 43.432, Porto Alegre, RS, CEP 91501-970, Brazil.
E-mail addresses: ccarlini@ufrgs.br, celia.carlini@pq.cnpq.br (C.R. Carlini), pasquali@cbiot.ufrgs.br (G. Pasquali).
[1] These authors share senior authorship.

0304-4165/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.11.010

(pepcanatox), that is released from the protein upon hydrolysis by insect cathepsin-like digestive enzymes [12–16]. Jaburetox-2Ec (Jbtx-2Ec), a recombinant peptide analog to pepcanatox, exhibited a potent insecticidal effect on two economically important crop pests: *Spodoptera frugiperda* (fall armyworm) and *Dysdercus peruvianus* (cotton stainer bug) [17,18]. Jbtx-2Ec was also shown to both permeabilize large unilamellar liposomes (LUVs) [19] and to affect transmembrane potential of insect Malpighian tubules, causing inhibition of diuresis [20]. A β-hairpin motif in the modeled structure of Jbtx-2Ec has been proposed [17,19] and its presence has been confirmed in the crystallographic structures of jack bean [21] and pigeon pea [9] ureases. This motif is present also in one class of pore-forming peptides and neurotoxic peptides [22] such as charybdotoxin, which affect ion channels [23]. A variant form of Jbtx-2Ec lacking the fused V5-antigen, here called simply Jbtx, also exhibited antifungal activity [8].

Aiming to identify motifs possibly involved in the different biological activities of Jbtx, here we described the cloning and expression of mutated versions of the Jbtx-encoding cDNA. Truncated versions of the peptide, with deletions of the regions of the β-hairpin motif, the N-terminal or the C-terminal halves of the molecule, were tested on LUV permeabilization, for insecticidal and other entomotoxic effects. Structural analyses of the truncated peptides were also carried out.

2. Materials and methods

2.1. Jbtx cDNA constructs

Jaburetox-2Ec, the first version of the recombinant urease-derived peptide cloned in [17], harbored a V5-antigen with 18 amino acids derived from the pET101/D-TOPO plasmid. In order to eliminate this foreign sequence, the jack bean urease truncated cDNA encoding 93 amino acids, called simply jaburetox (Jbtx), was cloned and expressed in *Escherichia coli* via pET-23a vector (Novagen), as described in [8]. This sequence was used as template for site-directed mutagenesis and PCR amplifications of the mutant forms as described below.

2.2. Jbtx lacking the internal β-hairpin (JbtxΔ-β)

In order to delete the β-hairpin motif (residues 61–74) of the Jbtx peptide, site-directed mutagenesis was performed using the QuickChange Site-directed Mutagenesis Kit (Stratagene). As this method is often used to generate a few nucleotide deletions, some modifications in the primers' design were made, as described by [24]. Pairs of complementary primers were designed (Table 1), and site-directed mutagenesis was performed according to the kit manufacturer's instructions. The deleted gene version was confirmed by sequencing on an ABI Prism 3100 automated sequencer (Applied Biosystems) platform (ACTGene Ltd, Center of Biotechnology, UFRGS). Sequence comparisons were performed using the BLASTx software [25], available at (http://www.ncbi.nlm.nih.gov). The resulting peptide was called JbtxΔ-β.

Table 1
Primers used in this study.

Primer	Size	Sequence
5′ Del_β-hairpin	40-Mer	AGTATGGTCCGACTATTGGTGAAAAGGATTTTGCCCT TTA
3′ Del_ β-hairpin	40-Mer	TAAAGGGCAAAATCCTTTTCACCAATAGTCGGACCAT ACT
5′ Del_ α-helix	40-Mer	CTTTCACCAAAGCCATTCCTTATGGTCCGACTATTGGTGA
3′ Del_ α-helix	40-Mer	TCACCAATAGTCGGACCATAAGGAATGGCTTTGGTGA AAG
5′ N-terminal	25-Mer	CCAACATATGGGTCCAGTTAAATGA
3′ N-terminal	25-Mer	CCCCCTCGAGGGTGAAAGGACAATC
5′ C-terminal	25-Mer	CCAACATATGAAAGCCATTCCTCGT
3′ C-terminal	25-Mer	CCCCCTCGAGTATAACTTTTCCACC

2.3. Jbtx N-terminal (Jbtx N-ter) and C-terminal (Jbtx C-ter) domain versions

The Jbtx gene regions corresponding to the N-terminal (residues 1–44) and C-terminal (residues 45–93) halves of the peptide were amplified by PCR with specifically designed primers (Table 1) and products were cloned into pET23a (Novagen). PCRs were performed in a final volume of 50 μL containing 50 ng of the template plasmid DNA, 200 ng of each primer, 200 μM each dNTPs, 2.5 U *Pfu* taq DNA polymerase (Fermentas) and 1× *Pfu* reaction buffer. Amplification was carried out under the following conditions: denaturation at 95 °C for 3 min, annealing at 55 °C for 30 s and elongation at 72 °C for 2 min. After a total of 35 cycles, the final products were digested with *Nde*I and *Xho*I (Fermentas), dephosphorylated with thermosensitive alkaline phosphatase (Promega) and ligated into the expression vector pET23a (Novagen). The inserts of the recombinant plasmids were fully sequenced in order to confirm their sequences essentially as described above. The resulting peptides were called Jbtx N-terminal (Jbtx N-ter) and Jbtx C-terminal (Jbtx C-ter). A schematic representation of all Jbtx-related peptides is shown in Fig. 1.

2.4. Expression and purification of Jbtx recombinant peptides

Recombinant pET23a plasmids were transformed into *E. coli* BL21-CodonPlus (DE3)-RIL cells (Stratagene) for Jbtx gene expressions following the provider's instructions. For the purification of the original Jbtx peptide and its mutated forms, 200 mL of Luria Bertani medium containing 100 μg/mL ampicillin and 40 μg/mL chloramphenicol were separately inoculated with 2 mL overnight cultures of each *E. coli* strain. Cells were grown for approximately 2 h at 37 °C under shaking until an optical density of 0.7 was reached. At this point, IPTG was added to cultures to a final concentration of 0.5 mM. After 3 h of additional culture, cells were harvested by centrifugation and suspended in 10 mL of lysis buffer (50 mM Tris buffer, pH 7.5, 500 mM NaCl, 5 mM imidazole), sonicated and centrifuged (14,000 *g*, 30 min). The supernatant was loaded onto a Ni^{+2} loaded Chelating Sepharose (GE Healthcare) column, previously equilibrated with the lysis buffer. After 30 min, the column was washed with 50 mL of the same buffer containing 50 mM imidazole. Bound protein was eluted with 200 mM imidazole in the lysis buffer. Samples were then dialyzed against buffer A (50 mM phosphate buffer, pH 7.5, 1 mM EDTA, 5 mM β-mercaptoethanol) in order to remove the imidazole. Protein concentration was measured by Bradford assay [26]. Predicted molecular mass of the peptides was obtained by submitting the deduced sequences to the ProtScale tool [27] available at the Expasy site (http://web.expasy.org/protscale).

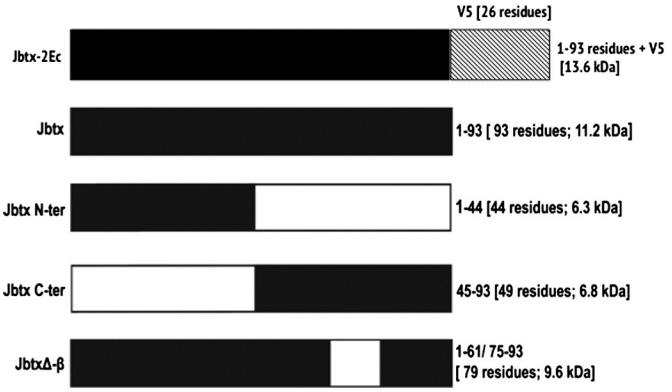


Fig. 1. Schematic representation of the sequences of jaburetox and mutants. The number of amino acid residues of each molecule (shown in black) is indicated on the right side.

Molar concentrations of peptides were calculated assuming a monomeric form in solution. Hydropathicity analysis of the peptides was carried out according to [28].

2.5. Tandem mass spectrometry (MS/MS)

In gel digestion of Jbtx peptides was performed for all samples analyzed, except for the positive control (C+), which was submitted to in solution digestion. For in gel digestion the protocol in [29] was followed, using 50 mM ammonium bicarbonate (AB) unless otherwise explained. Briefly, the bands on the gel were distained with 50 mM AB in 40% acetonitrile (ACN). Following, gel pieces were dehydrated with 100% ACN and lyophilized. Reduction was performed with 50 mM dithiothreitol (DTT) in 50 mM AB, for 30 min at 56 °C, in the dark, followed by the alkylation performed with 50 mM iodoacetamide (IAA) in 50 mM AB for 30 min, at room temperature, also in the dark. The gel pieces were washed with 50 mM AB + 40% ACN for 15 min, and a second dehydration step was done. Proteins in the gel were digested for 3 h (Jbtx N-ter and Jbtx C-ter) or 16 h (Jbtx and JbtxΔ-β) in 500 μL of 100 mM AB solution containing 100 ng of sequencing grade trypsin (Promega) at 37 °C. After digestion, supernatants were transferred to microcentrifuge tubes and gel pieces were washed with 1% formic acid in 60% ACN, and the supernatants were combined accordingly. Digested peptides were lyophilized and submitted to tandem mass spectrometry analyses. In the case of in solution digestion, sample of Jbtx peptide (30 μg in 20 mM sodium phosphate, 5 mM β-mercaptoethanol, 1 mM EDTA, pH 7.5) was reduced with 50 mM DTT and alkylated with 50 mM IAA (30 min each, at room temperature). Then, DTT to a final concentration of 10 mM was added for 15 min at room temperature. Digestion was performed in this case with 0.6 mg of trypsin (Promega) for 3 h at 37 °C. After the digestion process, the samples were desalted with ZipTip™ (Millipore®) according to the manufacturer's instructions. The eluted peptides were lyophilized and analyzed by tandem mass spectrometry.

The lyophilized digested peptides were suspended in 0.1% formic acid (10 μL) and 5 μL of each solution was subjected to reversed phase chromatography (NanoAcquity UltraPerformance LC-UPLC® chromatograph (Waters) using a Nanoease C18, 75 μm ID at 35 °C. The column was equilibrated with 0.1% trifluoroacetic acid (TFA) and the peptides were eluted in a 20 min gradient, ramping from 0 to 60% acetonitrile in 0.1% TFA at 0.6 nL/min constant flow. Eluted peptides were subjected to electrospray ionization and analyzed by mass spectrometry using a Q-TOF Micro™ spectrometer (Micromass). The voltage applied to the cone for the ionization step was 35 V. The three most intense ions in the range of m/z 200–2000 and +2 or +3 charges were selected for fragmentation. The acquired MS/MS spectra were processed using the Proteinlynx v.2.0 software (Waters) and the generated .mgf files were used to perform database searches using the MASCOT software version 2.4.00 (Matrix Science) against the NCBI database, and taxid was restricted to Viridiplantae (taxid:33090). Results were analyzed manually.

2.6. Electrophoresis

The peptide fractions of Jbtx and its mutant forms were visualized in SDS-Tricine gels [30]. The gels were stained with colloidal Coomassie G-250 (Sigma Chem. Co) according to [31].

2.7. Western blot

Western blots were performed according to [32]. Briefly, peptides were electrophoresed, transferred to PVDF membranes (Millipore) and immersed in a blocking buffer consisting of 5% nonfat dry milk in phosphate-buffered saline (PBS, 137 mM NaCl, 2.7 mM KCl and 4.3 mM $Na_2HPO_4 \cdot 7H_2O$, pH 7.3). After washing, the membrane was incubated with rabbit anti-jaburetox-2Ec polyclonal antibodies (1:7500 dilution) for 2 h at room temperature, followed by a 2 h incubation with anti-rabbit IgG (1:20,000 dilution) alkaline phosphatase conjugate (Sigma Chem. Co.). Colorimetric detection was carried out using 5-bromo-4-chloro-3-indolyl-phosphate p-toluidine salt and nitro-blue tetrazolium chloride.

2.8. Leakage experiment

Large unilamellar vesicles (LUVs) were produced and the leakage experiment was conducted as described previously [19]. LUVs were prepared using 10 mg of L-α-phosphatidic acid (egg chicken, Avant Polar Lipids), at a concentration of 20 mg/mL. The leakage promoted by Jbtx and its mutated forms at a final concentration of 5 μg/mL in 25 mM Tris, pH 7.0, was evaluated by the carboxyfluorescein release assay [19]. The concentration of LUVs in the experiment was estimated based on the absorbance of the fluorescent probe at 490 nm, and adjusted to a value of 0.1. In the leakage assays, fluorescence intensity of the reaction mixture (LUVs plus peptide or buffer) was recorded as a function of time. The samples were excited at 490 nm and the fluorescence was acquired at 518 nm. It was assumed that the absence of leakage (0%) corresponded to the fluorescence of the vesicles at time zero; 100% leakage was taken as the value of fluorescence intensity obtained after the addition of 1% (v/v) Triton X-100. All measurements were carried out in a Cary Eclipse fluorescence spectrophotometer (Varian).

2.9. Insecticidal activity

Fifth-instar *Rhodnius prolixus* were kindly provided by Dr. Hatisaburo Masuda and Dr. Pedro L. Oliveira (Institute of Medical Biochemistry, Universidade Federal do Rio de Janeiro, RJ, Brazil) and by Dr. Denise Feder (Universidade Federal Fluminense, RJ, Brazil). The phytophagous milkweed bugs (*Oncopeltus fasciatus*) were reared in our laboratory as previously described [15].

2.9.1. Injection assays

Fifth instars of *O. fasciatus* or *R. prolixus* were injected into the hemocoel using a Hamilton Microliter 900 series syringe (Hamilton). Group of 10 insects (*O. fasciatus*) or 5 insects (*R. prolixus*) were injected with 20 mM sodium phosphate buffer (pH 7.5) containing peptides at a final dose of 0.015 μg (*O. fasciatus*) or 0.05 μg (*R. prolixus*) per mg of insect body weight. Control insects received injections of buffer alone. Mortality rate within each group was recorded after 48 or 96 h. Two independent bioassays were carried out for each peptide on each insect model. Results shown are means ± standard errors.

2.9.2. Feeding assays

Fifth instars *R. prolixus* were fed on *R. prolixus* saline solution (150 mM NaCl, 8.6 mM KCl, 2.0 mM $CaCl_2$, 8.5 mM $MgCl_2$, 4.0 mM $NaHCO_3$, 34.0 mM glucose, 5.0 mM HEPES, pH 7.0) containing 1 mM ATP and enough peptide (tested individually) to give final doses of 0.1 μg per mg of body weight. Groups of 5 insects for each peptide were fed for approximately 30 min, at 37 °C, by placing their mouth apparatus inside glass capillaries containing the test solutions. Control insects fed solely on *R. prolixus* saline solution containing 1 mM ATP, under the same conditions. Mortality rate within each group was recorded after 24 h. One triplicated bioassay was carried out for each peptide. The results shown are means and standard errors.

2.10. Measurement of fluid secretion by Rhodnius prolixus Malpighian tubules

The assay was performed essentially as described in [20], using *R. prolixus* serotonin-stimulated Malpighian tubules. Secretion rate was expressed as the percentage of fluid secretion measured after the addition of Jbtx or mutated peptides as compared to serotonin (2.5×10^{-8} M) alone (control). For each peptide and dose, 5–6 replicates were done. The results shown are means ± standard error.

2.11. In vivo cockroach metathoracic coxal-adductor nerve–muscle preparation

The *in vivo* cockroach metathoracic coxal-adductor muscle preparation was used [33] to characterize further the entomotoxic activity of Jbtx and its mutated versions. Male adult *Phoetalia pallida* (3–4 months after molting) were reared in our laboratory at controlled temperature (22–25 °C) on a 12 h:12 h light:dark cycle. Animals were immobilized by chilling and mounted, ventral side up, in a Lucite holder covered with 1 cm soft rubber that restrained the body and provided a platform to which the metathoracic coxae could be firmly attached using entomologic needles. The left leg was then tied at the medial joint with a dentistry suture line connected to a 1 g force transducer (AVS Instruments, São Carlos, SP, Brazil). The transducer was mounted in a micromanipulator to allow adjustment of muscle length. The exoskeleton was removed from over the appropriated thoracic ganglion. Nerve 5, which includes the motor axon to the muscle, was exposed and a bipolar electrode was inserted to provide electrical stimulation. The nerve was covered with mineral oil to prevent dryness and stimulated at 0.5 Hz, 5 ms, with twice the threshold, during 120 min. Twitch tension was digitalized, recorded and retrieved using a computer based software AQCAD (AVS Instruments, São Carlos, SP, Brazil). Data were further analyzed using the software ANCAD (AVS Instruments, São Carlos, SP, Brazil). Jbtx and peptides were dissolved in insect physiological solution (214 mM NaCl, 3.1 mM KCl, 9 mM $CaCl_2$, 0.1 mM $MgSO_4$, and 5 mM HEPES, pH 7.2 [34]). The test solutions were prepared daily and 20 μL were injected into the insect's third abdominal segment using a Hamilton syringe.

2.12. Molecular modeling and simulation

The three-dimensional model for Jbtx was built by comparative modeling with MODELLER9v10 [35] employing the structure of the *C. ensiformis* major urease isoform (PDB ID: 3LA4), [21] as template. Ten models were built, stereochemically evaluated and theoretically validated for their three-dimensional profiles with PROCHECK [36] and Verify3D [37], respectively. The best scored model was then selected. The amino-terminal Met residue and the carboxy-terminal LEHHHHHH segment were added with SwissPDBviewer [38]. The Jbtx peptide was then subjected to molecular dynamics (MD) simulations with GROMACS 4.5 suite [39] using GROMOS96 53a6 force field [40] for 500 ns. The systems were solvated in triclinic boxes using periodic boundary conditions and SPC water models [41]. Counterions (Na^+) were added to neutralize the systems. The Lincs method [42] was applied to constrain covalent bond lengths, allowing an integration step of 2 fs after an initial energy minimization using Steepest Descents algorithm. Electrostatic interactions were calculated with Particle Mesh Ewald method [43]. Temperature and pressure were kept constant by coupling proteins, ions, and solvent to external temperature and pressure baths with coupling constants of $\tau = 0.1$ and 0.5 ps [44], respectively. The dielectric constant was treated as $\varepsilon = 1$, and the reference temperature was adjusted to 300 K. The system was slowly heated from 50 to 300 K, in steps of 5 ps, each step increasing the reference temperature by 50 K, allowing a progressive thermalization of the molecular system. The simulation was performed to 500 ns, with no restraint, considering a reference value of 3.5 Å between heavy atoms for a hydrogen-bond, and a cutoff angle of 30° between hydrogen-donor–acceptor [39].

2.13. Statistical analysis

Data were evaluated by ANOVA followed by the Bonferroni's or Student *t* test using GraphPad Prism software (Version 5.0 for Windows). See legends to figures for more details. A $p < 0.05$ was considered statistically significant.

3. Results

3.1. MD simulation of Jbtx

It has been previously suggested that a prominent β-hairpin in the predicted model of the urease-derived peptide Jbtx could be responsible at least in part for its membrane-disturbing activity and some of its biological properties [17,19]. The presence of this β-hairpin was confirmed by x-ray crystallographic data of jack bean urease [21], and short simulations of the crystal-derived peptide were performed [45].

In order to establish if this β-hairpin would still be present in the peptide once it has been released from the urease molecule, a 3D-model of Jbtx was constructed using the crystal structure of jack bean urease as template and subjected to molecular dynamics for 500 ns (Fig. 2, panels A and B). The MD simulation indicated that Jbtx becomes more globular when in aqueous solution (Fig. 2, panel B), changing its conformation along the simulation with an increase of RMSD, as compared to the initial crystal-derived structure (Supplementary Fig. 1). The secondary structure of Jbtx changed in solution, with loss of many helix turns and formation of a minute beta sheet. The β-hairpin at Jbtx's C-terminal half was conserved despite the increase in coil content (Fig. 2, panels B and D).

3.2. Expression of recombinant jaburetox (Jbtx) and mutated forms

Aiming to identify motifs probably involved in the biological activities of Jbtx, mutated forms of the peptide lacking the internal β-hairpin (JbtxΔ-β), the N-terminal half (Jbtx C-ter) or the C-terminal half (Jbtx N-ter) domains were constructed. A schematic representation of these peptides is shown in Fig. 1.

All the His-tagged peptides were purified and analyzed by SDS-PAGE (Fig. 3A and B). The predicted molecular masses of the peptides based on their deduced amino acid sequences are 11,193 Da for Jbtx, 9625.6 Da for JbtxΔ-β, 6325.8 Da for Jbtx N-ter and 6772.5 Da for Jbtx C-ter. As it can be observed from the SDS-PAGE results, all the recombinant peptides showed the expected mass, except for the Jbtx N-ter peptide which behaved as a dimer with an estimated molecular mass of approximately 12 kDa (Fig. 3B).

Anti-Jbtx-2Ec polyclonal antibodies recognized equally Jbtx-2Ec, Jbtx and JbtxΔ-β (result not shown) and although with a weaker reactivity, also interacted with the two half-peptides (Fig. 3C).

All bands seen in the lanes corresponding to each peptide were excised from the SDS-PAGE gels, digested with trypsin and submitted to MS/MS analysis. The identities of the peptides Jbtx, JbtxΔ-β, and Jbtx C-ter (and aggregated forms of the peptides) were confirmed by MS/MS analysis as shown in Fig. 3D. On the other hand, the peptide Jbtx N-ter was not identified in the MS/MS assay. The band corresponding to the dimer of the peptide Jbtx N-ter in the SDS-PAGE (Fig. 3B) reacted positively with the anti-Jbtx antibodies (Fig. 3C), thus confirming its identity. The tendency to form aggregates previously described for jaburetox-2Ec [19,46] persisted in Jbtx, as well as in all the mutated forms of this peptide, as confirmed by the MS/MS analysis. After a few days in aqueous solution, all the peptides formed insoluble precipitates. These aggregates did not revert to the monomeric state under a number of tested conditions [19]. High ionic strength accelerates the aggregation of Jbtx (data not shown), suggesting hydrophobic interactions as a driving force for the oligomerization process. Because it was not possible to ascertain the oligomeric state of each peptide in solution, their monomeric states were considered when expressing molar concentrations in the subsequent assays.

Since all mutated peptides retained considerable antigenicity towards anti-Jbtx-2Ec polyclonal antibodies, they probably kept their tridimensional structures, resembling the corresponding portions in Jbtx. The CD spectrum of Jbtx (not shown) indicated the presence mainly of irregular structures, with a minor contribution of β-sheets and helices. This type of CD spectrum has been observed for Chab I,

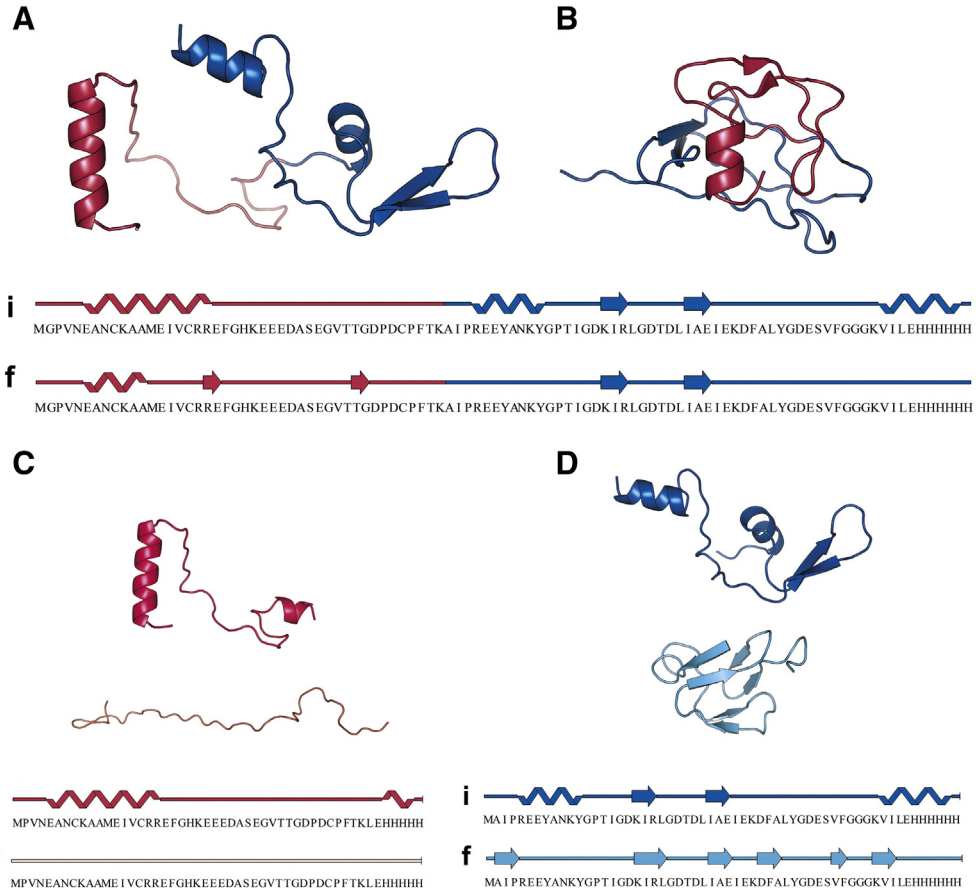


Fig. 2. Structural changes in jaburetox and its mutated versions after MD simulation of 500 ns. Three-dimensional representations of the full Jbtx peptide (A) initial and (B) final structures, with the N-terminal domain (residues 1–44) depicted in pink and the C-terminal domain in blue; (C) Jbtx N-ter (amino-terminal mutant): top, initial state; bottom, final state; (D) Jbtx C-ter (carboxy-terminal mutant): top, initial state; bottom, final state; Schematic representations of the secondary structure content of the (i) initial and (f) final structures are colored according to their three-dimensional counterparts. The corresponding amino acid sequences are also shown.

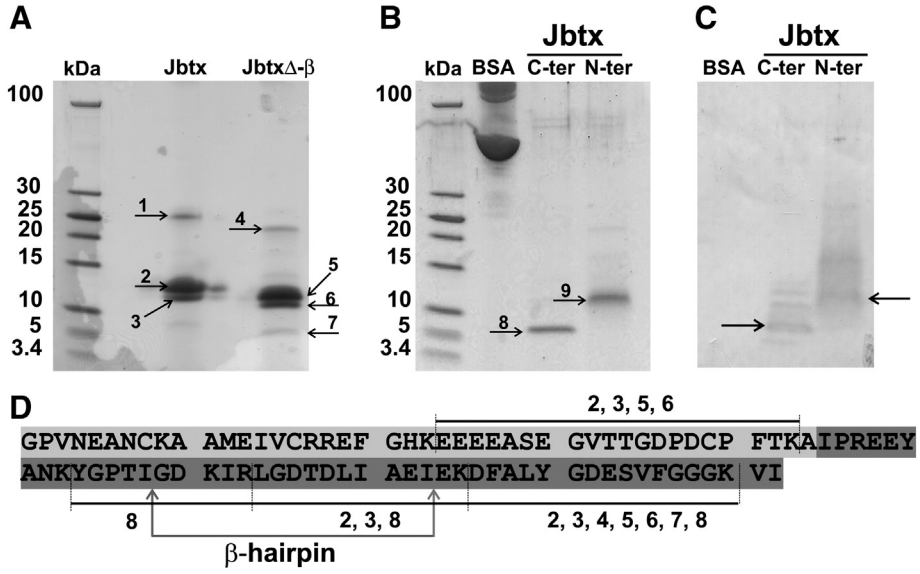


Fig. 3. (A and B) SDS-Tricine PAGE of jaburetox and their derived peptides. Numbered arrows indicate the bands that were excised and analyzed by mass spectrometry. *Lanes*: Jbtx, jaburetox; JbtxΔ-β, jaburetox with deleted β-hairpin motif; BSA, bovine serum albumin; C-ter, carboxy-terminal region of jaburetox; N-ter, amino-terminal region of jaburetox. (C) Western blot analysis with polyclonal anti-jaburetox antibodies. *Lanes*: BSA, bovine serum albumin as negative control; C-ter, carboxy-terminal region of jaburetox; N-ter, amino-terminal region of jaburetox; (D) amino acid sequence of jaburetox. The numbered lines above and below the sequence correspond to the arrows in panels A and B, showing parts of the jaburetox sequence identified by mass spectrometry. The sequence of Jbtx N-ter mutant is shown in light gray and that of the Jbtx C-ter in dark gray. The region corresponding to the β-hairpin is also indicated.

a charybdotoxin analog [47], and also for the acid unfolded state of equine β-lactoglobulin, which has residual helices and β-hairpins [48,49].

Simulations were carried out to establish the putative structures in solution of the mutated peptides representing the two half domains of Jbtx. The N-terminal mutant (residues 1 to 44) became completely unfolded after simulation (Fig. 2, panel C), while the C-terminal mutant (residues 45 to 93) showed propensity towards stabilization of a newly-formed β-sheet (Fig. 2, panel D).

3.3. Vesicle leakage promoted by Jbtx peptides

We employed LUVs composed by L-α-phosphatidic acid [19] as a membrane model to evaluate which part of the Jbtx molecule interacts with phospholipid membranes and induces vesicle leakage. Fig. 4 shows typical results. Vesicle leakage was more prominent when LUVs were treated with either Jbtx C-ter or Jbtx (5 μg/mL), although all peptides produced at least 80% of leakage at the end of the 10 min incubation period. Taken together, these findings showed that the β-hairpin is not essential for the membrane-disruptive activity of Jbtx. Moreover, the data indicated that all Jbtx-related peptides are able to induce LUV leakage, while the C-terminal region of the peptide seems to contribute the greatest effect. In fact, hydropathicity plots indicated the presence of prominent hydrophobic regions in both, the N-terminal and the C-terminal domains of Jbtx (Supplementary Fig. 2).

3.4. Insecticidal effect of Jbtx peptides

In order to compare the insecticidal activity of Jbtx to that previously described for Jbtx-2Ec [6,18], we tested the entomotoxic effect of Jbtx upon injection into *R. prolixus* nymphs. Employing a dose of 0.05 μg/mg of insect weight, 100% mortality was observed 48 h after injection (result not shown), indicating that the absence of the V5 epitope in Jbtx did not affect its insecticidal property. When the insecticidal activity of the β-hairpin deleted form (JbtxΔ-β) was assayed in *R. prolixus* nymphs, it produced an entomotoxic (mortality) effect equivalent to that of the original Jbtx, either by injection (Fig. 5A) or by feeding (Fig. 5B). Four days after injection into fifth instars *R. prolixus*, we have observed that the Jbtx N-term induced up to 60% mortality, while Jbtx C-ter caused less than 10% mortality

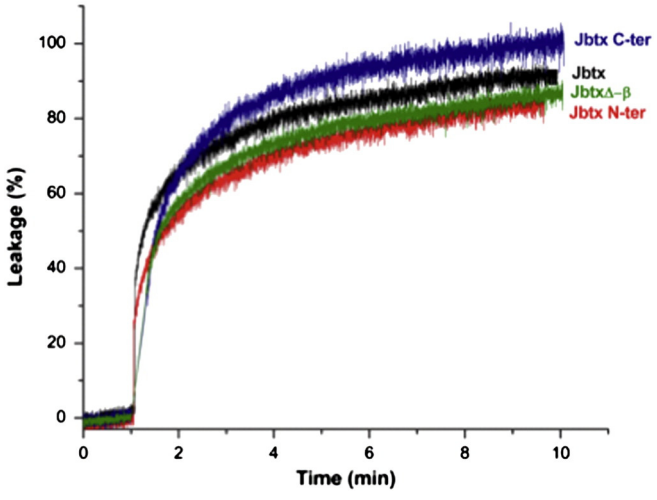


Fig. 4. Effect of jaburetox and derived mutants on LUVs composed by L-α-phospatidic acid. The carboxyfluorescein release assay was performed for each peptide at a final concentration of 5 μg/mL (Jbtx, 0.44 μM; JbtxΔ-β, 0.51 μM; Jbtx N-ter, 0.79 μM; and Jbtx C-ter, 0.73 μM) in 25 mM Tris, pH 7.0. The absence of leakage (0%) corresponds to the fluorescence of the vesicles at time zero; 100% leakage was taken as the value of fluorescence intensity obtained after addition of 1% (v/v) Triton X-100. The experiments were performed at 25 °C. The figure shows superimposed tracings of a typical result for each peptide to facilitate comparison.

(Fig. 5A). On the other hand, 24 h after feeding, both Jbtx N-ter and Jbtx C-ter had similar lethal effects on *R. prolixus* nymphs, ranging from 60 to 80% mortality (Fig. 5B).

Fifth instars *O. fasciatus* were also injected with Jbtx and its mutant variants. Similarly to what was observed for *R. prolixus* upon injections, Jbtx N-ter (Fig. 6A) and JbtxΔ-β (Fig. 6B) displayed lethal effects comparable to that of Jbtx, while Jbtx C-ter was near to inactive (Fig. 6A), suggesting that the N-terminal portion of the Jbtx carries its insecticidal domain.

3.5. Antidiuretic effect of Jbtx-related peptides on Malpighian tubules

We have previously described that, in the dose range of 10^{-16} to 10^{-15} M, Jbtx 2-Ec inhibited the serotonin-stimulated fluid secretion in *R. prolixus* Malpighian tubules [20]. Fig. 7 shows that Jbtx and all its variants, at a concentration of 1×10^{-15} M, were able to inhibit fluid secretion in the tubules producing similar antidiuretic effect.

3.6. In vivo neuromuscular blockade of cockroach nerve–muscle preparations induced by Jbtx-related peptides

The injection of Jbtx or its mutant versions (32 μg/g of animal weight) produced a time-dependent blockade of the cockroach nerve–muscle preparation (Fig. 8). Jbtx was the most effective and induced a complete neuromuscular paralysis at 35 ± 10 min followed by Jbtx N-ter at 80 ± 2 min (Fig. 8B). In contrast, the neuromuscular blockades induced by Jbtx Δ-β or Jbtx C-ter were only partial at the end of the 120 min recording time. The administration of insect saline alone did not interfere with normal neuromuscular responses during 120 min recordings (Fig. 8A). Thus, similar to what was observed in the case of the insecticidal activity (upon injection), these data suggest that the N-terminal half of Jbtx carries its entomotoxic domain. In this type of assay, however, there is a contribution of the β-hairpin to the effect.

4. Discussion

In this study we evaluated the jack bean urease-derived peptide Jbtx and three domain-deleted variants in order to identify the regions of the molecule that are critical for its entomotoxic activities. A previous version of Jbtx, harboring a large V5-antigen derived from the pET101/D-TOPO plasmid and called jaburetox-2Ec (Jbtx-2Ec), was shown to be lethal to *R. prolixus* by oral route and hemocoel injection [50,51] and to permeabilize vesicles composed of charged lipids [19]. Jbtx has the same 93 amino acid urease-derived sequence and the polyhistidine tail found in Jbtx-2Ec, but lacks the V5 epitope present in the later. Here we demonstrated that Jbtx displays insecticidal activity equivalent to that described for Jbtx-2Ec, evidencing that the epitope V5 is not implied in its entomotoxicity.

Comparing the structure obtained here for Jbtx with the model of jaburetox-2Ec generated by comparative modeling [19] (prior to the first description of the crystal structure of a plant urease [21]), important conformational similarities can be seen in the region correspondent to the short helix as well as in the large content of random coil conformation, even though jaburetox-2Ec exhibited a more well-defined β-hairpin than Jbtx. The differences in the models generated for these peptides might be attributed in part to the distinct initial structures used in the MD simulation, as *H. pylori* urease and jack bean urease served as template in the comparative modeling for jaburetox-2Ec [19] and Jbtx (this work), respectively.

Balasubramanian and Ponnuraj [21] were the first to report in 2010 the crystal structure of a plant (*C. ensiformis*, jack bean) urease at 2.05 Å of resolution. These authors confirmed the presence of an internal β-hairpin motif in the jack bean urease, previously suggested by our group to be present in the structure of jaburetox-2Ec [17,19], and proposed to be involved in the insecticidal activity of both urease and its derived entomotoxic peptide. The same group described insecticidal

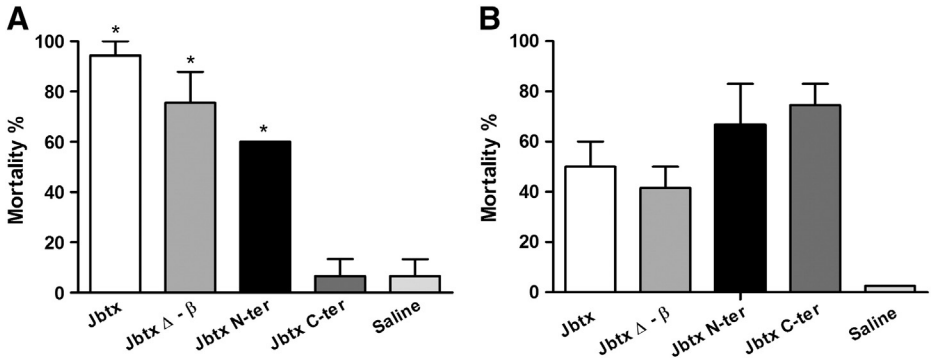


Fig. 5. Insecticidal effect of Jbtx and derived peptides on fifth instar *Rhodnius prolixus*. (A) Groups of 5 insects were injected with each peptide separately at final doses of 0.05 μg per mg of body weight. Control insects were injected with *Rhodnius* saline. The mortality was recorded after 96 h. Two independent bioassays were carried out for each peptide. Results shown are means and standard error. (B) Groups of 5 insects were fed on *R. prolixus* saline plus 1 mM ATP and the peptides separately at final doses of 0.1 μg per mg of body weight. Control insects were fed solely on *R. prolixus* saline plus 1 mM ATP. Mortality rate within each group was recorded after 24 h. Results shown are means and standard error.

and antifungal properties of the pigeon pea (*Cajanus cajan*) urease and reported the presence of a similar β-hairpin motif in the crystal structure of this urease [9]. Moreover, Balasubramanian and coworkers, using molecular modeling studies and short (5 ns) molecular dynamics simulations of Jbtx, suggested that its β-hairpin could self-associated into a β-barrel able to anchor into a membrane-like environment, and hypothesized an insecticidal mode of action of Jbtx based on pore formation [45].

Also present in bacterial ureases, the microbial β-hairpin motif is formed with contributions from the α- and the β urease chains while its counterpart in the single chain of plant ureases is formed exclusively by amino acids located in a region corresponding to the bacterial α-chain [21]. Our group reported that, contrasting to plant ureases, *Bacillus pasteurii* urease has no insecticidal activity against *Dysdercus peruvianus* [52]. Since the β-hairpin motif is present in the *B. pasteurii* urease as well [21], a plausible explanation for this could be the fact that part of the sequence corresponding to the N-terminal half of Jbtx is missing in bacterial ureases.

Here we demonstrated that Jbtx and JbtxΔ-β, its β-hairpin deleted version, behaved almost indistinguishably regarding LUV leakage, antidiuretic effect and insecticidal activity upon injection. These results strongly suggested that the β-hairpin motif is not involved in membrane-disturbing activity or in these biological properties of the peptide.

At that point we had no clues to any other possible motif in the Jbtx molecule that could be responsible for its biological properties, so we decided to produce two half-peptides, corresponding to the N-terminal and C-terminal half versions of Jbtx (Jbtx N-ter and Jbtx C-ter, respectively). The mutated peptides Jbtx N-ter and Jbtx C-ter were then tested for LUV leakage and for different types of entomotoxic activities. Two distinct groups of results were obtained depending on the assay: (i) Jbtx, JbtxΔ-β and Jbtx N-ter were equally active while and Jbtx C-ter was inactive or significantly less active; and (ii) all the peptides produced similar effects.

When tested for insecticidal activity upon injection into *R. prolixus* (Fig. 5A) or *O. fasciatus* (Fig. 6) nymphs, Jbtx, Jbtx Δ-β and Jbtx N-ter caused significant mortality after 96 h, while the survival rate of insects injected with Jbtx C-ter was equivalent to that of control group. It thus became clear from these experiments that the N-terminal half of Jbtx (Jbtx N-ter) has the insecticidal domain of Jbtx. This conclusion agrees with the fact that the deletion of the β-hairpin, which is present in the Jbtx C-ter, did not interfere on the entomotoxicity.

On the other hand, all the peptides were able to induce blockade of the cockroach neuromuscular junction *in vivo* (Fig. 8). The neuromuscular blockade induced by Jbtx resembles the effect of neurotoxins which act directly on receptor ion channels [53], among which are pore-forming neurotoxins [54]. In this work we did not attempt to elucidate the pharmacological interactions of Jbtx and related peptides at specific sites of insect neuromuscular junctions. The Jbtx N-ter peptide had an effect comparable to that of the intact peptide producing almost complete neuromuscular blockade after 40 min of recordings while JbtxΔ-β and Jbtx C-ter were clearly less active. The activity loss of JbtxΔ-β in this bioassay may reflect some critical alteration of the peptide 3D-structure affecting also its N-terminal domain, which alone is capable of producing full effect in the absence of the β-hairpin.

Upon feeding to *R. prolixus* (Fig. 5B) all the peptides were lethal, even Jbtx C-ter, contrasting with its lack of activity when injected into the hemolymph. This fact points to the presence of two active domains in the Jbtx molecule, with the amphipatic β-hairpin in the C-terminal domain probably interacting with insect's gut membranes as predicted,

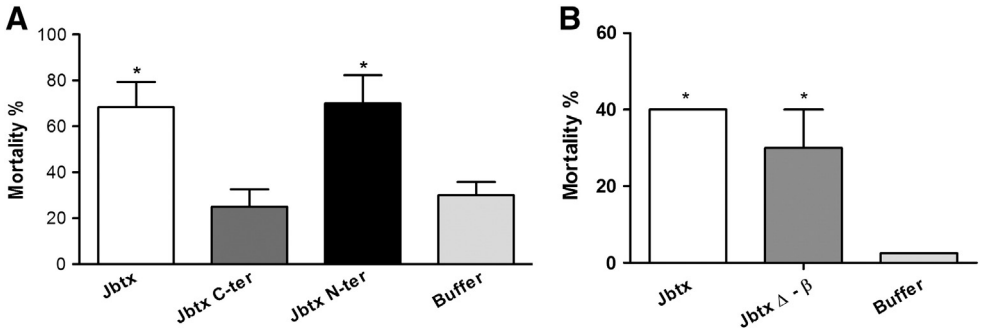


Fig. 6. Insecticidal effect of Jbtx and derived peptides on fifth instars *Oncopeltus fasciatus*. (A) Groups of 10 nymphs were injected with 1.5 μL of Jbtx, Jbtx N-ter or Jbtx C-ter into the hemocoel (dose of 0.015 μg/mg of insect body weight) or 20 mM phosphate buffer, pH 7.5 (control group). (B) Groups of 5 nymphs were injected with 1.5 μL of Jbtx or JbtxΔ-β peptides into the hemocoel (dose of 0.015 μg/mg of insect body weight) or 20 mM phosphate buffer, pH 7.5 (control group). The mortality rate was recorded after 96 h. Results are means ± standard error of triplicates of two independent experiments. (*) indicates statistically significant difference ($p \leq 0.05$) from the control group.

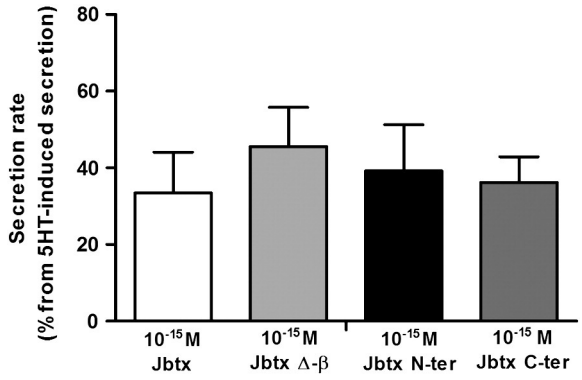


Fig. 7. Effect of jaburetox and mutants on secretion of *Rhodnius prolixus* Malpighian tubules. The assay was performed as described by Staniscuaski et al. [20]. Tubules were incubated with 2.5×10^{-8} M serotonin (5-hydroxy-tryptamine, 5-HT) for 20 min to record the maximal secretion. After washing, the tubules were incubated with the peptides (1×10^{-15} M) in the presence of serotonin for another 20 min. The secretion rate was expressed as a percentage from the control (serotonin without peptides). Results shown are means ± standard deviation of 5–6 replicates for each peptide.

when given by oral route. This conclusion could also be drawn from the facts that all the peptides had equivalent antidiuretic effects (Fig. 7) and that both terminal domains of Jbtx were able to induce leakage of vesicles (Fig. 4). The preliminary results using Planar Lipid Bilayers another artificial membrane composed only of lipids, also showed that all mutant versions of Jbtx form ion channels displaying membrane-disturbing properties (Piovesan A., unpublished data). On the other hand, Jbtx C-ter showed significantly lower activity than Jbtx or Jbtx N-ter, when its first contact within the insect was with the hemolymph probably due to a "saturating" effect of the lipid-rich medium on its membrane-disrupting ability.

All the peptides, including Jbtx C-ter, produced antidiuresis or were lethal given by oral route, circumstances where their first interaction happened with single cell layered tissues such as the Malpighian tubules [55] or the gut [56]. One hypothesis to explain the lack of specificity of these assays to discriminate the different Jbtx variants could be that biological multilayered tissue systems, such as the neuromuscular junction [57] and the whole insect (by injection, skipping the first contact with the gut), probably add additional levels of tissue- or cell specificity to the entomotoxic effects of Jbtx-related peptides. Altogether, our data indicate that the main entomotoxic domain of the urease-derived peptide Jbtx is located in its N-terminal half. However, depending on the bioassay, the C-terminal domain and/or its β-hairpin motif could also contribute part of the biological activity of Jbtx.

From the molecular dynamics simulation, it seems that the monomeric Jbtx peptide is mostly formed by coils (Fig. 2). Our simulation results confirm and expand previous theoretical observations [19], such as the compaction of the peptide in solution. Simulations of the half-peptides indicated that after 500 ns Jbtx N-ter adopts a random coil conformation while Jbtx C-ter acquires a newly-formed β-sheet (Fig. 2). These data may explain why Jbtx is highly prone to aggregation [19], and the instability of Jbtx N-ter in aqueous solution (unpublished results), possibly a consequence of the unfolding of the highly hydrophobic N-terminal of Jbtx that would require protein–protein (or protein–membrane) contact to stabilize.

Presently, to the best of our knowledge, it is not possible to compare the MD simulated structure of Jbtx to that of any other known insecticidal or membrane-disrupting peptide. The high level of coils, especially in the N-terminus, may be related to the peptide toxicity,

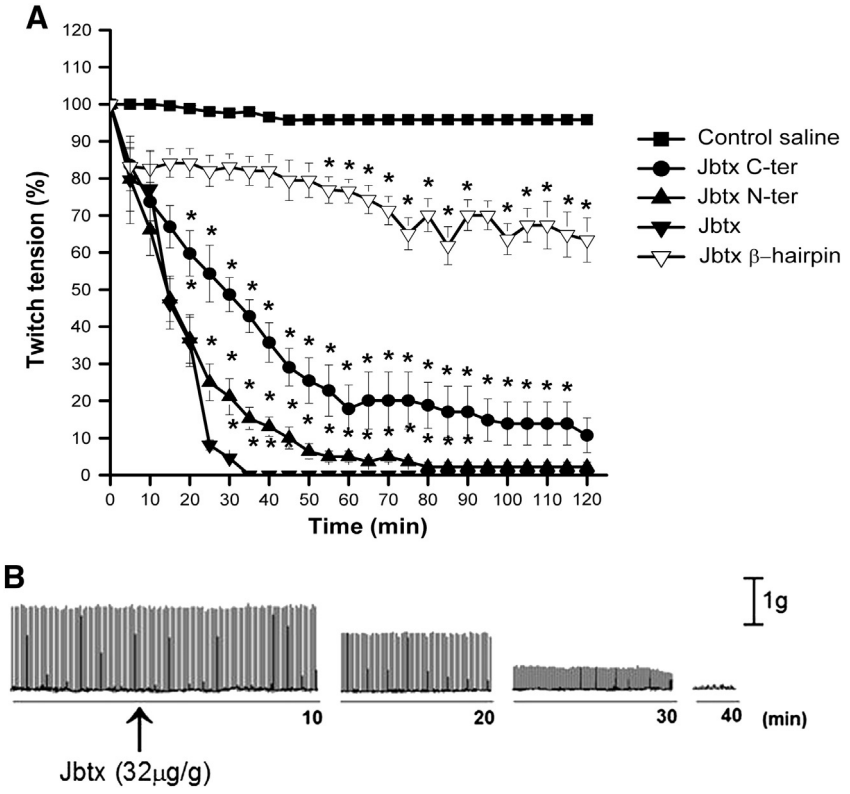


Fig. 8. Neuromuscular paralysis induced by Jbtx and peptides on *in vivo* cockroach coxal-adductor methatoracic nerve–muscle preparation. (A) Time course of the blockade of the neuromuscular activity in the presence of 32 μg/g of each version of Jbtx peptides against control insects treated only with saline (means ± standard error, n = 12). Note that Jbtx and Jbtx N-ter were able to induce complete paralysis. * indicates $p \leq 0.05$ in comparison to control saline, with ANOVA two way and Student *t* test. (B) Representative myographic 120 min recording of the coxal-adductor methatoracic nerve–muscle preparation of a Jbtx-treated cockroach.

since some toxins employ these unfolded states as recognition motifs. One example of such toxins is colicin, from *E. coli* [58]. These unfolded recognition domains may be advantageous for the toxins that carry them, since they allow these proteins to overcome steric restrictions while providing large average interaction surfaces per residue [58–60]. There are many reports in the literature of folding and oligomerization of proteins and peptides that acquire their biologically active state upon interaction with lipids or membranes. Examples are cecropin A, a 37-residue insect antimicrobial peptide [61,62], the Cyt1Aa toxin produced by *Bacillus thuringiensis* [63], anticancer β-hairpin peptides [64], antimicrobial, cell-penetrating peptides and fusion peptides such as the HIV fusion peptide FP23 [65], to cite a few.

The tendency to oligomerize and to interact with lipids exhibited by Jbtx brings the question whether the active form of the peptide (or its N-ter and C-ter versions) is an oligomer rather than a monomer. The oligomerization/aggregation phenomenon was also observed for Jbtx-2Ec, causing an enormous impact of the membrane-disruptive ability of the peptide [19].

Jbtx has promising biotechnological potential as a biopesticide. We are currently testing transgenic Jbtx expressing sugar cane (*Saccharum officinarum*) plants, and so far we have observed an increase of resistance to several species of lepidopterans in greenhouse conditions (Becker-Ritt et al., unpublished data). These data indicate the effectiveness of the peptide as an environment friendly insecticide with practical application, reducing crop losses while avoiding the use of chemical toxic agents.

We conclude that the urease-derived peptide Jbtx probably represents a new example of membrane-active peptide with insecticidal and fungitoxic activities. Its insecticidal activity was tracked down mostly to its N-terminal region and does not require the prominent β-hairpin present in the C-terminal region, although this part of the molecule probably contributes to its overall entomotoxic properties. Understanding the complex behavior of these peptides in solution as well as in the presence of lipids and biological membranes is a critical step towards unraveling their mechanisms of action and exploiting their potential as insecticidal agents.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbagen.2013.11.010.

Authors' contributions

A.H.S.M. and K.K. constructed the mutated peptides, M.S.D. helped in the insect bioassays, A.R.P. and C.F. carried out LUVs leakage assay, F.S. run the Malpighian tubules assay, D.R.D. performed MS assays, R.L-B. and H.V. conducted molecular modeling and simulations, C.A.D.B. and C.G.M.A. tested the peptides on the cockroach neuromuscular junction, C.R.C. wrote the paper and together with G.P., conceived and supervised all the work.

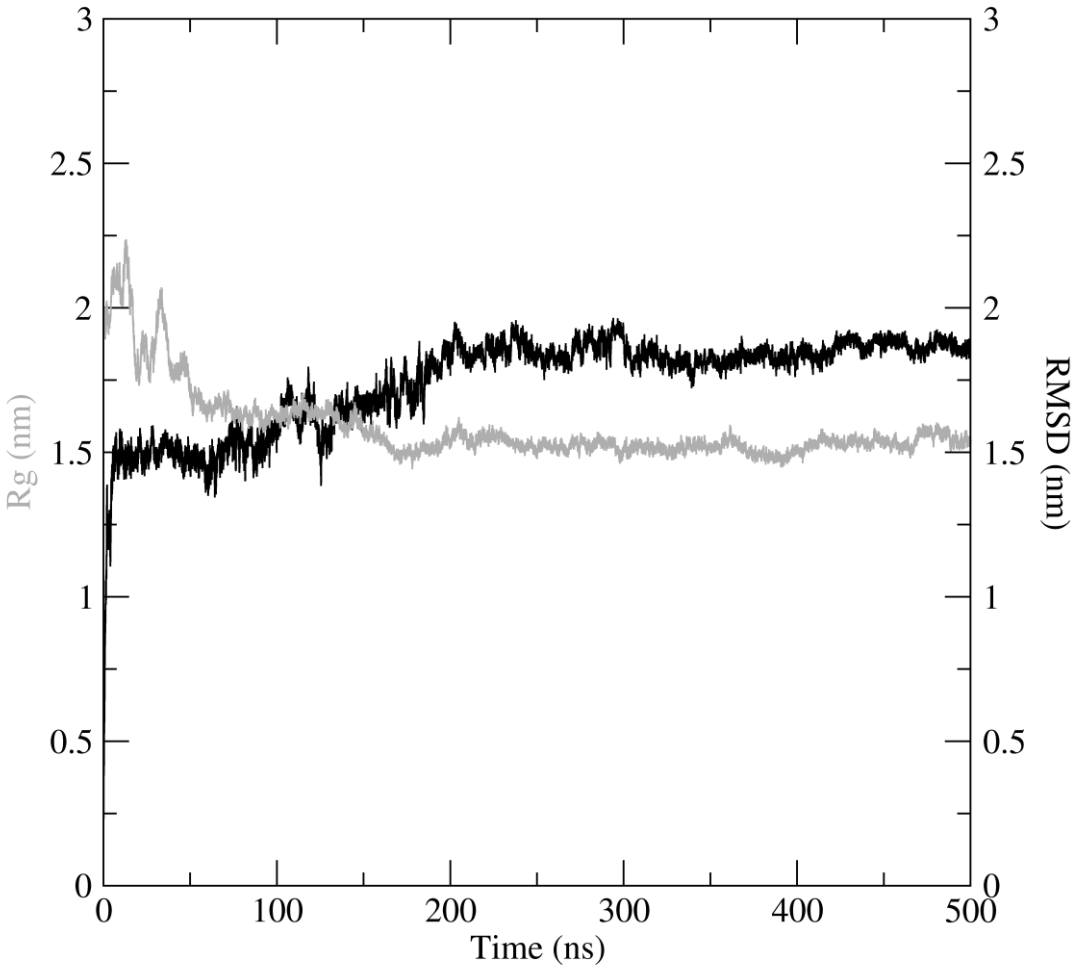
Acknowledgements

The authors wish to thank the Centro Nacional de Supercomputação at Universidade Federal do Rio Grande do Sul, for the assistance and access to the supercomputer; Dr. Yraima Cordeiro, Inst. Biophysiscs Carlos Chagas Filho, Universidade Federal do Rio Grande do Sul, for preliminary circular dichroism analyses of the peptides and B.Sc. Marinês de Avila Heberle, Unipampa, for helping collecting data on the cockroach preparation. This work was supported by grants from the Brazilian agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Edital de Toxinologia [proj 54/2011]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)—Edital Universal [proj. 47.0785/2011-47.5908/2012]; Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS)—PRONEX [proj. 10/0014-2]. The authors declare no conflicts of interest related to this work.

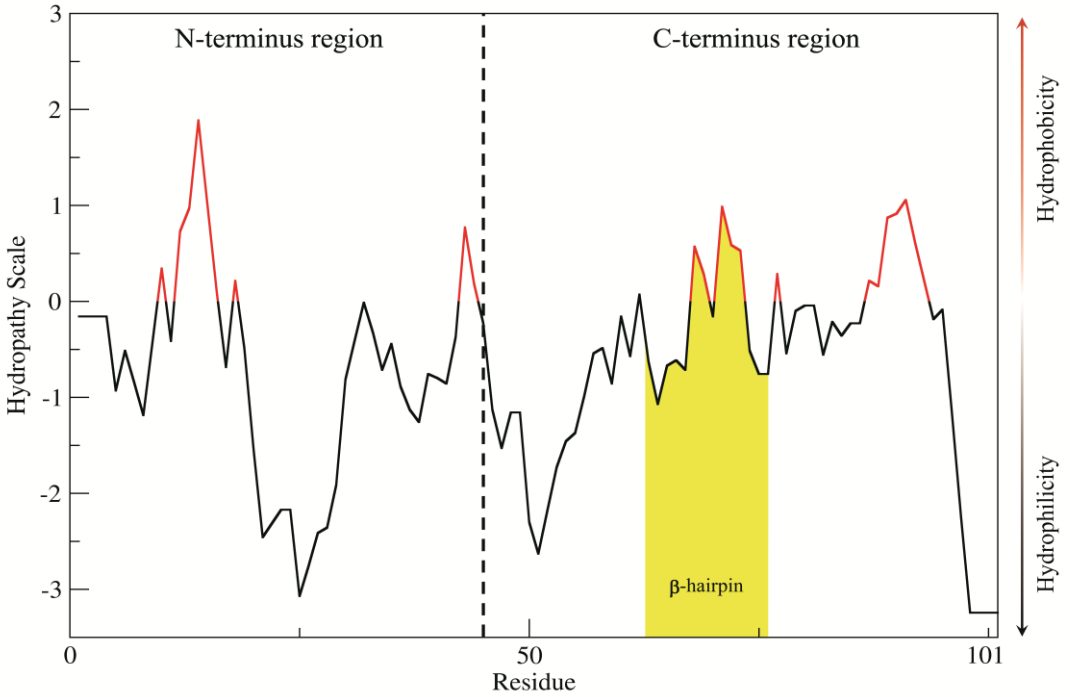
References

[1] R. Ligabue-Braun, F.C. Andreis, H. Verli, C.R. Carlini, 3-to-1: unraveling structural transitions in ureases, Naturwissenschaften 100 (2013) 459–467.
[2] J.C. Polacco, M.A. Holland, Roles of urease in plant cells, Int. Rev. Cytol. 145 (1993) 65–103.
[3] C.R. Carlini, J.C. Polacco, Toxic properties of urease, Crop Sci. 48 (2008) 1665–1672.
[4] C.R. Carlini, M.F. Grossi-de-Sa, Plant toxic proteins with insecticidal properties. A review on their potentialities as bioinsecticides, Toxicon 40 (2002) 1515–1539.
[5] A.B. Becker-Ritt, C.R. Carlini, Fungitoxic and insecticidal plant polypeptides, Biopolymers 38 (2012) 367–384.
[6] F. Staniscuaski, C.R. Carlini, Plant ureases and related peptides: understanding their entomotoxic properties, Toxins 4 (2012) 55–67.
[7] A.B. Becker-Ritt, A.H. Martinelli, S. Mitidieri, V. Feder, G.E. Wassermann, L. Santi, M.H. Vainstein, J.T. Oliveira, L.M. Fiuza, G. Pasquali, C.R. Carlini, Antifungal activity of plant and bacterial ureases, Toxicon 50 (2007) 971–983.
[8] M. Postal, A.H.S. Martinelli, A.B. Becker-Ritt, R. Ligabue-Braun, D.R. Demartini, S.F.F. Ribeiro, G. Pasquali, V.M. Gomes, C.R. Carlini, Antifungal properties of Canavalia ensiformis urease and derived peptides, Peptides 38 (2012) 22–32.
[9] A. Balasubramanian, V. Durairajpandian, S. Elumalai, N. Mathivanan, A.K. Munirajan, K. Ponnuraj, Structural and functional studies on urease from pigeon pea (*Cajanus cajan*), Int. J. Biol. Macromol. 58 (2013) 301–309.
[10] C.R. Carlini, J.A. Guimaraes, Isolation and characterization of a toxic protein from *Canavalia ensiformis* (jack bean) seeds, distinct from concanavalin A, Toxicon 19 (1981) 667–675.
[11] C. Follmer, G.B. Barcellos, R.B. Zingali, O.L. Machado, E.W. Alves, C. Barja-Fidalgo, J.A. Guimaraes, C.R. Carlini, Canatoxin, a toxic protein from jack beans (*Canavalia ensiformis*), is a variant form of urease (EC 3.5.1.5): biological effects of urease independent of its ureolytic activity, Biochem. J. 360 (2001) 217–224.
[12] C.R. Carlini, A.E. Oliveira, P. Azambuja, J. Xavier-Filho, M.A. Wells, Biological effects of canatoxin in different insect models: evidence for a proteolytic activation of the toxin by insect cathepsinlike enzymes, J. Econ. Entomol. 90 (1997) 340–348.
[13] C.T. Ferreira-DaSilva, M.E. Gombarovits, H. Masuda, C.M. Oliveira, C.R. Carlini, Proteolytic activation of canatoxin, a plant toxic protein, by insect cathepsin-like enzymes, Arch. Insect Biochem. Physiol. 44 (2000) 162–171.
[14] A.R. Piovesan, F. Staniscuaski, J. Marco-Salvadori, R. Real-Guerra, M.S. Defferrari, C.R. Carlini, Stage-specific gut proteinases of the cotton stainer bug *Dysdercus peruvianus*: role in the release of entomotoxic peptides from *Canavalia ensiformis* urease, Insect Biochem. Mol. Biol. 38 (2008) 1023–1032.
[15] M.S. Defferrari, D.R. Demartini, T.B. Marcelino, P.M. Pinto, C.R. Carlini, Insecticidal effect of *Canavalia ensiformis* major urease on nymphs of the milkweed bug *Oncopeltus fasciatus* and characterization of digestive peptidases, Insect Biochem. Mol. Biol. 41 (2011) 388–399.
[16] R. Real-Guerra, C.R. Carlini, F. Staniscuaski, Role of lysine and acidic amino acid residues on the insecticidal activity of jackbean urease, Toxicon 71 (2013) 76–83.
[17] F. Mulinari, F. Staniscuaski, L.R. Bertholdo-Vargas, M. Postal, O.B. Oliveira-Neto, D.J. Rigden, M.F. Grossi-de-Sa, C.R. Carlini, Jaburetox-2Ec: an insecticidal peptide derived from an isoform of urease from the plant *Canavalia ensiformis*, Peptides 28 (2007) 2042–2050.
[18] F. Staniscuaski, C.T. Ferreira-Dasilva, F. Mulinari, M. Pires-Alves, C.R. Carlini, Insecticidal effects of canatoxin on the cotton stainer bug *Dysdercus peruvianus* (Hemiptera: Pyrrhocoridae), Toxicon 45 (2005) 753–760.
[19] P.R. Barros, H. Stassen, M.S. Freitas, C.R. Carlini, M.A.C. Nascimento, C. Follmer, Membrane-disruptive properties of the bioinsecticide Jaburetox-2Ec: implications to the mechanism of the action of insecticidal peptides derived from ureases, Biochim. Biophys. Acta, Proteins Proteomics 1794 (2009) 1848–1854.
[20] F. Staniscuaski, V.T. Brugge, C.R. Carlini, I. Orchard, In vitro effect of *Canavalia ensiformis* urease and the derived peptide Jaburetox-2Ec on *Rhodnius prolixus* Malpighian tubules, J. Insect Physiol. 55 (2009) 255–263.
[21] A. Balasubramanian, K. Ponnuraj, Crystal structure of the first plant urease from jack bean: 83 years of journey from its first crystal to molecular structure, J. Mol. Biol. 400 (2010) 274–283.
[22] A. Menez, Functional architectures of animal toxins: a clue to drug design? Toxicon 36 (1998) 1557–1572.
[23] A. Nikouee, M. Khabiri, S. Grissmer, R. Ettrich, Charybdotoxin and margatoxin acting on the human voltage-gated potassium channel hKv1.3 and its H399N mutant: an experimental and computational comparison, J. Phys. Chem. B 116 (2012) 5132–5140.
[24] X.D. Li, Y.F. Qiu, Y. Shen, C. Ding, P.H. Liu, J.P. Zhou, Z.Y. Ma, Splicing together different regions of a gene by modified polymerase chain reaction-based site-directed mutagenesis, Anal. Biochem. 373 (2008) 398–400.
[25] S.F. Altschul, D.J. Lipman, Protein database searches for multiple alignments, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 5509–5513.
[26] M.M. Bradford, A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, Anal. Biochem. 72 (1976) 248–254.
[27] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, Protein identification and analysis tools on the ExPASy server, in: J.M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press, 2005, pp. 571–607.
[28] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[29] D.R. Demartini, C.R. Carlini, J.J. Thelen, Global and targeted proteomics in developing jack bean (*Canavalia ensiformis*) seedlings: an investigation of urease isoforms mobilization in early stages of development, Plant Mol. Biol. 75 (2011) 53–65.

[30] H. Schagger, G. Vonjagow, Tricine sodium dodecyl-sulfate polyacrylamide-gel electrophoresis for the separation of proteins in the range from 1-kDa to 100-kDa, Anal. Biochem. 166 (1987) 368–379.
[31] N. Dyballa, S. Metzger, Fast and sensitive colloidal Coomassie G-250 staining for proteins in polyacrylamide gels, J. Vis. Exp. 30 (2009).
[32] H. Towbin, T. Staehelin, J. Gordon, Electrophoretic transfer of proteins form polyacrylamide gels to nitrocellulose sheets. Procedure and some applications, Proc. Natl. Acad. Sci. U. S. A. 76 (1979) 4350–4354.
[33] R.J. Full, D.R. Stokes, A.N. Ahn, R.K. Josephson, Energy absorption during running by leg muscles in a cockroach, J. Exp. Biol. 201 (1998) 997–1012.
[34] K.A. Wafford, D.B. Sattelle, Effects of amino acid neurotransmitter candidates on an identified insect motoneurone, Neurosci. Lett. 63 (1986) 135–140.
[35] R. Sánchez, A. Šali, Comparative protein structure modeling: introduction and practical examples with Modeller, Methods Mol. Biol. 143 (2000) 97–129.
[36] R.A. Laskowski, M.W. Macarthur, D.S. Moss, J.M. Thornton, PROCHECK—a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283–291.
[37] R. Luthy, J.U. Bowie, D. Eisenberg, Assessment of protein models with 3-dimensional profiles, Nature 356 (1992) 83–85.
[38] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, Electrophoresis 18 (1997) 2714–2723.
[39] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation, J. Chem. Theory Comput. 4 (2008) 435–447.
[40] C. Oostenbrink, A. Villa, A.E. Mark, W.F. Van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6, J. Comput. Chem. 25 (2004) 1656–1676.
[41] H.J.C. Berendsen, J.R. Grigera, T.P. Straatsma, The missing term in effective pair potentials, J. Phys. Chem. 91 (1987) 6269–6271.
[42] B. Hess, H. Bekker, H.J.C. Berendsen, J. Fraaije, LINCS: a linear constraint solver for molecular simulations, J. Comput. Chem. 18 (1997) 1463–1472.
[43] T. Darden, D. York, L. Pedersen, Particle Mesh Ewald—an N. Log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.
[44] H.J.C. Berendsen, J.P.M. Postma, W.F. Vangunsteren, A. Dinola, J.R. Haak, Molecular dynamics coupling to an external bath, J. Chem. Phys. 81 (1984) 3684–3690.
[45] A. Balasubramanian, N. Balaji, N. Gautham, K. Ponnuraj, Molecular dynamics simulation and molecular modelling studies on the insecticidal domain from jack bean urease, Mol. Simul. 39 (2012) 357–366.
[46] K. Kappaun, Estudos com o Jaburetox: efeito tóxico de *E. coli* liofilizadas carregadas com o peptídeo e análise da influencia do epitopo V5 na formação de agregados, (M.Sc. dissertation) Cellular and Molecular Biology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, 2012.
[47] E. Drakopoulou, J. Vizzavona, J. Neyton, V. Aniort, F. Bouet, H. Virelizier, A. Menez, C. Vita, Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation, Biochemistry 37 (1998) 1292–1301.
[48] K. Nakagawa, A. Tokushima, K. Fujiwara, M. Ikeguchi, Proline scanning mutagenesis reveals non-native fold in the molten globule state of equine beta-lactoglobulin, Biochemistry 45 (2006) 15468–15473.
[49] M. Yamamoto, K. Nakagawa, M. Ikeguchi, Importance of polypeptide chain length for the correct local folding of a beta-sheet protein, Biophys. Chem. 168 (2012) 40–47.
[50] G. Tomazetto, F. Mulinari, F. Staniscuaski, B. Settembrini, C.R. Carlini, M.A.Z. Ayub, Expression kinetics and plasmid stability of recombinant *E-coli* encoding urease-derived peptide with bioinsecticide activity, Enzyme Microb. Technol. 41 (2007) 821–827.
[51] F. Staniscuaski, V.T. Brugge, C.R. Carlini, I. Orchard, Jack bean urease alters serotonin-induced effects on *Rhodnius prolixus* anterior midgut, J. Insect Physiol. 56 (2010) 1078–1086.
[52] C. Follmer, R. Real-Guerra, G.E. Wasserman, D. Olivera-Severo, C.R. Carlini, Jackbean, soybean and *Bacillus pasteurii* ureases: biological effects unrelated to ureolytic activity, Eur. J. Biochem. 271 (2004) 1357–1363.
[53] G. Corzo, E. Villegas, F. Gomez-Lagunas, L.D. Possani, O.S. Belokoneva, T. Nakajima, Oxyopinins, large amphipathic peptides isolated from the venom of the wolf spider Oxyopes kitabensis with cytolytic properties and positive insecticidal cooperativity with spider neurotoxins, J. Biol. Chem. 277 (2002) 23627–23637.
[54] Z. Andreeva-Kovalevskaya, A.S. Solonin, E.V. Sineva, V.I. Ternovsky, Pore-forming proteins and adaptation of living organisms to environmental conditions, Biochem. Mosc. 73 (2008) 1473–1492.
[55] S.H.P. Maddrell, Secretion by the Malpighian tubules of Rhodnius. The movements of ions and water, J. Exp. Biol. 51 (1969) 71–97.
[56] W.R. Terra, Evolution of digestive systems of insects, Annu. Rev. Entomol. 35 (1990) 181–200.
[57] G.A. Edwards, H. Ruska, E. de Harven, Neuromuscular junctions in flight and tymbal muscles of the Cicada, J. Biochem. Biophys. Cytol. 4 (1958) 251–256.
[58] G. Anderluh, Q. Hong, R. Boetzel, C. MacDonald, G.R. Moore, R. Virden, J.H. Lakey, Concerted folding and binding of a flexible colicin domain to its periplasmic receptor TolA, J. Biol. Chem. 278 (2003) 21860–21868.
[59] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.R. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, M. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.
[60] V.N. Uversky, Natively unfolded proteins: a point where biology waits for physics, Protein Sci. 11 (2002) 739–756.
[61] L. Silvestro, P.H. Axelsen, Membrane-induced folding of cecropin A, Biophys. J. 79 (2000) 1465–1477.
[62] L. Otvos, Antibacterial peptides isolated from insects, J. Pept. Sci. 6 (2000) 497–511.
[63] C. Rodriguez-Almazan, I. Ruiz de Escudero, P. Emiliano Canton, C. Munoz-Garay, C. Perez, S.S. Gill, M. Soberon, A. Bravo, The amino- and carboxyl-terminal fragments of the Bacillus thuringensis Cyt1Aa toxin have differential roles in toxin oligomerization and pore formation, Biochemistry 50 (2011) 388–396.
[64] C. Sinthuvanich, A.S. Veiga, K. Gupta, D. Gaspar, R. Blumenthal, J.P. Schneider, Anticancer beta-hairpin peptides: membrane-induced folding triggers activity, J. Am. Chem. Soc. 134 (2012) 6210–6217.
[65] P. Wadhwani, J. Reichert, J. Bürck, A.S. Ulrich, Antimicrobial and cell-penetrating peptides induce lipid vesicle fusion by folding and aggregation, Eur. Biophys. J. 41 (2012) 177–187.



Supplementary Fig. 1. Time course of conformational changes of Jbtx in aqueous solution. All-atom root mean square deviation (RMSD, gray line), and radius of gyration (black line) of the Jbtx molecular dynamic simulation during 500 ns.



Supplementary Fig. 2. Hydropathicity plot for jaburetox. The profile was calculated considering the Kyte–Doolittle scale (see Materials and methods) and a window size of 7 residues. The position of the β-hairpin motif in the C-terminal domain is indicated.

5. DISCUSSÃO GERAL

"In the 'discussion' you adopt the ludicrous pretense of asking yourself if the information you have collected actually means anything."

Peter Medawar

A reconstrução filogenética apresentada nesta tese foi a primeira a incluir, de forma sistemática, dados de todos os *taxa* sintetizadores de urease. Apesar disso, a história evolutiva proposta apresenta algumas limitações. A principal delas envolve a impossibilidade de incluir um grupo externo sem excluir grandes segmentos das sequências alinhadas. A dihidroorotase, grupo externo mais próximo de ureases, apresenta correspondência apenas com o domínio (ou subunidade) ativo da urease. Entretanto, ao mapearmos na estrutura tridimensional as regiões que foram mantidas no alinhamento empregado na construção das árvores deste trabalho (Figura 11), observamos que a maior conservação se deu, justamente, na região catalítica.

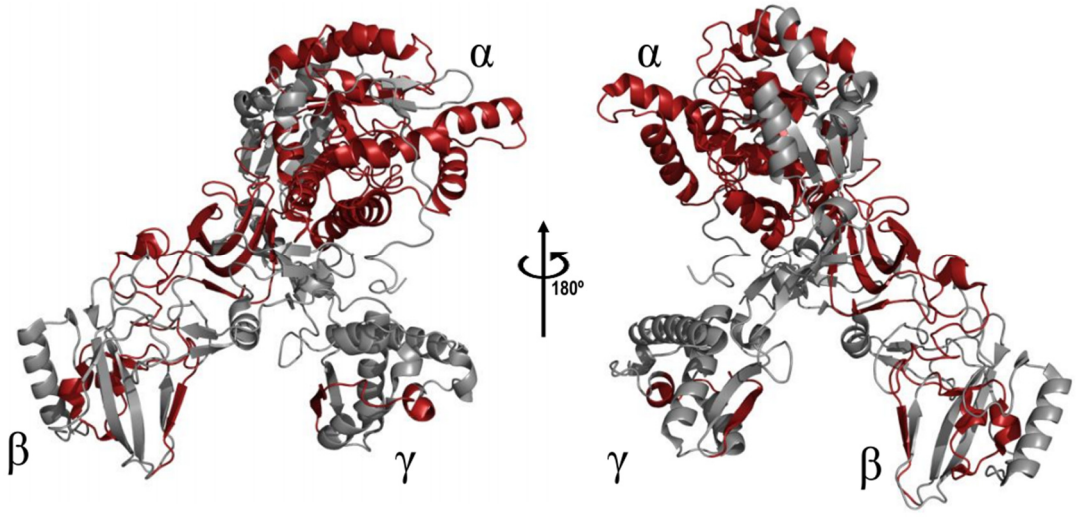


Figura 11. Regiões de variabilidade de sequência identificadas com SimPlot. Em vermelho, regiões conservadas nos alinhamentos; em cinza, regiões altamente variáveis. Os dados foram transpostos à estrutura tridimensional da urease de *C. ensiformis* (PDB ID 3LA4).

Assim, é possível que a reconstrução filogenética empregando apenas o domínio catalítico seja recomendada para essas enzimas. Por outro lado, a análise de conservação evidencia que as regiões não envolvidas em catálise são muito divergentes, e podem explicar as grandes diferenças observadas para as propriedades *moonlighting* de ureases. Outras abordagens possíveis para a reconstrução filogenética de casos extremos incluem aumentar ao máximo possível o número de sequências amostradas e a correção dos valores de *bootstrap* (conforme proposto por SANDERSON & SHAFFER, 2002).

Outro aspecto importante para análise evolutiva de ureases diz respeito à sua distribuição na natureza. Sua presença em bactérias, arqueas, plantas e fungos está bem estabelecida, enquanto alguns autores defendem a existência de ureases em (alguns) animais. Na década de 1990, enquanto FUJIWARA & NOGUCHI (1995) apontavam evidências de que o ancestral dos animais tinha perdido essa enzima, PEDROZO ET AL. (1996AB) defendiam sua existência no molusco *Aplysia californica*, atuando como coadjuvantes no processo de formação de inclusões de carbonato de cálcio. Artigos mais antigos caracterizaram ureases em diferentes invertebrados, como vermes, crustáceos e equinodermos (BROOKBANK & WHITELEY, 1954; SIMMONS, 1961). Em nenhum desses casos foram realizados testes para excluir a possibilidade de que a urease fosse oriunda de algum organismo simbionte, como ocorre em nematoides, por exemplo (SALVADORI ET AL., 2012).

Foi demonstrado que a urease do bicho-da-seda (*Bombyx mori*) é exógena, sendo absorvida da dieta (constituída exclusivamente por folhas da amoreira *Morus alba*), indicando que a urease vegetal passa intacta do intestino à hemolinfa (HIRAYAMA ET AL., 2000AB). Mais recentemente, foi identificada a expressão de uma proteína similar à UreG em anfioxo (*Branchiostoma belcheri*) (XUE ET AL., 2006), reacendendo a ideia de que animais possuem ureases ainda não descritas. A UreG, contudo, é a menos conservada das proteínas acessórias de ureases (ZAMBELLI ET AL., 2007) e nenhum sinal da enzima ativa foi detectado no animal, limitando a significância dessa descoberta.

Em uma descoberta um tanto surpreendente, BAI ET AL. (2013) associaram a presença de níquel na nogueira-pecã (*Carya illinoinensis*) a uma transição de atividade na RNAse de xilema, que passa a ter atividade ureásica na presença do metal. Para demonstrar que RNAses podem atuar como ureases, os mesmos

autores converteram a RNAse pancreática bovina a uma urease com 40% da atividade específica observada para a enzima de *C. ensiformis (*BAI ET AL., 2013). A transição de atividade, associada ao requerimento pouco compreendido de níquel em animais domésticos (SPEARS, 1984), poderia explicar o motivo dos animais terem perdido a urease ao longo de sua história evolutiva sem ter nenhum prejuízo aparente no metabolismo de nitrogênio.

Considerando a proposta de que a transição de três subunidades para uma unidade fundida em ureases possa ter ocorrido por transferência horizontal, fica evidente a necessidade de análises envolvendo dados genéticos (i.e. sequências de ácidos nucleicos em vez de aminoácidos). A transferência horizontal (seja por meio de elementos móveis ou vírus e incluindo a transferência entre organelas e o núcleo) tem se mostrado uma das mais importantes forças evolutivas, especialmente por promover novidades genéticas entre *taxa* não relacionados (GRIBALDO & BROCHIER, 2009; RAOULT, 2010; SCHAACK ET AL., 2010; DUNNING HOTOPP, 2011). Estima-se que pelo menos um terço, e talvez a totalidade, dos genes tenha sofrido transferência horizontal em algum ponto de sua história (DAGAN & MARTIN, 2007), entretanto, a inclusão desses eventos em árvores filogenéticas é problemática, sendo propostas alternativas, como a formação de redes, gráficos em anel ou tridimensionais (MCINERNEY ET AL., 2011). No momento não há consenso quanto à melhor forma de incluir eventos de transferência horizontal em análises filogenéticas tradicionais, mantendo as árvores como a representação de escolha (O'MALLEY & KOONIN, 2011).

Eventos de fusão gênica, provocando união de domínios proteicos originalmente independentes, foi demonstrada para algumas proteínas bacterianas (PASEK ET AL., 2006; ENRIGHT ET AL., 1999). Nesses casos, foi frequente a fusão de genes adjacentes. É interessante destacar que a fusão observada em ureases não ocorreu em bactérias (ou arqueas), que possuem enzimas de duas ou três subunidades, o que poderia excluir esse tipo de mecanismo como fonte das fusões observadas. Em animais, a descrição da fusão de domínios é mais desafiadora (BABUSHOK ET AL., 2007), e parece ter envolvido a duplicação gênica na maioria dos casos (BULJAN ET AL., 2010). A fusão de domínios esbarra em um tema muito mais complexo e sobre o qual existem debates em andamento, a origem de novos genes (TAUTZ & DOMAZET-

LOŠO, 2011; BULJAN ET AL., 2010). Além disso, a presença de um encadeossomo ("spliceossomo") como um caráter ancestral em linhagens procarióticas, conforme proposto por LANE ET AL. (2007) e FORTERRE (2011), também poderia explicar a fusão de subunidades observada em ureases, apesar de ser pouco suportada para outros casos.

As propostas estruturais para os intermediários de ativação da urease de *K. aerogenes* propostos nesta tese foram pioneiros, ao serem os primeiros a oferecer resolução em nível atômico a um oligômero antes observado apenas em termos gerais de forma e volume. Um estudo posterior (BIAGI ET AL., 2013), também baseado em atracamento, propôs estruturas para os intermediários de ativação em *H. pylori*. Seus resultados apresentam boa concordância com os resultados apresentados para *K. aerogenes*, apesar das grandes diferenças observadas na estrutura quaternária de *Helicobacter* (dodecamérica) em comparação à de *Klebsiella* (trimérica). Recentemente, FONG ET AL. (2013) publicaram uma estrutura cristalográfica para o complexo UreD(H)-Ure-F-UreG em *Helicobacter*, que difere daquelas propostas por atracamento, especialmente quanto à ligação de UreG (Figura 12).

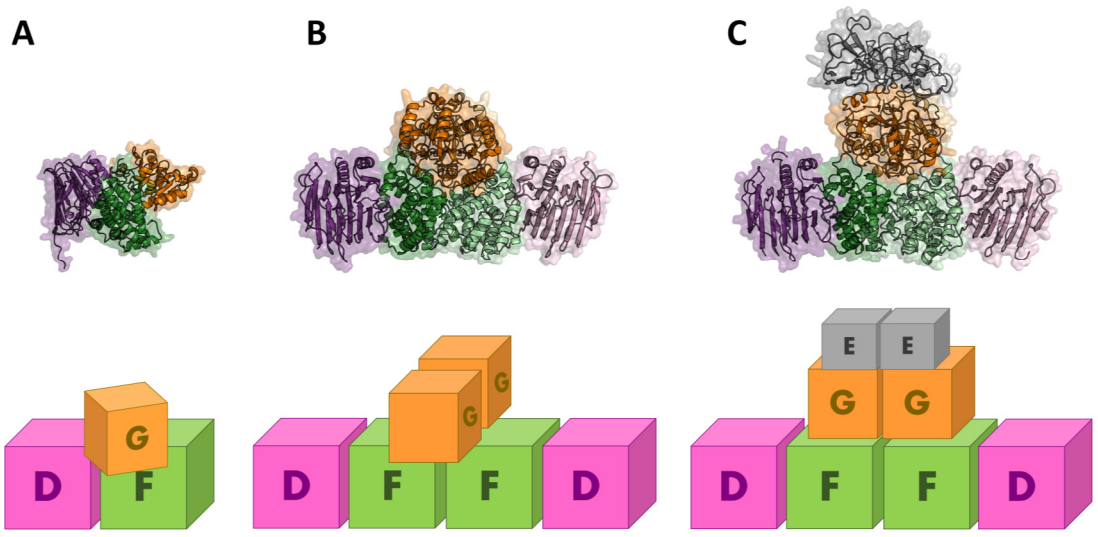


Figura 12. Estruturas para complexos de ativação de urease. (A) Complexo UreDFG de *K. aerogenes* derivado de docagem (esta tese); (B) Complexo UreDFG de *H. pylori* descrito por cristalografia de raios-X (FONG ET AL., 2013); (C) Complexo UreDFGE de *H. pylori* derivado de docagem (BIAGI ET AL., 2013). A representação em cubos (baseada em FONG ET AL., 2013) ilustra esquematicamente as diferenças observadas nas diferentes propostas para os complexos de ativação. Estrutura B gentilmente cedida por F. Musiani, estrutura C gentilmente cedida por K-B Wong.

Outro aspecto importante diz respeito à ligação das proteínas à apourease. O estudo do complexo ativador em *K. aerogenes* iniciou com a docagem à apoenzima trimérica, enquanto BIAGI ET AL. (2013) SE basearam na estrutura do complexo UreD-UreF publicado anteriormente por FONG ET AL. (2011), sem calcular sua ligação à apozenzima dodecamérica de *H. pylori.* Da mesma forma, o cristal obtido por FONG ET AL. (2013) não incluiu a urease. A comparação dos dados de atracamento de UreD a UreF e então a UreG filtrados por perfis de SAXS mostrou-se equivalente a perfis do complexo empregando UreDF baseados em FONG ET AL. (2011). A resolução do SAXS poderia explicar as diferenças observadas na orientação das proteínas (especialmente UreG) entre os resultados mostrados na Figura 12.

BIAGI ET AL. (2013) levantaram a possibilidade de que a orientação observada para UreG no complexo cristalográfico de FONG ET AL. (2013), que havia sido divulgado preliminarmente por FARRUGIA ET AL. (2013), poderia ser diferente daquela observada por atracamento por retratar a forma inativa da proteína, que requer alteração conformacional para realizar sua atividade. Entretanto, ao publicarem a descrição final da estrutura cristalográfica de UreD-UreF-UreG de *H. pylori*, FONG ET AL. (2013) incluíram ensaios de atividade, demonstrando que a UreG retratada no cristal estava ativa. É importante destacar o caráter desordenado da proteína acessória UreG, que pode limitar a aplicação de estruturas cristalográficas, reforçando a necessidade de analises dinâmicas no estudo desses sistemas.

Outro aspecto importante diz respeito ao estado de oligomerização das proteínas acessórias e sobre seu papel em cada etapa. A Tabela 1 resume as diferenças observadas em duas revisões recentes, oriundas dos dois principais grupos trabalhando no tema. Em nossa proposta de docagem, empregamos proteínas monoméricas, conforme QUIROZ-VALENZUELA ET AL. (2008) para *K. aerogenes*, enquanto BIAGI ET AL. (2013) empregaram proteínas diméricas, como descrito para *H. pylori*. Além disso, nossa estrutura foi montada sequencialmente sobre o trímero de unidades funcionais de ureases, enquanto BIAGI ET AL. (2013) não utilizaram referências à apourease em seus cálculos. Da mesma forma, FONG ET AL. (2013) cristalizaram o complexo UreD-UreF-UreG sem a presença do oligômero de ureases.

O complexo proposto por BIAGI ET AL. (2013) inclui a proteína UreE, a última a interagir com o complexo de ativação. Essa proteína apresenta um caráter parcialmente desordenado (ZAMBELLI ET AL., 2013), evidenciando a necessidade de estudos dinâmicos, como sugerido para UreG. Além disso, por ser a responsável pela "entrega" do níquel que eventualmente será inserido no sítio ativo da urease, a descrição adequada desses fenômenos provavelmente requerirá cálculos quânticos, do tipo QM-MM, associado a outros experimentos não cristalográficos, alguns deles já publicados (ZAMBELLI ET AL., 2013).

Tabela 1: Diferenças quanto às proteínas acessórias, conforme descrito em duas revisões recentes.

	ZAMBELLI ET AL., 2009	**CARTER ET AL., 2009**
UreD	Homodimérica	Monomérica
	Considera sem relevância biológica a fusão UreD-MBP.	Desenvolveu proteína híbrida UreD-proteína de ligação a maltose (MBP) na expectativa de caracterizar melhor a estrutura da proteína
UreF	Homodimérica	Monomérica
	É uma GAP (proteína ativadora de GTPase)	Pode ser GAP
UreG	Monomérica ou homodimérica (dependendo do organismo)	Monomérica
	GTPase, intrinsicamente desordenada (única enzima com tal característica)	Possui alto teor de estrutura secundária em solução
UreE	Homodimérica	Homodimérica
	Metalo-chaperona, entrega Ni	Metalo-chaperona, entrega Ni
Montagem do Complexo	Seqüencial (o complexo de ativação se forma sobre a apourease)	Seqüencial ou não (pode haver encontro do complexo de ativação pré-formado com a apourease)
	Não necessariamente simultâneo (unidades podem ou não ser ativadas ao mesmo tempo)	Simultâneo (todas as unidades ativadas ao mesmo tempo)

O Jaburetox demonstra a tendência a não possuir estrutura secundária em solução, apesar de ser capaz de atuar como toxina em organismos-modelo. Tal característica permite considera-lo uma proteína intrinsicamente desordenada. Estas proteínas não apresentam estruturas organizadas, apenas pequenos fragmentos com maior propensão a se organizarem em elementos curtos do tipo hélice ou fita sob determinadas circunstâncias (UVERSKY, 2002; DUNKER ET AL., 2001). Para estas proteínas não parece existir uma única estrutura responsável pela função analisada, mas um conjunto (um *ensemble*) de estruturas necessárias para que o efeito final seja obtido (DYSON & WRIGHT, 2005). Todavia, a aceitação de que o paradigma *estrutura-função* poderia não ser tão estrito quanto se pensava só tomou força nos últimos dez anos (TOMPA, 2012).

Os motivos para um domínio (ou uma proteína inteira, como observado em alguns casos) ser desorganizado são tão variados quanto as proteínas estudadas, não havendo um padrão que seja discernível atualmente (DYSON, 2011; UVERSKY & DUNKER, 2010). É proposto, entretanto, que a desordem proteica constitua uma ferramenta adaptativa para a conquista de novos ambientes, sendo evolutivamente conservada, apesar da manutenção da desorganização ser altamente não-trivial (SCHLESSINGER ET AL., 2011). Os mesmos autores especulam que o aumento da desordem proteica tenha sido um fator crucial na transição de procariotos a eucariotos. Especula-se que muitos dos genes órfãos (ORFans ou TRGs, genes taxonomicamente restritos), genes que parecem ser exclusivos de uma linhagem, sem homólogos em linhagens evolutivamente próximas, apresentem estruturas desordenadas, o que dificultaria sua identificação pela maioria dos métodos atualmente empregados para identificação gênica automatizada (TAUTZ & DOMAZET-LOŠO, 2011; SCHLESSINGER ET AL., 2011).

A urease, uma enzima comum, facilmente obtida, tem-se apresentado cada mais como uma biomolécula modelo, seja por suas diferentes propriedades, seja por suas aplicações biotecnológicas. Sua reação de catálise enganadoramente simples, dependente de um elemento raro em sistemas biológicos, já seria em si motivo de interesse, mas as ureases oferecem muito mais. Apresentam várias propriedades independentes da atividade catalítica,

sendo o protótipo de proteína *moonlighting*, além de serem capazes de originar peptídeos intrinsicamente desenovelados ao serem ingeridas por organismos suscetíveis. Sua cristalização levou quase um século para ser efetivada, gerando um prêmio Nobel e definindo o que hoje conhecemos como enzima; enquanto suas aplicações biotecnológicas, seja no controle de pragas agrícolas ou na construção de abrigos de luxo em dunas, está limitada apenas pelo número de novas propriedades a serem descritas para essas proteínas. Cento e cinquenta anos depois de sua descoberta, o potencial para inovações envolvendo ureases parece longe de ser esgotado.

6. CONCLUSÕES

"What is wanted is not the will-to-believe, but the wish to find out, which is its exact opposite."

Bertrand Russell

Considerando os objetivos propostos, o presente trabalho permitiu:

► reconstruir a possível história evolutiva de ureases, propondo relações de ancestralidade e derivação das ureases com diferentes organizações estruturais;

► propor estruturas tridimensionais para os intermediários de ativação de ureases que estão de acordo com dados obtidos anteriormente;

► caracterizar o comportamento conformacional de peptídeos derivados da urease de *C. ensiformis*, evidenciando sua tendência a um comportamento diferente do previamente descrito para toxinas similares.

Além disso, permitiu aplicar e consolidar diferentes metodologias computacionais no estudo de ureases, associado os mesmos a diferentes experimentos realizados em bancada. De forma geral, os resultados obtidos auxiliam na compreensão das relações estrutura-função em ureases, contribuindo em sua aplicação biotecnológica além de aprofundar o conhecimento de aspectos evolutivos a respeito dessas enzimas.

7. PERSPECTIVAS

"Urease! What a picture this word conjures up for all those who have had even the most fleeting flirtation with chemistry."

Burt Zerner

Considerando-se os procedimentos computacionais empregados no presente trabalho, especialmente relacionados ao estudo de ureases, as seguintes perspectivas podem ser traçadas:

- Ampliação a análise evolutiva de ureases, incorporando dados genéticos. Nesse caso, seria interessante guiar os alinhamentos de nucleotídeos com base nos alinhamentos de aminoácidos, além de incorporar dados de organização gênica (organização em operons, separação de éxons e íntrons, inclusão de regiões reguladoras). Além disso, a separação da análise nos diferentes domínios/subunidades e a análise da filogenia das proteínas acessórias poderiam revelar histórias evolutivas diferentes daquela observada para ureases inteiras.
- Obtenção de consenso quanto ao estado de oligomerização das proteínas acessórias de ureases, o que permitiria um ajuste mais fino das propostas para estruturas de intermediários de ativação dessas enzimas. Adicionalmente, poderia ser buscada a realização de novos cálculos de atracamento associados a experimentos de bancada que possam restringir as soluções obtidas.
- Análise conformacional dos peptídeos derivados de urease por diferentes métodos, buscando aprofundar o conhecimento sobre a relação estrutura-função nessas toxinas putativas. De maneira mais específica, a amostragem conformacional dos peptídeos poderá ser ampliada com o emprego de novas técnicas de simulação (especialmente troca de réplicas com têmpera do solvente, REST2), além de técnicas biofísicas, como microcalorimetria e RMN.

8. REFERÊNCIAS BIBLIOGRÁFICAS

"In order to know soup, it is not necessary to climb into a pot and be boiled."

Oliver Heaviside

Abascal, F.; Zardoya, R.; Posada, D.: ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. **2005**, *21*, 2104-2105.

Aguetoni Cambuí, C., Gaspar, M., Mercier, H.: Detection of urease in the cell wall and membranes from leaf tissues of bromeliad species. *Physiol. Plant*., **2009**, *136*, 86-93

Alonso, A.; Almendral, M.J.; Baez, M.D.; Porras, M.J.; Alonso, C.: Enzyme immobilization on an epoxy matrix. Determination of L-arginine by flow-injection techniques. *Anal. Chim. Acta*, **1995**, *308*, 164-169.

Altschul, S.F., Lipman, D.J.: Protein database searches for multiple alignments, *Proc. Natl. Acad. Sci. USA*, **1990**, *87*, 5509-5513.

Anderluh, G., Hong, Q., Boetzel, R., MacDonald, C., Moore, G.R., Virden, R., Lakey, J.H.: Concerted folding and binding of a flexible colicin domain to its periplasmic receptor TolA. *J. Biol. Chem*., **2003**, *278*, 21860-21868.

Andreeva-Kovalevskaya, Z., Solonin, A.S., Sineva, E.V., Ternovsky, V.I., Pore-forming proteins and adaptation of living organisms to environmental conditions. *Biochem. Mosc*., **2008**, *73*, 1473-1492.

Arango, C.P.: Molecular approach to the phylogenetics of sea spiders (Arthropoda: Pycnogonida) using partial sequences of nuclear ribosomal DNA. *Mol. Phylogenet. Evol*, **2003**, *28*, 588-600.

Babushok, D.V., Ostertag, E.M., Kazazian, H.H. Jr.: Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol. Life Sci*., **2007**, *64*, 542-554.

Bai, C., Liu, L., Wood, B.W.: Nickel affects xylem Sap RNase A and converts RNase A to a urease. *BMC Plant Biol*., **2013**, *13*, 207.

Balasubramanian, A.; Ponnuraj, K.: Crystal structure of the first plant urease from Jack bean: 83 years of journey from its first crystal to molecular structure. *J. Mol. Biol*., **2010**, *400,* 274-283.

Balasubramanian, A.; Durairajpandian, V.; Elumalai, S.; Mathivanan, N.; Munirajan, A.K.; Ponnuraj, K.: Structural and functional studies on urease from pigeon pea (*Cajanus cajan*). *Int. J. Biol. Macromol*., **2013A**, *58*, 301-309.

Balasubramanian, A.; Balaji, N., Gautham, N., Ponnuraj, K.: Molecular dynamics simulation and molecular modelling studies on the insecticidal domain from jack bean urease. *Molecular simulation*, **2013B**, *39*, 357-366.

Baldauf, S.L.: Phylogeny for the faint of heart: a tutorial. *Trends Genet*., **2003**, *19*, 345-351.

Barros, P.R., Stassen, H., Freitas, M.S., Carlini, C.R., Nascimento, M.A., Follmer, C.: Membrane-disruptive properties of the bioinsecticide Jaburetox-2Ec: implications to the mechanism

of the action of insecticidal peptides derived from ureases. *Biochim. Biophys. Acta*, **2009**, 1794, 1848-1854.

Becker-Ritt, A. B.; Martinelli, A. H.; Mitidieri, S.; Feder, V.; Wassermann, G. E.; Santi, L.; Vainstein, M. H.; Oliveira, J. T.; Fiuza, L. M.; Pasquali, G.; Carlini, C. R.: Antifungal activity of plant and bacterial ureases. *Toxicon*, **2007**, *50*, 971-983.

Becker-Ritt, A.B., Carlini, C.R.: Fungitoxic and insecticidal plant polypeptides. *Biopolymers*, **2012**, *38*, 367-384.

Benini, S.; Rypniewski, W. R.; Wilson, K. S.; Miletti, S.; Ciurli, S.; Mangani, S.: A new proposal for urease mechanism based on the crystal structures of the native and inhibited enzyme from *Bacillus pasteurii*: why urea hydrolysis costs two nickels. *Structure*, **1999**, *7*, 205-216.

Benini, S.; Rypniewski, W. R.; Wilson, K. S.; Ciurli, S.; Mangani S.: Structure-based rationalization of urease inhibition by phosphate: novel insights into the enzyme mechanism. *J. Biol. Inorg. Chem*., **2001**, *6*, 778-790.

Benini, S.; Kosikowska, P.; Cianci, M.; Mazzei, L.; Vara, A.G.; Berlicki, Ł.; Ciurli, S.: The crystal structure of *Sporosarcina pasteurii* urease in a complex with citrate provides new hints for inhibitor design. *J. Biol. Inorg. Chem.*, **2013**, *18*, 391-399.

Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R.: Molecular-dynamics with coupling to an external bath. *J. Chem. Phys*., **1984**, *81*, 3684-3690.

Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P.: The missing term in effective pair potentials. *J. Phys. Chem*., **1987**, *91*, 6269-6271.

Biagi, F., Musiani, F., Ciurli, S.: Structure of the UreD-UreF-UreG-UreE complex in *Helicobacter pylori*: a model study. *J. Biol. Inorg. Chem*., **2013**, *18*, 571-577.

Boer, J.L., Quiroz-Valenzuela, S., Anderson, K.L., Hausinger, R.P.: Mutagenesis of *Klebsiella aerogenes* UreG to probe nickel binding and interactions with other urease related proteins. *Biochemistry*, **2010**, *49*, 5859-5869.

Boer, J.L.; Hausinger, R.P.: *Klebsiella aerogenes* UreF: identification of the UreG binding site and role in enhancing the fidelity of urease activation. *Biochemistry*, **2012**, *51*, 2298-2308.

Boer, J.L.; Mulrooney, S.B.; Hausinger R.P.: Nickel-dependent metalloenzymes. Arch Biochem Biophys. 2013 Sep 10. pii: S0003-9861(13)00271-3. doi: 10.1016/j.abb.2013.09.002. [No prelo]

Bradford, M.M.: A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem*., **1976**, *72*, 248-254.

Bremner, J. M.; Krogmeier, M. J.: Evidence that the adverse effect of urea fertilizer on seed germination in soil is due to ammonia formed through hydrolysis of urea by soil urease. *Proc. Natl. Acad. Sci. U.S.A*., **1989**, *86*, 8185-8188.

Broll, V. *Purificação e caracterização de urease recombinante de 'Proteus mirabilis'*. Porto Alegre: UFRGS, 2013. Dissertação (Mestrado em Biologia Celular e Molecular), Centro de Biotecnologia do Estado do Rio Grande do Sul, Universidade Federal do Rio Grande do Sul, **2013**.

Brookbank, J.W.; Whiteley, A.H.: Studies on the urease of the eggs and embryos of the sea urchin, *Strongylocentrotus purpuratus. Biological Bulletin*, **107**, *1954*, 57-63.

Buljan, M., Frankish, A., Bateman, A.: Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.*, **2010**, *11*, R74.

Burne, R. A.; Chen, Y. Y.: Bacterial ureases in infectious diseases. *Microbes Infect.*, **2000**, *2*, 533-542.

Câmara, M.P., Palm, M.E., van Berkum, P., O'Neill, N.R.: Molecular phylogeny of *Leptosphaeria* and *Phaeosphaeria. Mycologia*, **2002**, *94*, 630-640

Carlini, C. R.; Guimarães, J. A.: Isolation and characterization of a toxic protein from *Canavalia ensiformis* (jack bean) seeds, distinct from concanavalin A. *Toxicon*, **1981**, *19*, 667-676.

Carlini, C.R.; Guimarães, J.A.; Ribeiro, J.M.: Platelet release reaction and aggregation induced by canatoxin, a convulsant protein: evidence for the involvement of the platelet lipoxygenase pathway. *Br. J. Pharmacol.*, **1985**, *84*, 551-560.

Carlini, C. R.; Oliveira, A. E.; Azambuja, P.; Xavier-Filho, J.; Wells, M. A.: Biological effects of canatoxin in different insect models: evidence for a proteolytic activation of the toxin by insect cathepsin-like enzymes. *J. Econ. Entomol.*,**1997**, *90*, 340-348.

Carlini, C.R., Ferreira-DaSilva, C.T., Gombarovits, M.E.C.: Peptídeo Entomotóxico da Canatoxina: Processo de Produção. *Instituto Nacional de Propriedade Industrial*, Brasil, Patente Nº. 0003334–0, **2000**.

Carlini, C. R.; Grossi-de-Sá, M. F.: Plant toxic proteins with insecticidal properties. A review on their potentialities as bioinsecticides. *Toxicon*, **2002**, *40*, 1515-1539.

Carlini, C. R.; Polacco, J. C.: Toxic properties of ureases. *Crop Science*, **2008**, *48*, 1665-1672.

Carter, E.L.; Flugga, N.; Boer, J.L.; Mulrooney, S.B.; Hausinger, R.P.: Interplay of metal ions and urease. *Metallomics*, **2009**, *1*, 207-221.

Carter, E.L., Boer, J.L., Farrugia, M.A., Flugga, N., Towns, C. L., Hausinger, R.P.: Function of UreB in *Klebsiella aerogenes* urease. *Biochemistry*, **2011**, *50*, 9296-9308.

Carter, E.L.; Tronrud, D.E.; Taber, S.R.; Karplus, P.A.; Hausinger, R.P.: Iron-containing urease in a pathogenic bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, **2011**, *108*, 13095-13099.

Castanier, S.; Métayer-Levrel, G. L.; Perthuisot, J. P.: Ca-carbonates precipitation and limestone genesis - the microbiogeologist point of view. *Sediment. Geol.*, **1999**, *126*, 9-23.

Chang, Z., Kuchar, J., Hausinger, R.P.: Chemical cross-linking and mass spectrometric identification of sites of interaction for UreD, UreF, and urease. *The Journal of Biological Chemistry*, **2004**, *279*, 15305-15313.

Comeau, S.R.; Gatchell, D.W.; Vajda, S.; Camacho, C.J.: ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **2004**, *20*, 45-50.

Contreras-Rodriguez, A., Quiroz-Limon, J., Martins, A.M., Peralta, H., Avila-Calderon, E., Sriranganathan, N., Boyle, S.M., Lopez-Merino, A.: Enzymatic, immunological and phylogenetic characterization of *Brucella suis* urease. *BMC Microbiol.*, **2008**, *8*, 121.

Corzo, G., Villegas, E., Gomez-Lagunas, F., Possani, L.D., Belokoneva, O.S., Nakajima, T.: Oxyopinins, large amphipathic peptides isolated from the venom of the wolf spider *Oxyopes kitabensis* with cytolytic properties and positive insecticidal cooperativity with spider neurotoxins, *J. Biol. Chem*., **2002**, *277*, 23627-23637.

Cox, G. M.; Mukherjee, J.; Cole, G. T.; Casadevall, A.; Perfect, J. R. Urease as a virulence factor in experimental cryptococcosis. *Infect. Immun*., **2000**, *68*, 443-448.

Cullen, D.C.; Sethi, R.S.; Lowe, C.R.: Multi-analyte miniature conductance biosensor. *Anal. Chim. Acta*, **1990**, *231*, 33-40.

Dagan, T., Martin, W.: Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA*, **2007**, *104*, 870-875.

Darden, T.; York, D.; Pedersen, L.: Particle Mesh Ewald – an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys*., **1993**, *98*, 10089-10092.

Das, N.; Kayastha, A. M.; Srivastava, P. K.: Purification and characterization of urease from dehusked pigeonpea (*Cajanus cajan* L) seeds. *Phytochemistry.* **2002**, *61*, 513-521.

de Groot, B. L.; Grubmüller, H.: Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science*, **2001**, *294*, 2353-2357.

de Sant'Anna, C.M.R.: Glossário de termos usados no planejamento de fármacos (recomendações IUPAC 1997). *Quim. Nova*, **2002**, *25*, 505-512.

Defferrari, M.S., Demartini, D.R., Marcelino, T.B., Pinto, P.M., Carlini, C.R.: Insecticidal effect of *Canavalia ensiformis* major urease on nymphs of the milkweed bug *Oncopeltus fasciatus* and characterization of digestive peptidases. *Insect Biochem. Mol. Biol*., **2011**, *41*, 388-399.

Demartini, D.R., Carlini, C,R., Thelen, J.J.: Global and targeted proteomics in developing jack bean (*Canavalia ensiformis*) seedlings: an investigation of urease isoforms mobilization in early stages of development. *Plant Mol Biol.,* **2011**, *75*, 53-65.

Dixon, N. E.; Gazzola, C.; Blakeley, R. L.; Zerner B.: Jack bean urease (EC 3.5.1.5). A metalloenzyme. A simple biological role for nickel? *J. Am. Chem. Soc*., **1975**, *97*, 4131-4133.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.R., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C.H., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, M., Garner, E.C., Obradovic, Z.: Intrinsically disordered protein. *J. Mol. Graph. Model*., **2001**, *19*, 26-59.

Drakopoulou, E., Vizzavona, J., Neyton, J., Aniort, V., Bouet, F., Virelizier, H., Menez, A., Vita, C.: Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation. *Biochemistry*, **1998**, *37*, 1292-1301.

Du, L., Damoiseaux, R., Nahas, S., Gao, K., Hu, H., Pollard, J.M., Goldstine, J., Jung, M.E., Henning, S.M., Bertoni, C., Gatti, R.A.: Nonaminoglycoside compounds induce readthrough of nonsense mutations. *J. Exp. Med*., **2009**, *206*, 2285-2297.

Dunning Hotopp, J.C.: Horizontal gene transfer between bacteria and animals. *Trends Genet*., **2011**, *27*, 157-163.

Dyballa, N., Metzger, S.: Fast and sensitive colloidal Coomassie G-250 staining for proteins in polyacrylamide gels. *J. Vis. Exp*., **2009**, *30*, 1431.

Dyson, H.J.; Wright, P.E.: Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **2005**, *6*, 197-208.

Dyson, H.J.: Expanding the proteome: disordered and alternatively folded proteins. *Q. Rev. Biophys*., **2011**, *44*, 467-518.

Eaton, K. A.; Brooks C. L.; Morgan, D. R.; Krakowka, S.: Essential role of urease in pathogenesis of gastritis induced by *Helicobacter pylori* in gnotobiotic piglets. *Infect. Immun*., **1991**, *59,* 2470-2475.

Edwards, G.A., Ruska, H., de Harven, E.: Neuromuscular junctions in flight and tymbal muscles of the Cicada. *J. Biochem. Biophys. Cytol*., 1958, *4*, 251-256.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A.: Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **1999**, *402*, 86-90.

Estiu, G.; Merz, K. M. Jr. The hydrolysis of urea and the proficiency of urease. *J. Am. Chem. Soc*., **2004**, *126*, 6932-6944.

Estiu, G.; Merz, K. M. Jr.: Competitive hydrolytic and elimination mechanisms in the urease catalyzed decomposition of urea. *J. Phys. Chem. B*., **2007**, *111*, 10263-10274.

Farrugia, M.A., Macomber, L., Hausinger, R.P.: Biosynthesis of the urease metallocenter. *J. Biol. Chem*., **2013**, *288*, 13178-13185.

Fearon, W. R.: XII. Urease. Part I. The chemical changes involved in the zymolysis of urea. *Biochem. J*., **1923**, *17*, 84-93.

Feig, M., Karanicolas, J., Brooks, C.L. III: MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model*., **2004**, *22*, 377-395.

Ferreira-DaSilva, C. T.; Gombarovits, M. E.; Masuda, H.; Oliveira, C. M.; Carlini, C. R.: Proteolytic activation of canatoxin, a plant toxic protein, by insect cathepsin-like enzymes. *Arch. Insect. Biochem. Physiol*., **2000**, *44*, 162-171.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., Searle, S.M.: Ensembl 2012, *Nucleic Acids Res*., **2012**, *40*, D84-D90.

Follmer, C.; Barcellos, G. B.; Zingali, R. B.; Machado, O. L.; Alves, E. W.; Barja-Fidalgo, C.; Guimarães, J. A.; Carlini, C. R.: Canatoxin, a toxic protein from Jack beans (*Canavalia*

ensiformis), is a variant form of urease (EC 3.5.1.5): biological effects of urease independent of its ureolytic activity. *Biochem. J*., **2001**, *360*, 217-224.

Follmer, C.; Carlini, C.R.; Yoneama, M.-L.; Dias, J.F.: PIXE analysis of urease isoenzymes isolated from *Canavalia ensiformis* (jack bean) seeds. *Nucl. Instrum. Meth. Phys. Res. B*, **2002**, *189*, 482-486.

Follmer, C.; Real-Guerra, R.; Wasserman, G.E.; Olivera-Severo, D.; Carlini, C. R.: Jackbean, soybean and *Bacillus pasteurii* ureases: biological effects unrelated to ureolytic activity. *Eur. J. Biochem*., **2004**, *271*, 1357-1363.

Follmer, C.: Insights into the role and structure of plant ureases. *Phytochemistry*. **2008**, *69*, 18-28

Follmer, C: Ureases as a target for the treatment of gastric and urinary infections. *J. Clin. Pathol*., **2010**, *63*, 424-430.

Fong, Y.H.; Wong, H.C.; Chuck, C.P.; Chen, Y.W.; Sun, H.; Wong, K.B.: Assembly of preactivation complex for urease maturation in *Helicobacter pylori*: crystal structure of UreF-UreH protein complex. *J Biol Chem*., **2011**, *286*, 43241-43249.

Fong, Y.H., Wong, H.C., Yuen, M.H., Lau, P.H., Chen, Y.W., Wong, K.B.: Structure of UreG/UreF/UreH complex reveals how urease accessory proteins facilitate maturation of *Helicobacter pylori* urease. *PLoS Biol*., **2013**, *11*, e1001678.

Forster, M. K.: Molecular modelling in structural biology. *Micron*, **2001**, *33*, 365-384.

Forterre, P.: A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. *Res Microbiol*., **2011**, *162*, 77-91.

Fujiwara, S.; Noguchi, T.: Degradation of purines: only ureidoglycollate lyase out of four allantoin-degrading enzymes is present in mammals. *The Biochemical journal*, **1995**, *312*, 315-318.

Full, R.J., Stokes, D.R., Ahn, A.N., Josephson, R.K.: Energy absorption during running by leg muscles in a cockroach. *J. Exp. Biol*., **1998**, *201*, 997-1012.

Gasper, R.; Scrima, A.; Wittinghofer, A.: Structural insights into HypB, a GTP-binding protein that regulates metal binding. *J. Biol. Chem*., **2006**, *281*, 27492-27502.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A.: Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (Ed.). *The Proteomics Protocols Handbook*. Nova Iorque: Humana Press, **2005**.

Gaylarde, P.M., Crispim, C.A., Neilan, B.A., Gaylarde, C.C.: Cyanobacteria from Brazilian building walls are distant relatives of aquatic genera. *Omics*, **2005**, *9*, 30-42.

Gianfreda, L.; Cristofaro, A.; Rao, M. A.; Violante, A.: Kinetic Behavior of Synthetic Organo- and Organo-Mineral-Urease Complexes. *Soil Sci. Soc. Am. J*., **1995**, *59*, 811-815.

Gernert, D.: Ockham's razor and its improper use. *J. Sci. Explor*., **2007**, *21*, 135-140.

Graur, D.; Li, W.H.: Molecular phylogenetics. In: _____. *Fundamentals of Molecular Evolution*. 2.ed. Sunderland: Sinauer, **2000**.

Gray, J.J.: High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.*, **2006**, *16*, 183-193.

Gregory, T.R.: Understanding Evolutionary Trees. *Evo. Edu. Outreach*, **2008**, *1*, 121-137.

Gribaldo, S.; Brochier, C.: Phylogeny of prokaryotes: does it exist and why should we care? *Res Microbiol.*, **2009**, *160*, 513-521.

Gueneau, P., Loiseaux-De Goër, S.: *Helicobacter*: molecular phylogeny and the origin of gastric colonization in the genus. *Infect. Genet. Evol*., **2002**, *1*, 215-223.

Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **2002**, *320*, 369-387.

Guex, N., Peitsch, M.C.: SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **1997**, *18*, 2714-2723.

Ha, N.-C.; Oh, S.-T.; Sung, J. Y.; Cha, K. A.; Lee, M. H.; Oh, B.-H.: Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. *Nature Struct. Biol*., **2001**, *8*, 505-509.

Hausinger, R.P.: Urease. In: _____. *Biochemistry of nickel*. Nova Iorque: Plenum, **1993**.

Hernández-Pinzón, I., de Jesús, E., Santiago, N., Casacuberta, J.M.: The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. *J. Mol. Evol*., **2009**, *68*, 269-278.

Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M.: LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem*., **1997**, *18*, 1463-1472.

Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E.: GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation, *J. Chem. Theory Comput*., **2008**, *4*, 435-447.

Hess, P.N., Russo, C.A.M.: An empirical test of the midpoint rooting method. *Biol. J. Linn. Soc*., **2007**, *92*, 669-674.

Hirai, M.; Kawai-Hirai, R.; Hirai, T.; Ueki, T.: Structural change of jack bean urease induced by addition of surfactants studied with synchrotron-radiation small angle X-ray scattering. *Eur. J. Biochem*., **1993**, *215*, 55-61.

Hirayama, C.; Sugimura, M.; Saito, H.; Nakamura, M.: Host plant urease in the hemolymph of the silkworm, *Bombyx mori. J. Insect Physiol*., **2000A**, *46*, 1415-1421.

Hirayama, C.; Sugimura, M.; Saito, H.; Nakamura, M.: Purification and properties of urease from the leaf of mulberry, *Morus alba. Phytochemistry*. **2000B**, *53*, 325-330.

Holder, M.; Lewis, P.O.: Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, **2003**, *4*, 275-284.

Holm, L., Sander, C.: An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **1997**, *28*, 72-82.

Huber, C., Eisenreich, W., Hecht, S., Wächtershäuser, G.: A possible primordial peptide cycle. *Science*, **2003**, *301*, 938-940.

Huelsenbeck, J.P.; Ronquist, F.: MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. **2001**, *17*, 754-755.

Hunt, T., Vogler, A.P.: A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. *Mol. Phylogenet. Evol*., **2008**, **47**, 289-301.

Jabri, E.; Carr, M. B.; Hausinger, R. P.; Karplus, P. A.: The crystal structure of urease from *Klebsiella aerogenes. Science*, **1995**, *268*, 998-1004.

Jeffery, C. J.: Moonlighting proteins. *Trends Biochem. Sci*., **1999**, *24*, 8-11.

Jeffery, C. J.: Moonlighting proteins – an update. *Mol. Biosyst*., **2009**, *5*, 345-350.

Jin, M.; Rosario, W.; Watler, E; Calhoun, D. H.: Development of a large-scale HPLC based purification for the urease from *Staphylococcus leei* and determination of subunit structure. *Protein Expr. Purif*., **2004**, *34*, 111-117.

Jones, C.D., Begun, D.J.: Parallel evolution of chimeric fusion genes. *Proc Natl Acad Sci USA*, **2005**, *102*, 11373-11378.

Jones, B.D., Mobley, H.L.: *Proteus mirabilis* urease: nucleotide sequence determination and comparison with jack bean urease. *J. Bacteriol*., **1989**, *171*, 6414-6422.

Kaessman, H.: Origins, evolution, and phylogenetic impact of new genes. *Genome Res*., 2010, *20*, 1313-1326.

Kappaun, K. *Estudos com o Jaburetox: efeito tóxico de 'E. coli' liofilizadas carregadas com o peptídeo e análise da influência do epitopo V5 na formação de agregados.* Porto Alegre: UFRGS, 2013. Dissertação (Mestrado em Biologia Celular e Molecular), Centro de Biotecnologia do Estado do Rio Grande do Sul, Universidade Federal do Rio Grande do Sul, **2012**.

Karmali, A.; Domingos, A.: Monoclonal antibodies against urease from *Canavalia ensiformis*. *Biochimie*, **1993**, *75*, 1001-1006.

Karplus, P. A.; Pearson, M. A.; Hausinger, R. P.: 70 Years of Crystalline Urease: What Have We Learned?. *Acc. Chem. Res*., **1997**, *30*, 330-337.

Kodama, S.; Yotsuzuka, F.: Acid Urease: reduction of ethyl carbamate formation in sherry under simulated baking conditions. *J. Food Sci*., **1996**, *61*, 304-307.

Konieczna, I.; Zarnowiec, P.; Kwinkowski, M.; Kolesinska, B.; Fraczyk, J.; Kaminski, Z.; Kaca, W.: Bacterial urease and its role in long-lasting human diseases. *Curr. Protein. Pept. Sci*., **2012**, *13*, 789-806.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., Detours, V.: Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull*., **2001**, *58*, 19-42.

Kozakov, D., Brenke, R., Comeau, S.R., Vajda, S.: PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **2006**, *65*, 392-406.

Krajewska, B.: Ureases I. Functional, catalytic and kinetic properties: A review. *J. Mol. Catal. B: Enzym*., **2009A**, *59*, 9-21.

Krajewska, B.: Ureases II. Properties and their customizing by enzyme immobilizations: A review. *J. Mol. Catal. B: Enzym*., **2009B**, *59*, 22-40.

Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **1982**, *157*, 105-132.

Lane, C.E.; van den Heuvel, K.; Kozera, C.; Curtis, B.A.; Parsons, B.J.; Bowman, S.; Archibald, J.M.: Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*, 19908-19913.

Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; Thompson, J.D.; Gibson, T.J.; Higgins, D.G.: Clustal W and Clustal X version 2.0. *Bioinformatics*. **2007**, 23, 2947-2948.

Larsson, D.S.; Liljas, L.; van der Spoel, D.: Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Comput. Biol*. **2012**, *8*, e1002502.

Larsson, M. *Dune (AA Thesis 07-08)*. Disponível em <http://www.magnuslarsson.com/architecture/dune.asp > Acesso em: 26 jul. 2013.

Laskowski, R.A.; Macarthur, M.W.; Moss, D.S.; Thornton, J.M.: PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **1993**, *26*, 283-291.

Leach, A. R.: *Molecular Modelling: Principles and Applications*, 2.ed. Cingapura: Longman, **2001**.

Lee, M.H., Mulrooney, S.B., Renner, M.J., Markowicz, Y., Hausinger, R.P: *Klebsiella aerogenes* urease gene cluster: Sequence of ureD and demonstration that four accessory genes (ureD, ureE, ureF, and ureG) are involved in nickel metallocenter biosynthesis. *Journal of Bacteriology*, **1992**, *174*, 4324-4330.

Li, X.D., Qiu, Y.F., Shen, Y., Ding, C., Liu, P.H., Zhou, J.P., Ma, Z.Y.: Splicing together different regions of a gene by modified polymerase chain reaction-based site-directed mutagenesis, *Anal. Biochem*., **2008**, *373*, 398-400.

Liang, H., Sandberg, W.S., Terwilliger, T.C.: Genetic fusion of subunits of a dimeric protein substantially enhances its stability and rate of folding. *Proc. Natl. Acad. Sci. USA*, **1993**, *90*, 7010-7014.

Ligabue-Braun, R., Andreis, F.C., Verli, H., Carlini, C.R.: 3-to-1: unraveling structural transitions in ureases. *Naturwissenschaften*, **2013**, *100*, 459-467.

Lole, K.S.; Bollinger, R.C.; Paranjape, R.S.; Gadkari, D.; Kulkarni, S.S.; Novak, N.G.; Ingersoll, R.; Sheppard, H.W.; Ray, S.C.: Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol*. **1999**, *73*, 152-160.

Lubbers, M. W.; Rodriguez, S.B.; Honey, N.K.; Thornton, R.J.: Purification and characterization of urease from *Schizosaccharomyces pombe. Can. J. Microbiol*., **1996**, *42*,132-140.

Lüthy, R., Bowie, J.U., Eisenberg, D.: Assessment of protein models with three-dimensional profiles. *Nature*, **1992**, *356*, 83-85.

Ma, J., Lemieux, L., Gennis, R.B.: Genetic fusion of subunits I, II, and III of the cytochrome bo ubiquinol oxidase from *Escherichia coli* results in a fully assembled and active enzyme. *Biochemistry*, **1993**, *32*, 7692-7697.

Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.D., Ritchie, D.W.: HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.*, **2010**, *38*, W445-449.

Maddrell, S.H.P.: Secretion by the Malpighian tubules of *Rhodnius*. The movements of ions and water. *J. Exp. Biol*., **1969**, *51*, 71-97.

Maggin, E. J.; Elliott, J. R.: Historical perspective and current outlook for Molecular Dynamics as a chemical engineering tool. *Ind. Eng. Chem. Res*., **2010**, *49*, 3059-3078.

Manchester, K.L.: The crystallization of enzymes and virus proteins: laying to rest the colloidal concept of living systems. *Endeavour*, **2004**, *28*, 25-29.

Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Šali A.: Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophis. Biomol. Struct*., **2000**, *29*, 291-325.

McCammon, J. A.; Gelin, B. R.; Karplus, M.: Dynamics of folded proteins. *Nature*, **1977**, 267, 585-590.

McInerney, J.O., Pisani, D., Bapteste, E., O'Connell, M.J.: The public goods hypothesis for the evolution of life on Earth. *Biol Direct*., **2011**, *6*, 41.

Medeiros-Silva, M. *Papel de ureases na nodulação de 'Glycine max' L. Merr por 'Bradyrhizobium japonicum'*. Porto Alegre: UFRGS, 2012. Tese (Doutorado em Biologia Celular e Molecular), Centro de Biotecnologia do Estado do Rio Grande do Sul, Universidade Federal do Rio Grande do Sul, **2012**.

Melquiond, A.S.J.; Bonvin, A.M.J.J.: Data-driven docking: using external information to spark the biomolecular rendez-vous. In Zacharias, M. (Ed), *Protein-Protein Complexes: Analysis, Modeling and Drug Design*. Londres: Imperial College Press, **2008**.

Menegassi, A.; Wassermann, G. E.; Olivera-Severo, D.; Becker-Ritt, A. B.; Martinelli, A. H.; Feder, V.; Carlini, C. R.: Urease from cotton (*Gossypium hirsutum*) seeds: isolation, physicochemical characterization, and antifungal properties of the protein. *J. Agric. Food Chem*., **2008**, *56*, 4399-4405.

Menez, A.: Functional architectures of animal toxins: a clue to drug design? *Toxicon*, **1998**, *36*, 1557-1572.

Mertens, H.D., Svergun, D.I.: Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol*., **2010**, *172*, 128-141.

Métayer-Levrel, G. L.; Castanier, S.; Orial, G.; Loubière, J. F.; Perthuisot, J. P.: Applications of bacterial carbonatogenesis to the protection and regeneration of limestones in buildings and historic patrimony. *Sediment. Geol.*, **1999**, *126*, 25-34.

Meyer-Bothling, L. E.; Polacco, J. C.: Mutational analysis of the embryo-specific urease locus of soybean. *Mol. Gen. Genet*., **1987**, *209*, 439-444.

Mirbod, F; Schaller, R. A.; Cole, G. T.: Purification and characterization of urease isolated from the pathogenic fungus *Coccidioides immitis. Med. Mycol*., **2002**, *40*, 35-44.

Mobley, H. L.; Island, M. D.; Hausinger, R. P.: Molecular biology of microbial ureases. *Microbiol. Rev*., **1995**, *59*, 451-480.

Mulinari, F.; Stanisçuaski, F.; Bertholdo-Vargas, L. R.; Postal, M.; Oliveira-Neto, O. B.; Rigden, D. J.; Grossi-de-Sá, M. F.; Carlini, C. R. Jaburetox-2Ec: Na insecticidal peptide derived from an isoform of urease from the plant *Canavalia ensiformis. Peptides*, **2007**, *28*, 2042-2050.

Mulinari, F.; Becker-Ritt, A.B.; Demartini, D.R.; Ligabue-Braun, R.; Stanisçuaski, F.; Verli, H.; Fragoso, R.R.; Schroeder, E.K.; Carlini, C.R.; Grossi-de-Sá, M.F.: Characterization of JBURE-IIb isoform of *Canavalia ensiformis* (L.) DC urease. *Biochim. Biophys. Acta*, **2011**, *1814*, 1758-1768.

Mulrooney, S.B., Hausinger, R.P.: Nickel uptake and utilization by microorganisms. *FEMS Microbiology Reviews*, **2003**, *27*, 239-261.

Mulrooney, S.B., Ward, S.K., Hausinger, R.P.: Purification and properties of the *Klebsiella aerogenes* UreE metal-binding domain, a functional metallochaperone of urease. *Journal of Bacteriology*, **2005**, *187*, 3581-3585.

Musiani, F., Bellucci, M., Ciurli, S.: Model structures of *Helicobacter pylori* UreD(H) domains: A putative molecular recognition platform. *Journal of Chemical Information and Modeling*, **2011**, *51*, 1513-1520.

Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C.S., Stinson, J., Peters, B.A., Modrusan, Z., Jung, K., Seshagiri, S., Wu, T.D.: Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics*, **2011**, *4*, 11.

Nagano, N.; Orengo, C. A.; Thornton, J. M.: One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **2002**, *321*, 741-765.

Nakagawa, K., Tokushima, A., Fujiwara, K., Ikeguchi, M.: Proline scanning mutagenesis reveals non-native fold in the molten globule state of equine beta-lactoglobulin. *Biochemistry*, **2006**, *45*, 15468-15473.

Navarathna, D.H., Harris, S.D., Roberts, D.D., Nickerson, K.W.: Evolutionary aspects of urea utilization by fungi. *FEMS Yeast Res.*, **2010**, *10*, 209-213.

Nedelcu, A.M., Lee, R.W., Lemieux, C., Gray, M.W., Burger, G.: The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome Res.*, **2000**, *10*, 819-831.

Nikouee, A., Khabiri, M., Grissmer, S., Ettrich, R.: Charybdotoxin and margatoxin acting on the human voltage-gated potassium channel hKv1.3 and its H399N mutant: an experimental and computational comparison. *J. Phys. Chem. B*, **2012**, *116*, 5132-5140.

O'Malley, M.A., Koonin, E.V.: How stands the Tree of Life a century and a half after The Origin? *Biol Direct.*, **2011**, *6*, 32.

Olivera-Severo, D.; Wassermann, G.; Carlini, C. R.: Ureases display biological effects independent of enzymatic activity. Is there a connection to diseases caused by urease-producing bacteria? *Braz. J. Med. Bio. Res.*, **2006**, *39*, 851-861.

Oostenbrink, C., Villa, A., Mark, A.E., van Gunsteren, W.F.: A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, **2004**, *25*, 1656-1676.

Otvos, L.: Antibacterial peptides isolated from insects. *J. Pept. Sci.*, **2000**, *6*, 497-511.

Pasek, S., Risler, J.L., Brézellec, P.: Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **2006**, *22*, 1418-1423.

Peabody, D.S.: Subunit fusion confers tolerance to peptide insertions in a virus coat protein. *Arch. Biochem. Biophys*., **1997**, *347*, 85-92.

Pearson, M.A.; Michel, L.O.; Hausinger, R.P.; Karplus, P.A.: Structures of Cys319 variants and acetohydroxamate-inhibited *Klebsiella aerogenes* urease. *Biochemistry*. **1997**, *36*, 8164-8172.

Pedrozo, H. A.; Schwartz, Z.; Dean, D. D.; Wiederhold, M. L.; Boyan, B. D.: Regulation of statoconia mineralization in *Aplysia californica* in vitro. *Connect. Tissue Res*., **1996A**, *35*, 317-323.

Pedrozo, H. A.; Schwartz, Z.; Luther, M.; Dean, D. D.; Boyan, B.D.; Wiederhold, M. L.:. A mechanism of adaptation to hypergravity in the statocyst of *Aplysia californica. Hear Res*., **1996B**, *102*, 51-62.

Petoukhov, M.V., Svergun, D.I.: Applications of small-angle X-ray scattering to biomacromolecular solutions. *Int. J. Biochem. Cell. Biol*., **2013**, *45*, 429-437.

Piatigorsky, J.; Wistow, G.J.: Enzyme/crystallins: gene sharing as an evolutionary strategy. *Cell*, **1989**, *57*, 197-199.

Piatigorsky, J.: *Gene Sharing and Evolution: The Diversity of Protein Functions*. Cambridge: Harvard University Press, **2007**.

Piovesan, A.R., Stanisçuaski, F., Marco-Salvadori, J., Real-Guerra, R., Defferrari, M.S., Carlini, C.R.: Stage-specific gut proteinases of the cotton stainer bug *Dysdercus peruvianus*: role in the release of entomotoxic peptides from *Canavalia ensiformis* urease. *Insect Biochem. Mol. Biol.,* **2008**, *38*, 1023-1032.

Pires-Alves, M.; Grossi-de-Sá, M. F.; Barcellos, G. B.; Carlini, C. R.; Moraes, M. G.: Characterization and expression of a novel member (JBURE-II) of the urease gene family from jackbean [*Canavalia ensiformis* (L.) DC]. *Plant Cell Physiol*., **2003**, *44*: 139-145.

Pol-Fachin, L., Verli, H.: Structural glycobiology of the major allergen of *Artemisia vulgaris* pollen, Art v 1: O-glycosylation influence on the protein dynamics and allergenicity. *Glycobiology*, **2012**, *22*, 817-825.

Polacco. J. C.; Havir, E. A.: Comparisons of soybean urease isolated from seed and tissue culture. *J. Biol. Chem*., **1979**, *254*, 1707-1715.

Polacco, J. C.; Holland, M. A.: Roles of urease in plant cells. In Jeon, K.W.; Jarvik, J. (Eds.). *International Review of Cytology*, vol. 145. San Diego: Academic Press, **1993**.

Polacco, J.C.; Holland, M. A.: Genetic control of plant ureases. In Setlow, J.K. (Ed.). *Genetic Engineering*, vol. 16. Nova Iorque: Plenum Press, **1994**.

Polacco, J.C.; Mazzafera, P.; Tezotto, T.: Opinion: nickel and urease in plants: still many knowledge gaps. *Plant Sci*., **2013**, *199-200*, 79-90.

Ponder, J. W.; Case, D. A.: Force fields for protein simulations. In Richards, F. M.; Eisenberg, D. S.; Kuriyan, J. (Eds.). *Advances in Protein Chemistry*, vol. 66. San Diego: Elsevier Academic Press, **2003**.

Postal, M.; Martinelli, A.H.; Becker-Ritt, A.B.; Ligabue-Braun, R.; Demartini, D.R.; Ribeiro, S.F.; Pasquali, G.; Gomes, V.M.; Carlini, C.R.: Antifungal properties of *Canavalia ensiformis* urease and derived peptides. *Peptides*. **2012**, *38*, 22-32.

Prakash, S.; Chang, T.M.S.: Preparation and in vitro analysis of microencapsuleated genetically engineered *E. coli* DH5 cells for urea and ammonia removal", *Biotechnol. Bioeng*., **1995**, *46*, 621-626.

Qin, Y.; Cabral, J.M.S.: Properties and applications of urease. *Biocatalysis and biotransformation*, **2002**, *20*, 1-14.

Quiroz-Valenzuela, S.; Sukuru, S.C.; Hausinger, R.P.; Kuhn, L.A;; Heller, W.T.: The structure of urease activation complexes examined by flexibility analysis, mutagenesis, and small-angle X-ray scattering. *Arch Biochem Biophys*., **2008**, *480*, 51-57.

Ragsdale, S.W.: Nickel-based Enzyme Systems. *J. Biol. Chem*., **2009**, *284*, 18571-18575.

Rambaut, A.: FigTree. Disponível em <http://tree.bio.ed.ac.uk/software/figtree/> Acesso em: 26 jul. 2013.

Raoult, D.: The post-Darwinist rhizome of life. *Lancet*, **2010**, *375*, 104-105.

Raut SH, Sarode DD, Lele SS: Biocalcification using *B. pasteurii* for strengthening brick masonry civil engineering structures. *World J. Microbiol. Biotechnol*., **2014**, 30, 191-200.

Real-Guerra, R., Carlini, C.R., Staniscuaski, F.: Role of lysine and acidic amino acid residues on the insecticidal activity of jackbean urease. *Toxicon*, **2013**, *71*, 76-83.

Riddles, P. W.; Whan, V.; Blakeley, R. L.; Zerner, B.: Cloning and sequencing of a jack bean urease-encoding cDNA. *Gene*, **1991**, *108*, 265-267.

Ritchie, D.W.: Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci*., **2008**, *9*, 1-15.

Rodriguez-Almazan, C., Ruiz de Escudero, I., Emiliano Canton, P., Munoz-Garay, C., Perez, C., Gill, S.S., Soberon, M., Bravo, A.: The amino- and carboxyl-terminal fragments of the *Bacillus thuringensis* Cyt1Aa toxin have differential roles in toxin oligomerization and pore formation. *Biochemistry*, **2011**, *50*, 388-396.

Sachett, L.G., Verli, H.: Dynamics of different arachidonic acid orientations bound to prostaglandin endoperoxide synthases. *Eur. J. Med. Chem*., **2011**, *46*, 5212-5217.

Sacristán, M.; Millanes, A. M.; Legaz, M. E.; Vicente, C.: A lichen lectin specifically binds to the α-1,4-polygalactoside moiety of urease located in the cell wall of homologous algae. *Plant Signal. Behav*., **2006**, *1*, 23-27.

Saladin, A.; Prevost, C.: Protein-protein docking. In Zacharias, M. (Ed), *Protein-Protein Complexes: Analysis, Modeling and Drug Design*. Londres: Imperial College Press, **2008**.

Salvadori, J.D.M; Defferrari, M.S.; Ligabue-Braun, R.; Lau, E.Y.; Salvadori, J.R.; Carlini, C.R.: Characterization of entomopathogenic nematodes and symbiotic bacteria active against *Spodoptera frugiperda* (Lepidoptera: Noctuidae) and contribution of bacterial urease to the insecticidal effect. *Biological Control*, **2012**, *63*, 253-263.

Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S.: Simulating movement of tRNA into the ribosome during decoding. *Proc. Natl. Acad. Sci. U.S.A*., **2005**, *102*, 15854-15859.

Sánchez, R.; Šali, A.: Comparative protein structure modeling. Introduction and practical examples with Modeller. *Methods Mol. Biol.*, **2000**, *143*, 97-129.

Sanderson, M.J.; Shaffer, H.B.: Troubleshooting molecular phylogenetic analysis. *Ann. Rev. Ecol. Syst.*, **2002**, *33*, 49-72.

Sansubrino, A.; Mascini, M.: Development of an optical fiber sensor for the ammonia, urea, urease and IgG. *Biosensors Bioelectron.*, **1994**, *9*, 207-216.

Sayers, E.W.; Barrett, T.; Benson, D.A.; Bolton, E.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Fingerman, I.M.; Geer, L.Y.; Helmberg, W.; Kapustin, Y.; Krasnov, S.; Landsman, D.; Lipman, D.J.; Lu, Z.; Madden, T.L.; Madej, T.; Maglott, D.R.; Marchler-Bauer, A.; Miller, V.; Karsch-Mizrachi, I.; Ostell, J.; Panchenko, A.; Phan, L.; Pruitt, K.D.; Schuler, G.D.; Sequeira, E.; Sherry, S.T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T.A.; Wagner, L.; Wang, Y.; Wilbur, W.J.; Yaschenko, E.; Ye, J.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **2012**, *40*, D13-D25.

Schaack, S., Gilbert, C., Feschotte, C.: Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol*, **2010**, *25*, 537-546.

Schäfer, U. K.; Kaltwasser, H.: Urease from *Staphylococcus saprophyticus*: purification, characterization and comparison to *Staphylococcus xylosus* urease. *Arch. Microbiol.*, **1994**, *161*, 393-399.

Schagger, H., Vonjagow, G.: Tricine sodium dodecyl-sulfate polyacrylamide-gel electrophoresis for the separation of proteins in the range from 1-kDa to 100-kDa. *Anal. Biochem.*, **1987**, *166*, 368-379.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., Rost, B.: Protein disorder--a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.*, **2011**, *21*, 412-418.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J.: PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, **2005**, *33*, W363-W367.

Schneidman-Duhovny, D., Hammel, M., Sali, A.: FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.*, **2010**, *38*, W540-544.

Schneidman-Duhovny, D., Hammel, M., Šali, S.: Macromolecular docking restrained by a small angle X-ray scattering profile. *Journal of Structural Biology*, **2011**, *173*, 461-471.

Schlick, T.: *Molecular Modeling and Simulation: an Interdisciplinary Guide*. Nova Iorque: Springer, **2006**.

Schrödinger, LLC. *The PyMOL Molecular Graphics System*, Version 1.5.0.4

Schwede, T.; Šali, A.; Eswar, N.; Peitsch, M.C.: Protein Structure Modeling. In: Schwede, T.; Peitsch, M.C. (Eds). *Computational Structural Biology: Methods and Applications*. Cingapura: World Scientific Publishing Company, **2008**.

Serdyuk, I.N.; Zaccai, N.R.; Zaccai, J.: Molecular dynamics. In: _____. *Methods in Molecular Biophysics: Structure, Dynamics, Function*. Cambridge: Cambridge University Press, **2007**.

Sheridan, L., Wilmot, C.M., Cromie, K.D., van der Logt, P., Phillips, S.E.: Crystallization and preliminary X-ray structure determination of jack bean urease with a bound antibody fragment. *Acta Crystallographica Section D Biological Crystallography*, **2002**, *58*, 374-376.

Silvestro, L., Axelsen, P.H.: Membrane-induced folding of cecropin A. *Biophys. J*., **2000**, *79*, 1465-1477.

Simmons, J.E. Jr.: Urease activity in trypanorhynch cestodes. *Biological Bulletin*, **121**, *1961*, 535-546.

Singh, A.; Panting, R.J.; Varma, A.; Saijo, T.; Waldron, K.J.; Jong, A.; Ngamskulrungroj, P.; Chang, Y.C.; Rutherford, J.C.; Kwon-Chung, K.J.: Factors required for activation of urease as a virulence determinant in *Cryptococcus neoformans*. *MBio*, **2013**, *4*, e00220-13.

Sinthuvanich, C., Veiga, A.S., Gupta, K., Gaspar, D., Blumenthal, R., Schneider, J.P.: Anticancer beta-hairpin peptides: membrane-induced folding triggers activity. *J. Am. Chem. Soc*., **2012**, *134*, 6210-6217.

Sirko, A.; Brodzik, R.: Plant ureases: roles and regulation. *Acta Bioch. Pol*., **2000**, *4*, 1189-1195.

Skjaerven, L., Martinez, A., Reuter, N.: Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins*, **2011**, *79*, 232-243.

Spears, J.W.: Nickel as a "newer trace element" in the nutrition of domestic animals. *J. Anim. Sci*., **1984**, *59*, 823-835.

Stanisçuaski, F.; Ferreira-DaSilva, C. T.; Mulinari, F.; Pires-Alves, M.; Carlini, C. R.: Insecticidal effects of canatoxin on the cotton stainer bug *Dysdercus peruvianus* (Hemiptera: Pyrrhocoridae). *Toxicon*, **2005**, *45*, 753-760.

Stanisçuaski, F.; Te Brugge, V.; Carlini, C.R.; Orchard, I.: In vitro effect of *Canavalia ensiformis* urease and the derived peptide Jaburetox-2Ec in *Rhodnius prolixus* Malpighian tubules. *J. Insect Physiol*., **2009**, *55*, 255-263.

Stanisçuaski, F., Brugge, V.T., Carlini, C.R., Orchard, I.: Jack bean urease alters serotonin-induced effects on *Rhodnius prolixus* anterior midgut. *J. Insect Physiol*., **2010**, *56*, 1078-1086.

Stanisçuaski, F.; Carlini, C.R.: Plant ureases and related peptides: understanding their entomotoxic properties. *Toxins*, **2012**, *4*, 55-67.

Suhre, K., Sanejouand, Y.H.: ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **2004**, *32*, W610-614.

Sumner J.B.: The isolation and crystallization of the enzyme urease. *J. Biol. Chem*.,**1926**, *69*, 435-441.

Sumner, J. B.: The story of urease. *J. Chem. Educ*., **1937**, *14*, 255-259.

Svergun, D.I., Koch, M.H.: Advances in structure analysis using small-angle scattering in solution. *Curr. Opin. Struct. Biol*., **2002**, *12*, 654-660.

Sydor, A.M.; Zamble, D.B.: Nickel metallomics: general themes guiding nickel homeostasis. *Met. Ions Life Sci*., **2013**, *12*, 375-416.

Takishima, K.; Suga, T.; Mamiya, G.: The structure of jack bean urease. The complete amino acid sequence, limited proteolysis and reactive cysteine residues. *Eur. J. Biochem*., **1988**, *175*, 151-165.

Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S.: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731-2739.

Tange, Y.; Niwa, O.: Identification of the *ure1+* gene encoding urease in fission yeast. *Curr. Genet*., **1997**, *32*, 244-246.

Tautz, D., Domazet-Lošo, T.: The evolutionary origin of orphan genes. *Nat. Rev. Genet*., **2011**, *12*, 692-702.

Terra, W.R.: Evolution of digestive systems of insects. *Annu. Rev. Entomol*., **1990**, *35*, 181-200.

Thauer, R. K.: Nickel to the fore. *Science*, **2001**, *293*, 1264-1265.

Tomazetto, G., Mulinari, F., Staniscuaski, F., Settembrini, B., Carlini, C.R., Ayub, M.A.Z..: Expression kinetics and plasmid stability of recombinant *E. coli* encoding urease-derived peptide with bioinsecticide activity. *Enzyme Microb. Technol.*, **2007**, *41*, 821-827.

Tompa, P.: Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci*., **2012**, *37*, 509-516.

Tonegawa, S.: Somatic generation of antibody diversity. *Nature*, **1983**, *302*, 575-581.

Towbin, H., Staehelin, T., Gordon, J.: Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets. Procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A*., **1979**, *76*, 4350-4354.

Uberti AF, Olivera-Severo D, Wassermann GE, Scopel-Guerra A, Moraes JA, Barcellos-de-Souza P, Barja-Fidalgo C, Carlini CR.: Pro-inflammatory properties and neutrophil activation by *Helicobacter pylori* urease. *Toxicon*, **2013**, *69*, 240-249.

Uversky, V.N.: Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*., **2002**, *11*, 739-756.

Uversky, V.N.; Dunker, A.K.: Understanding protein non-folding. *Biochim. Biophys. Acta*, **2010**, *1804*, 1231-1264.

von der Haar, T., Tuite, M.F.: Regulated translational bypass of stop codons in yeast. *Trends Microbiol*, **2007**, *15*, 78-86.

Wadhwani, P., Reichert, J., Bürck, J., Ulrich, A.S.: Antimicrobial and cell-penetrating peptides induce lipid vesicle fusion by folding and aggregation, *Eur. Biophys. J*., **2012**, *41*, 177-187.

Wafford, K.A., Sattelle, D.B.: Effects of amino acid neurotransmitter candidates on an identified insect motoneurone. *Neurosci. Lett*., **1986**, *63*, 135-140.

Wang, G.; Xia, Y.; Cui, J.; Gu, Z.; Song, Y.; Chen, Y.Q.; Chen, H.; Zhang, H.; Chen, W.: The roles of moonlighting proteins in bacteria. *Curr. Issues Mol. Biol.*, **2013**, *16*, 15-22.

Wassermann, G. E.; Olivera-Severo, D.; Uberti, A. F.; Carlini, C.R.: *Helicobacter pylori* urease activates blood platelets through a lipoxygenase-mediated pathway. *J. Cell. Mol. Med*., **2010**, *14*, 2025-2034

Whelan, S.; Goldman, N.: A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol*. **2001**, *18*, 691-699.

Witte CP, Tiller S, Isidore E, Davies HV, Taylor MA.: Analysis of two alleles of the urease gene from potato: polymorphisms, expression, and extensive alternative splicing of the corresponding mRNA. *J. Exp. Bot*., **2005**, *56*, 91-99.

Witte, C.P.: Urea metabolism in plants. *Plant Sci*., **2011**, *180*, 431-438.

Woese, C.R., Kandler, O., Wheelis, M.L.: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA,* **1990**, *87*, 4576-4579.

Wöhler, F.: Ueber künstliche bildung des harnstoffs. *Annalen der Physik und Chemie*, **1828**, *88*, 253-256.

Xue, J.Y.; Zhang, S.C.; Liu, N.G.; Liu, Z.H.: Verification, characterization and tissue-specific expression of UreG, a urease accessory protein gene, from the amphioxus *Branchiostoma belcheri. Acta Biochim. Biophys. Sin*., **2006**, *38*, 549-555.

Yamamoto, M., Nakagawa, K., Ikeguchi, M.: Importance of polypeptide chain length for the correct local folding of a beta-sheet protein, *Biophys. Chem*., **2012**, *168*, 40-47.

Yang, L., Song, G., Jernigan, R.L.: How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.*, **2007**, *93*, 920-929.

Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., Abraham, P.E., Lankford, P.K., Adams, R.M., Shah, M.B., Hettich, R.L., Lindquist, E., Kalluri, U.C., Gunter, L.E., Pennacchio, C., Tuskan, G.A.: Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res*., **2011**, *21*, 634-641.

Yang, Z.; Rannala, B.: Molecular phylogenetics: principles and practice. *Nat. Rev. Genet*., **2012**, *13*, 303-314.

Yu, J. J.; Smithson, S. L.; Thomas, P. W.; Kirkland, T. N.; Cole, G. T.: Isolation and characterization of the urease gene (*URE*) from the pathogenic fungus *Coccidioides immitis. Gene*, **1997**, *198*, 387-391.

Zambelli, B.; Musiani, F.; Savini, M.; Tucker, P.; Ciurli, S.: Biochemical studies on *Mycobacterium tuberculosis* UreG and comparative modeling reveal structural and functional conservation among the bacterial UreG family. *Biochemistry*, **2007**, *46*, 3171-3182.

Zambelli, B.; Musiani, F.; Benini, S.; Ciurli, S.: Chemistry of Ni^{2+} in urease: sensing, trafficking, and catalysis. *Acc. Chem. Res*., **2011**, *44*, 520-530.

Zambelli, B., Banaszak, K., Merloni, A., Kiliszek, A., Rypniewski, W., Ciurli, S.: Selectivity of Ni(II) and Zn(II) binding to *Sporosarcina pasteurii* UreE, a metallochaperone in the urease

assembly: a calorimetric and crystallographic study. *J. Biol. Inorg. Chem*., **2013**, *18*, 1005-1017.

Zerner B.: Recent advances in the chemistry on an old enzyme, urease. *Bioorg. Chem*., **1991**, *19*, 116-131.

Zonia, L. E.; Stebbins, N. E.; Polacco, J. C.: Essential role of urease in germination of nitrogen-limited *Arabidopsis thaliana* seeds. *Plant Physiol*., **1995**, *107*, 1097-1103.

REFERÊNCIAS PARA AS CITAÇÕES

Crichton, M. Introduction: What kind of world do we live in? In Crichton, M.; Preston, R. *Micro*. Nova Iorque: Harper, **2011**.

Fifield, W.: Pablo Picasso: A Composite Interview. *The Paris Review*, 32, Summer-Fall, **1964**.

Medawar, P.B.: Is the scientific paper fraudulent? *The Saturday Review*, ago. 1, 42-43, **1964**.

Preston, R.: Introduction: Adventures in nonfiction writing. In:_____. *Panic in level 4: Cannibals, killer viruses, and other journeys to the edge of science*. Nova Iorque: Random House, **2009**.

Russell, B.: Free thought and official propaganda. In:_____. *Sceptical essays*, **1928**.

Sumner, J. B.: The story of urease. *J. Chem. Educ*., **1937**, *14*, 255-259.

Zerner, B.: Recent advances in the chemistry of an old enzyme, urease. *Bioorganic chemistry*, **1991**, *19*, 116-131.

9. APÊNDICES

Apêndice A

Tabela A: Estruturas 3D de ureases depositadas no Protein Data Bank (http://www.rcsb.org/pdb).

PDB ID	Organismo-fonte*	Comentário	Resolução (Å)
1FWJ	*Klebsiella aerogenes*	Nativa	2,20
1EJR	*Klebsiella aerogenes*	Mutante D221A	2,00
1EJS	*Klebsiella aerogenes*	Mutante H219N	2,00
1EJT	*Klebsiella aerogenes*	Mutante H219Q	2,00
1EJU	*Klebsiella aerogenes*	Mutante H320N	2,00
1EJV	*Klebsiella aerogenes*	Mutante H320Q	2,40
1EJW	*Klebsiella aerogenes*	Tipo selvagem a 298K	1,90
1EJX	*Klebsiella aerogenes*	Tipo selvagem a 100K	1,60
1EF2	*Klebsiella aerogenes*	Substituição por manganês	2,50
1A5K	*Klebsiella aerogenes*	Mutante K217E	2,20
1A5L	*Klebsiella aerogenes*	Mutante K217C	2,20
1A5M	*Klebsiella aerogenes*	Mutante K217A	2,00
1A5N	*Klebsiella aerogenes*	Mutante K217A, quimicamente resgatado por formato e níquel	2,40
1A5O	*Klebsiella aerogenes*	Mutante K217C, quimicamente resgatado por formato e níquel	2,50
1FWA	*Klebsiella aerogenes*	Mutante C319A em pH 7,5	2,00
1FWB	*Klebsiella aerogenes*	Mutante C319A em pH 6,5	2,00
1FWC	*Klebsiella aerogenes*	Mutante C319A em pH 8,5	2,00
1FWD	*Klebsiella aerogenes*	Mutante C319A em PH 9,4	2,00
1FWE	*Klebsiella aerogenes*	Mutante C319A, com ácido acetohidroxâmico ligado	2,00
1FWF	*Klebsiella aerogenes*	Mutante C319D	2,00
1FWG	*Klebsiella aerogenes*	Mutante C319S	2,00
1FWH	*Klebsiella aerogenes*	Mutante C319Y	2,00
1FWI	*Klebsiella aerogenes*	Mutante H134A	2,00
1KRA	*Klebsiella aerogenes*	Apoenzima	2,30
1KRB	*Klebsiella aerogenes*	Mutante H219A	2,50

1KRC	*Klebsiella aerogenes*	Mutante H320A	2,50
4EPD	*Klebsiella aerogenes*	Danificada por radiação a 300 K, estrutura inicial	1,70
4EPE	*Klebsiella aerogenes*	Danificada por radiação a 300 K, estrutura final	2,05
4EP8	*Klebsiella aerogenes*	Danificada por radiação a 100 K, estrutura inicial	1,55
4EPB	*Klebsiella aerogenes*	Danificada por radiação a 100 K, estrutura final	1,75
2UBP	*Bacillus pasteurii*	Nativa	2,00
1UBP	*Bacillus pasteurii*	Inibida por beta-mercaptoetanol	1,65
3UBP	*Bacillus pasteurii*	Inibida por diamidofosfato	2,00
4UBP	*Bacillus pasteurii*	Inibida por ácido acetohidroxâmico	1,55
1S3T	*Bacillus pasteurii*	Inibida por borato	2,10
1IE7	*Bacillus pasteurii*	Inibida por fosfato	1,85
4A7C	*Bacillus pasteurii*	Complexada com citrato	1,50
1E9Z	*Helicobacter pylori*	Nativa	3,00
1E9Y	*Helicobacter pylori*	Complexada com ácido acetohidroxâmico	3,00
3QGA	*Helicobacter mustelae*	Nativa	3,00
3QGK	*Helicobacter mustelae*	Nativa, sem solvente ordenado	3,00
2FVH	*Mycobacterium tuberculosis*	Subunidade Gamma	1,80
4FUR	*Brucella melitensis*	Subunidade Gamma	2,10
3LA4	*Canavalia ensiformis*	Nativa	2,05
4H9M	*Canavalia ensiformis*	Nativa	1,52
4GY7	*Canavalia ensiformis*	Nativa	1,49
4GOA	*Canavalia ensiformis*	Inibida por fluoreto	2,20
4G7E	*Cajanus cajan*	Nativa	2,20

***** alguns deles sofreram revisão taxonômica: *Klebsiella aerogenes* é sinônimo *de Enterobacter aerogenes*; *Bacillus pasteurii* é sinônimo de *Sporosarcina pasteurii*; *Helicobacter mustelae* é sinônimo de *Campylobacter mustelae*.

Apêndice B

Venomous mammals: a review.

Ligabue-Braun R, Verli H, Carlini CR.

Toxicon, 2012, 59, 680-595.

O artigo a seguir foi elaborado considerando o interesse pessoal exercido pelo tema e a ausência de trabalhos elencando avanços recentes na área. O mesmo recebeu a distinção “Top 25 Hottest Articles”, por ter sido o terceiro trabalho mais acessado no periódico Toxicon em 2012.

Novas descobertas foram feitas após a publicação desta revisão. Em especial, dois artigos caracterizando melhor a draculina, um inibidor dos fatores IXa e Xa da cascata de coagulação, foram publicados quase simultaneamente, enquanto um novo inibidor (desmolaris) foi identificado:

Francischetti IM, Assumpção TC, Ma D, Li Y, Vicente EC, Uieda W, Ribeiro JM. The "Vampirome": Transcriptome and proteome analysis of the principal and accessory submaxillary glands of the vampire bat *Desmodus rotundus*, a vector of human rabies. *J Proteomics*. **2013**, 82, 288-319.

Low DH, Sunagar K, Undheim EA, Ali SA, Alagon AC, Ruder T, Jackson TN, Pineda Gonzalez S, King GF, Jones A, Antunes A, Fry BG. Dracula's children: molecular evolution of vampire bat venom. *J Proteomics*. **2013**, 89, 95-111.

Ma D, Mizurini DM, Assumpção TC, Li Y, Qi Y, Kotsyfakis M, Ribeiro JM, Monteiro RQ, Francischetti IM. Desmolaris, a novel Factor XIa anticoagulant from the salivary gland of the vampire bat (*Desmodus rotundus*) inhibits inflammation and thrombosis in vivo. *Blood,* **2013** (no prelo).

Além destes, um artigo extenso foi publicado acerca do mais enigmático dos mamíferos venenosos, o lóris. Na referida publicação são feitas propostas arrojadas quanto ao possível mimetismo dos lóris, buscando simular serpentes do gênero *Naja*.

Nekaris KA, Moore RS, Rode EJ, Fry BG. Mad, bad and dangerous to know: the biochemistry, ecology and evolution of slow loris venom. *J Venom Anim Toxins Incl Trop Dis*., **2013**, 19, 21.

Finalmente, mais evidências fósseis foram analisadas, levando à conclusão de que não é possível alegar um “passado venenoso” para mamíferos com base nos mamíferos venenosos atuais:

Folinsbee KE. Evolution of venom across extant and extinct eulipotyphlans. *Comptes Rendus Palevol*, **2013** (no prelo).

Toxicon 59 (2012) 680–695

Contents lists available at SciVerse ScienceDirect

Toxicon

journal homepage: www.elsevier.com/locate/toxicon

Review

Venomous mammals: A review

Rodrigo Ligabue-Braun [a], Hugo Verli [a,b], Célia Regina Carlini [a,c,*]

[a] Graduate Program in Cellular and Molecular Biology, Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
[b] Center of Biotechnology and Faculty of Pharmacy, UFRGS, Porto Alegre, RS, Brazil
[c] Department of Biophysics-IB, UFRGS, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 14 October 2011
Received in revised form 19 January 2012
Accepted 21 February 2012
Available online 3 March 2012

Keywords:
Shrew
Vampire bat
Solenodon
Mole
Platypus
Loris
Venom components
Toxic effects

ABSTRACT

The occurrence of venom in mammals has long been considered of minor importance, but recent fossil discoveries and advances in experimental techniques have cast new light into this subject. Mammalian venoms form a heterogeneous group having different compositions and modes of action and are present in three classes of mammals, Insectivora, Monotremata, and Chiroptera. A fourth order, Primates, is proposed to have venomous representatives. In this review we highlight recent advances in the field while summarizing biochemical characteristics of these secretions and their effects upon humans and other animals. Historical aspects of venom discovery and evolutionary hypothesis regarding their origin are also discussed.

1. Introduction

The fact that mammals can be venomous was neglected by mainstream science for centuries, and only recently has received an increasing interest in literature. Such lag in the unravel of mammalian venoms may be a reflection of advances on experimental techniques, previously unavailable for precise characterization of individual molecules, combined with the discovery of fossil mammals with putative envenomation apparatus. Additionally, this renewed attention may also draw from a review article written by Dufton (1992), describing venomous properties of the saliva in the mammalian order Insectivora. Besides these Insectivora representatives (some shrew species and *Solenodon* spp.), venom is also found in the orders Monotremata (crural gland in the male platypus), Chiroptera (salivary glands in vampire bats), and arguably in Primates (brachial gland in slow and pygmy lorises).

When dealing with venomous animals, one has to comply with a definition for it. Not surprisingly, there is no single definition for venom or venomous animal. Bücherl (1968) states that venomous animals must possess at least one venom gland, a mechanism for excretion or extrusion of the venom, as well as apparatus with which to inflict wounds. When distinguishing venomous from poisonous, Mebs (2002) states that venomous animals produce venom in a group of cells or gland, and have a tool, the venom apparatus, which delivers the venom by injection during a bite or sting. The venom apparatus in this definition encompasses both the gland and the injection device, which must be directly connected. Mebs (2002) also defines venoms as compounds that are deleterious to another organism at certain dosage, being composed mainly by proteins or peptides. For the present review we

* Corresponding author. Dept Biophysics & Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500, Prédio 43431, Campus do Vale, 91501-970, Porto Alegre, RS, Brazil. Tel.: +55 51 3308 7606; fax: +55 51 3308 7603.
E-mail addresses: rodrigobraun@cbiot.ufrgs.br (R. Ligabue-Braun), hverli@cbiot.ufrgs.br (H. Verli), ccarlini@ufrgs.br (C.R. Carlini).

0041-0101/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.toxicon.2012.02.012

follow the broader definition proposed by Fry et al. (2009). According to this definition, a venom is a secretion produced in a specialized gland in one animal and delivered to a target animal through the infliction of a wound. This secretion must contain molecules that disrupt normal physiological processes so as to facilitate feeding or defense by the producing animal. Additionally, according to these authors, the feeding secretion of hematophagous specialists may be regarded as a specialized subtype of venom.

Even though they have different compositions and act by different means, all mammalian venoms share the same history of disregard by science. For centuries there has been a widespread belief that mammals could be as venomous as reptiles (Dufton, 1992). This belief, however, remained overlooked by orthodox mammalogists and was treated as folklore (Pournelle, 1968). Considering the renaissance of interest on mammalian venoms, in the next sections we will outline some historical aspects related to these animals and their venoms, review the pharmacological effects of these secretions while pointing out several difficulties associated with their study, and address how the venom relates to the animals' habits. Based on such panorama, previous theories regarding the origin and importance of venom(s) in mammals will be discussed.

2. Insectivora venom

2.1. Introduction

The taxonomically complex group Insectivora (Lipotyphla) holds most of the venomous mammals. With the exception of vampire bats, these are the only mammals so far observed to produce toxic saliva. The American short-tailed shrew (*Blarina brevicauda*), the Hispaniolan solenodon (*Solenodon paradoxus*), the European water shrew (*Neomys fodiens*) and the Mediterranean water shrew (*Neomys anomalus*) provided the most definite evidences for salivary venom (Pournelle, 1968). All four species have significantly enlarged and granular submaxillary salivary glands from which the toxic saliva is produced (Dufton, 1992). Preliminary studies suggest that the Cuban solenodon (*Solenodon cubanus*) and the Canarian shrew (*Crocidura canariensis*) also produce venomous saliva (Gundlach, 1877; Lopez-Jurado and Mateo, 1996). Suspicion remains upon untested insectivores, including the American shrew (*Sorex cinereus*) and the European mole (*Talpa europaea*). Moles have large and granular submaxillary glands and are known for storing worms in "paralyzed" state in their burrows (Dufton, 1992). This hoarding habit is also present in *B. brevicauda* (Martin, 1981; Merrit, 1986). Hedgehogs were once considered venomous, but subsequent tests revealed no toxicity in their saliva (Mebs, 1999).

2.2. Historical background

One of the first records of venomous insectivores is from a classic frightening description of the European shrew in Historie of Foure-Footed Beasts: "*It is a ravening beast, feigning itself gentle and tame, but being touched it biteth deep, and poisoneth deadly. It beareth a cruel mind, desiring to hurt anything, neither is there any creature it loveth*" (Topsell, 1607). English beliefs associated 'shrew' with matters of depravity, wickedness, evil, ill omen and malignancy (Dufton, 1992). Even the Latin name for the European shrew, *Sorex araneus*, encompasses the idea that shrews would have poisonous bites like a spider (*aranea* in latin). The belief that the bite of a shrew is highly poisonous is usually associated with Old World folklore, but is also common in certain regions of the United States (Pournelle, 1968).

The West Indian natives have long considered the Hispaniolan solenodon (*S. paradoxus*) (Fig. 1) as having a poisonous bite (Pournelle, 1968). In 1877, J. Gundlach described the effects of almiqui (Cuban solenodon, *S. cubanus*) bites and made some comparisons with those from venomous snakes (Gundlach, 1877). Some years later, C.J. Maynard described in detail the effects of a bite on the hand that he had received from an American short-tailed shrew (*B. brevicauda*) (Maynard, 1889). His report was largely ignored and no other reliable record of adverse

Fig. 1. General aspect of a Hispaniolan solenodon (*Solenodon paradoxus*) (photograph by Guillermo Armenteros).

reactions from shrew bites emerged for fifty years (Pournelle, 1968).

The presence of toxic substances in the saliva of insectivores found new evidences when Pearson (1942) injected mice and rabbits with extracts from submaxillary glands of *B. brevicauda*. Many studies followed, including other shrew species and the Hispaniolan solenodon. Despite the renewed interest on these venomous mammals, no paper was written on the subject from the 1960s to the 1990s, when M.J. Dufton published a major review including his own research data on Insectivora venoms (Dufton, 1992). In the 2000s this subject was once again on the spotlight, propelled by the purification and characterization of blarina toxin from *B. brevicauda* saliva (Kita et al., 2004), and by the discovery of fossil shrews with envenomation apparatus (Cuenca-Bescós and Rofes, 2007).

2.3. Venom effects on humans and other animals

European folk-tales focused on the effects of shrew bites upon cattle and horses, describing the affected animals as paralyzed and deprived of feeling, hinting to systemic effects. The same tales did not consider shrews as a serious threat to humans, who experienced only discomfort after the bite (Dufton, 1992).

Gundlach, in his description of the effects of a Cuban solenodon bite, reported that the lower incisor punctures elicited inflammation at the site of the wound, while the upper incisors did not, adding that intense inflammation was regarded by the natives as a common effect of these bites (Gundlach, 1877).

In his account of a *B. brevicauda* bite to his hand, C.J. Maynard described an immediate burning sensation at the wound, swelling, impossibility of using the affected hand for three days due to intense pain, and a week-long discomfort after that (Maynard, 1889). A cautionary article was published in 1969, indicating the potential hazard imposed by *B. brevicauda* venom, especially for children (Chadwick, 1969). Many mammalogists, however, reported receiving bites from these shrews without ill effects (Pournelle, 1968).

Comprehensive assays regarding the effects of *Blarina* (Pearson, 1942; Tomasi, 1978), *Neomys* (Pucek, 1968) and *Solenodon* (Rabb, 1959) submaxillary extracts were carried out on mice, rabbits and cats. The symptoms were generally similar, with a sequence of general depression, breathing disturbance, paralysis and convulsions. Dosage and route of administration also play important roles, with intracerebral and intravenous injections being far more effective than intraperitoneal or subcutaneous (for a detailed review on the pharmacology of these venoms see Dufton, 1992).

Despite the importance of small vertebrate prey (mice and voles) in some insectivore diets, it is the invertebrate contribution that makes up most of their dietary intake. Assays with insect prey (crickets and roaches) and *B. brevicauda* showed that its venom had an immobilizing effect upon the insects and that the immobilized preys were stored by the shrew for later consumption (Martin, 1981). In its natural habitat, *B. brevicauda* is reported to cache a variety of animal preys in comatose state, including earthworms, insects, snails and small mammals (Merrit, 1986).

2.4. Toxic components

For many years the difficulty in obtaining submaxillary gland material, in adequate quantities and freshness, was considered the main reason for the lack of progress in the study of Insectivora venom. However, some other factors also contributed on this regard (Dufton, 1992), including the difficulty in maintaining shrews in captivity and the endangered status of the much larger Cuban and Hispaniolan solenodons (Soy and Mancina, 2008; Turvey and Incháustegui, 2008), which prohibits any invasive studies with these animals.

Most of the studies concerning the toxic principle of this venom were based on very impure mixtures obtained from *B. brevicauda* submaxillary glands (Pucek, 1968; Dufton, 1992). The active element targeted in almost all studies was a putative neurotoxin that would be held responsible for the observed effects on vertebrates and invertebrates. Similarities between the effects of shrew and cobra venoms (Lawrence, 1945) encouraged the search for the major enzymatic components of the ophidian venom in the insectivore saliva, but no significant resemblance was found (Dufton, 1992).

The purification of the toxic component of the *B. brevicauda* saliva, blarina toxin (BLTX), was achieved in 2004 (Kita et al., 2004). The active mature BLTX is a glycosylated (N-glycosylations at Asn80 and Asn93) protein composed of 253 amino acids with a kallikrein-like protease activity. This toxin cleaves kininogens producing kinins, including bradykinin, an inflammation mediator which increases vascular permeability and lowers blood pressure. These kinins may be the main toxic agents related to Insectivora venom, explaining some of the symptoms observed experimentally, such as dyspnea, hypotension and hypokinesia (Kita et al., 2004). Blarinasin, another tissue kallikrein-like protease from *B. brevicauda* salivary glands, is highly similar to BLTX, but is not toxic to mice, suggesting that minor differences (such as differential glycosylation) may be important for BLTX toxicity (Kita et al., 2005). Additionally, BLTX is microheterogeneous regarding its glycosylation state (occurring as similar-sized proteins with different pI values). The contribution of other, undefined toxic components that would act synergistically with BLTX in the shrew saliva is not discarded (Kita et al., 2004).

The LD_{50} value for intraperitoneal injection of purified BLTX into male mice was calculated to be 1 mg per kg (Kita et al., 2004). Previous attempts to establish a LD_{50} for partially purified toxin from *Blarina* in mice reached the value of 3.4 mg per kg (Pucek, 1968). For rabbits, the LD_{50} for fresh extract of salivary glands was 5.85–7.90 mg per kg, and for partially purified toxin was 1.50–1.95 mg per kg (Pucek, 1968). Regarding the toxicity of *N. fodiens* saliva, 20 mg per kg were lethal for rabbits. Intraperitoneal injection of 21 mg per 20 g of body weight of *Microtus agrestis* (field vole) produced only a slight reaction, while intracerebral injection of 0.2–0.4 mg per 20 g was lethal to these animals. For mice, intracerebral injection of 0.5–1.0 mg per 20 g of body weight was lethal (Pucek, 1968). The extracts of submaxillary glands of *Solenodon* were lethal when injected intravenously into white mice at doses

of 0.38–0.55 mg per g of body weight (death ensued in 2–6 min). Intraperitoneal injection of 0.56–0.66 mg per g killed mice in 12 h (Rabb, 1959).

2.5. Biological and evolutionary aspects

Long before the blarina toxin characterization, there was the proposition of similarities between Insectivora and reptilian venoms. Most of these inferred similarities associated the venom of shrews and solenodons with snake venoms (Gundlach, 1877; Lawrence, 1945). In his attempt to identify the toxic components of shrew saliva, Pearson argued that these animals have modified salivary glands which produce venom in a way that closely resembles snake venoms. He also observed, however, that snakes have their parotid glands modified to produce venom, while shrews have their submaxillary glands as a venom source (Pearson, 1942). Venom-producing submaxillary glands are found in only two lizards, the Gila monster (*Heloderma suspectum*) and the Mexican beaded lizard (*Heloderma horridum*). The link between the venoms from *Heloderma* spp. and shrews was briefly proposed by Lawrence (1945) but was only confirmed after the characterization of BLTX (Kita et al., 2004). BLTX is similar to gila toxin (GTX) and horridum toxin (33.6% and 32.4% identical, respectively), two toxins from the Mexican beaded lizard (Utaisincharoen et al., 1993; Datta and Tu, 1997). GTX and BLTX also share similar functions and effects upon prey (Kita et al., 2004). These toxins also have similar, nonhomologous residue insertions (Fig. 2). The relevance of these insertions was asserted recently, by means of structural modeling and molecular dynamics simulations (Aminetzach et al., 2009). These authors were able to predict structural modifications that would increase the catalysis of non-toxic kallikreins to BLTX levels. Predicted alterations included increased flexibility, differences in loop length and polarity, as well as general differences in surface charges. BLTX matches all the predicted criteria for the transition from a non-toxic to a toxic kallikrein. However, the most remarkable observations of these authors is that GTX also made the transition from non-toxic to toxic kallikrein by comparable means, with locally different alterations leading to a globally similar structures (Aminetzach et al., 2009). This is one outstanding example of convergent evolution at the molecular level, with two independently evolved serine protease venoms converging to highly similar protein structures (Aminetzach et al., 2009; Brodie, 2010).

Fig. 2. Lower jaw of a *S. paradoxus* showing the characteristic grooved second lower incisor (arrow) (photograph from specimen USNM 537794, housed at the Department of Paleobiology, National Museum of Natural History, Smithsonian Institution).

Other important aspect of Insectivora venom concerns its function. There is an ongoing debate on this subject, defined as “hunting big or hoarding small” by Furió et al. (2010). It is well known that shrews cache various preys in a comatose state, including earthworms, insects, snails, and to a lesser extent, small mammals such as voles and mice (Pournelle, 1968; Merrit, 1986). This habit is part of an adaptive winter profile, which includes utilization of elaborate nests, foraging confined to a stable thermal regime and reduced activity during periods of cold (Merrit, 1986). In this context, the venom would act as a tool to sustain a living hoard, thus ensuring food supply when capturing prey is difficult (e.g. cold winter). This is especially significant considering the high metabolic rate of shrews (Pournelle, 1968). Arguments against this theory sustain that the venom is used as a tool to hunt bigger prey. Dufton (1992) proposed that insectivores have an enhanced dependence on vertebrate food material, which is larger and more dangerous than their power to weight ratio would allow, thus requiring an extra asset to overcome these difficulties. Extant shrews do not have specialized venom delivery apparatus. Their teeth do not have channels, but a concavity on the first incisors may collect and transmit saliva from the submaxillary ducts, which open near the base of these teeth (Pournelle, 1968). The discovery of an extinct giant shrew (*Beremendia* sp.) with an envenomation apparatus (Cuenca-Bescós and Rofes, 2007) injected fuel in the “hunting big or hoarding small” debate. The giant *Beremendia* (three to four times larger than the extant *N. fodiens*) had specialized lower incisors with a gutter-like grove along the medial side of the crown (Cuenca-Bescós and Rofes, 2007; Rofes and Cuenca-Bescós, 2009), resembling the grooved incisors of *Solenodon* spp. (Dufton, 1992) (Fig. 3) where the grooves act conducting the saliva from the salivary gland to the wound inflicted onto the prey. Cuenca-Bescós and Rofes (2007) argue that the envenomation apparatus was an evolutionary adaptation, probably related to increase in body mass and hunting of larger-sized prey. The venomous nature of the species was reassessed by Furió et al. (2010), with different implications. In a paleoenvironmental reconstruction study, these authors concluded that the *Beremendia* diet did not rely on larger prey, but on coleopterans and gastropods, and the venom was required for hoarding live prey. This behavior would be a reflection of the highly unpredictable environment where these giant shrews lived (Furió et al., 2010). These contradicting explanations point to the absence of consensus on this subject. It is likely that insectivores may still take advantage of venom by both means, including combining them to hoard larger prey. It is also interesting to consider venom as a weapon for intraspecific competition. Rabb (1959) observed that death was frequent among Hispaniolan solenodons kept together in the same enclosure, with bite marks on their feet being the only observable cause. Such use in competition may be a secondary aspect of the insectivore venom.

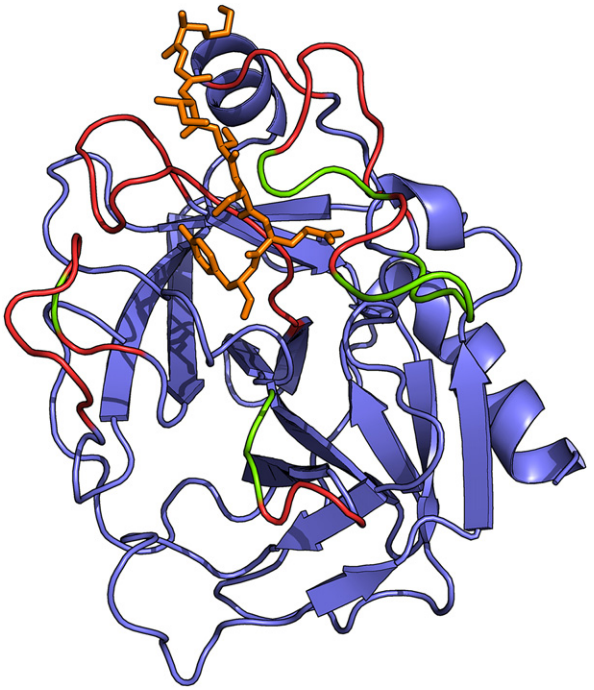


Fig. 3. Structural model of BLTX (built based on PDB ID 2ZCK, Ménez et al., 2008). Regulatory loops are colored in red, and the insertions discussed in the text are colored in green. A fluorogenic substract analog is shown in orange. Depicted based on Aminetzach et al. (2009). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Monotremata venom

3.1. Introduction

The platypus (*Ornithorhynchus anatinus*) is thought to be the only venomous Monotremata representative (Fig. 4). Suspicion remains upon echidnas, but no proper venom has been ascribed to these mammals so far. Platypus are found in fresh water rivers and streams in most of the eastern coast of Australia (Grant and Temple-Smith, 1998). These semi-fossorial, semi-aquatic, egg-laying mammals have amused people since their discovery in 1798. Among their uncommon features is the possession of venom-delivering keratinized spurs in their hind legs (Fig. 5). Both males and females are born with these spurs, but the latter lose them during development (Calaby, 1968). The spurs are connected to the venom-producing crural glands, forming the crural system (Whittington and Belov, 2007). During the mating season these glands become highly active, producing venom to be delivered by the channeled spur. In echidnas, both males and females seem to have degenerates spurs (in a way similar to the female platypus), but the male glands show cyclic activity as observed in male platypus (Krause, 2009).

3.2. Historical background

There are less than twenty records of platypus envenomation in literature, but at least one case is reported each year to the specialized agencies in Australia (Tonkin and Negrine, 1994). The first account on platypus envenomation was made in the early 1800s (Jamison, 1818), followed by a detailed analysis on the anatomy and use of the venom glands (Spicer, 1876), and experimentation with the venom effect upon test animals and tissues (Martin and Tidswell, 1895; Kellaway and LeMessurier, 1935). Thirty years later, Calaby reviewed the venomous characteristics of the platypus (Calaby, 1968). Attacks on humans started to reappear in specialized literature twenty years later (Petrie and Pearn, 1987). Fenner et al. (1992) published the most detailed case report on platypus envenomation. This was followed by the last report of platypus attack (Tonkin and Negrine, 1994). Considering the (apparently) facetious editor's note that accompanies this last report, platypus envenomation was not considered a real threat to humans by the medical community.

The interest on Monotremata venom increased from 1995 onwards, following the first “modern” biochemical investigation of its components (de Plater et al., 1995). At least four major components were identified since (as reviewed by Whittington and Belov, 2007). The Platypus

Fig. 4. General aspect of a duck-billed platypus (*Ornithorhynchus anatinus*) (photograph by Caleb McElrea).

Fig. 5. Male platypus hind-leg skeleton with crural spur (arrow) (photograph from a specimen housed at the Museum of Life Sciences, King's College London).

Genome Project (Warren et al., 2008) allowed a most in-depth analysis of the venom.

3.3. Venom effects on humans and other animals

Platypus envenomation was fairly common when the animal was still hunted for its fur (Calaby, 1968). Nowadays any close contact with the animal is rare and restricted to biologists, zookeepers and fishermen (who occasionally catch them in lines or nets) (Whittington and Belov, 2007).

When attacking, the hind legs of the animal are driven toward one another with considerable force so that the spurs are embedded in the flesh caught between and if venom is being produced, a few milliliters are injected by repeated jabbing (Calaby, 1968; Whittington and Belov, 2007). The spurs have enough strength to support the weight of the platypus, which often hangs from the victim, requiring assistance for removal (Fenner et al., 1992). In humans, all recorded strikes have been on the hand or wrist. Envenomation results in immediate and acute pain and swelling. General first-aid procedures and drugs such as morphine are ineffective in relieving the pain. The only effective treatment is localized anesthetic blockade supplemented by intravenous narcotic infusion (Fenner et al., 1992). The symptoms have duration of two weeks to several months. It is also noteworthy that the envenomation aftermath is dependent on the season and gland usage prior to the attack (Calaby, 1968). Considering this variation, there can be no effect at all (Kellaway and LeMessurier, 1935) to excruciating pain "worse than shrapnel wounds" (Fenner et al., 1992). No fatalities have been recorded (Tonkin and Negrine, 1994).

Attacks of platypus to hunting dogs were somewhat common (Martin and Tidswell, 1895). Most dogs were struck on the face, with their head becoming swollen and apparently sore afterward, but death was infrequent (Calaby, 1968). Attacks among platypus are seldom observed, but males are often found in the wild with punctures in their bodies during the mating season (Grant and Temple-Smith, 1998).

The earliest experimental studies on platypus venom included its effects upon test animals (Martin and Tidswell, 1895; Kellaway and LeMessurier, 1935). When gland secretion was injected subcutaneously into a rabbit it produced localized swelling and tenderness. When injected intravenously into rabbits ($n = 3$), a rapid fall in blood pressure and respiratory distress were observed, being followed by death. Two animals probably died from extensive intravascular coagulation. The third animal, in which the material was injected more slowly, died somewhat later than the others. Its blood did not exhibit signs of clotting in the vessels and a sample of this blood was found to clot abnormally slowly (Martin and Tidswell, 1895). When Kellaway and LeMessurier (1935) injected the material obtained from Martin and Tidswell (1895) intravenously into two rabbits, it resulted in severe dyspnea. Fresh material produced a rapid and profound fall in blood pressure and death when intravenously injected into rabbits.

3.4. Toxic components

Studies on the composition of platypus venom are restricted by two major reasons. The first reason is the limited amount of venom that is produced exclusively by males during the breeding season. The second reason is the difficulty in obtaining tissue and venom samples from wild animals, which are protected (de Plater et al., 1998a; Whittington and Belov, 2007). Nevertheless, the availability of new protein characterization techniques from the 1990s onwards allowed identification of some of the venom components.

The *O. anatinus* venom is a complex mixture of 19 different fractions. The peptide fractions include C-type natriuretic peptides (CNPs), defensin-like peptides (DLPs), nerve growth factors (NGFs), isomerases, hyaluronidase, protease, and uncharacterized proteins (de Plater et al., 1995; Kourie, 1999a; Whittington and Belov, 2007). The platypus venom transcriptome identified 88 putative venom genes when filtered against transcriptomes of non-venomous tissues (Whittington et al., 2010). These genes were analogous to venom genes from other species, including reptiles, insectivores, fish, and invertebrates.

Natriuretic peptides take part in the control of blood pressure, exhibiting hypotensive and vasorelaxant properties. In mammals, three classes have been described: atrial natriuretic peptide (ANP, predominantly produced by the cardiac atria), brain natriuretic peptide (BNP, isolated from brain but predominantly produced by the cardiac ventricules) and C-type natriuretic peptide (CNP, found in brain and endothelium). CNPs lack natriuretic activity, suggesting a unique physiological role for these peptides (de Plater et al., 1998a; de Plater et al., 1998b). OvCNPs are CNPs found in the *O. anatinus* venom that are homologous to CNP-53 and CNP-22 (de Plater et al., 1998a). They are the most biologically active peptides in the venom, and may be responsible for the systemic signs produced by envenomation (e.g. hypotension) (Whittington and Belov, 2007). When injected into rats, OvCNPs causes edema and histamine release from mast cells, also causing relaxation of isolated rat uterus and vas deferens (de Plater et al., 1998b). OvCNPs can also form cation-selective channels in cell membranes (Kourie, 1999b), inducing calcium influx in neuroblastoma cells (Kita et al., 2009). Whole venom was also able to induce calcium-dependent currents in cultured dorsal root ganglion cells from rats, suggesting Ca^{2+} release from intracellular stores and involvement of tyrosine or serine–threonine kinases (de Plater et al., 2001). This effect seems to be related to putative nociceptors. The interaction

of venom peptides with membranes and its disruption of ion transport pathways are consistent with the observed envenomation symptoms (Whittington and Belov, 2007).

Defensin-like peptides constitute a family of four peptides structurally similar to anti-microbial β-defensins found in mammals, with the main structural feature being six paired cysteine residues (Fig. 6) (Torres et al., 1999, 2000). β-defensins take part in innate immunity, disrupting bacterial cell membranes and acting as chemokines, being produced by epithelial cells and neutrophils (Torres and Kuchel, 2004). These functions are not shared with DLPs, possibly due to absence of sequence similarity between these peptide groups (Torres et al., 1999). In view of the fact that DLPs are the most abundant peptides in the *O. anatinus* venom, it is proposed that they could produce the venom-induced pain, either alone or synergistically with the less-studied venom NGFs (Torres et al., 2000; Whittington and Belov, 2007). The latter are proposed to have immunogenic effects instead of a putative neuronal action (Whittington and Belov, 2009).

One intriguing aspect of the platypus venom is the presence of different forms of peptides that are identical in sequence. This variation was first shown for OvCNPa and OvCNPb (de Plater et al., 1998a) and later for DLP-2 and DLP-4 (Torres et al., 2005). The reason for such difference is the unusual presence of a D-amino acid in the second position of OvCNPb and DLP-2 (Torres et al., 2002, 2005). The residue conversion agent was later identified as an L-to-D-amino-acid-residue isomerase (Torres et al., 2006). It is unclear whether D-amino acids confer functional differences to the venom peptides, but protease resistance whilst in the crural gland has been proposed (Torres et al., 2002, 2006).

3.5. Biological and evolutionary aspects

Despite being only part of a bizarre animal, the crural system was considered one of the most mysterious anatomical features of the platypus, eliciting many theories to explain its function (Calaby, 1968). Such theories, as reviewed by Grant and Temple-Smith (1998), included the use of the crural system to climb river banks, collect nesting materials, water-proof the fur, hold the female during copulation or act as a defensive weapon. Most of the evidence now supports the proposition of Martin and Tidswell (1895), that the apparatus is used by males on one another as a weapon when competing for females, taking part in sexual selection. It was observed that testicles and crural glands of the platypus increase and decrease in size before and after the mating season. During this season males become more aggressive and are found with punctures in their bodies, especially in the tail region. Adult males also largely avoid each other. These observations led Grant and Temple-Smith (1998) to propose a polygynous mating system for platypuses, with male–male interactions directing the access to females and thus justifying the development and retention of the crural apparatus in males.

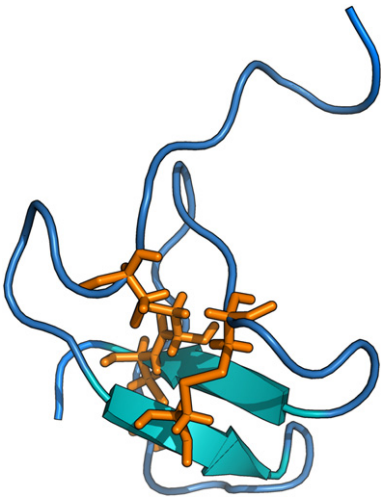


Fig. 6. Structure of DLP-2 from *O. anatinus* venom (PDB ID 1ZUE, Torres et al., 2002). The three disulphide bridges are depicted in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Even though homologous, the crural system of echidnas is much less studied than that of platypuses. Spicer (1876) reported lack of reaction from a male echidna that was constantly touched, even with "considerable pressure" near the crural spur, while Calaby (1968) concluded that the crural gland in these animals is "never very active". Recently, Krause (2009) performed a comparative histochemical and morphological study with echidna and platypus crural systems, providing insights to this overlooked subject. As observed for platypus, male short-beaked echidnas (*Tachyglossus aculeatus*) are seasonal in their reproductive activities, and their crural glands undergo changes consistent with their levels of sexual hormones. These animals, however, lack aggressive use of the spurs and have a poor mechanism for locking the spur to the tibia during an attack, thus not supporting the use of the apparatus as a weapon. Nonetheless, the seasonality of the crural glands suggests a putative role for them as scent glands. Since nothing is known about the secretion produced by echidnas, its function remains unclear (Krause, 2009).

Recently available fossil evidences suggest that crural spurs may have been a widespread feature in primitive mammals. According to this proposition, primitive mammals used their venom-delivering apparatus as a defensive structure while sharing habitats with dinosaurs (Hurum et al., 2006; Kielan-Jaworowska and Hurum, 2006). This proposition, however, relies on comparison with extant animals since it is not possible to ascertain venom-delivery based solely on fossil evidence (Orr et al., 2007; Folinsbee et al., 2007).

As observed for Insectivora, Monotremata venom also shares many features with reptilian venoms via convergent evolution (Whittington et al., 2008). Duplications of β-defensin, CNP and NGF gene families and subsequent co-option of these non-toxic homologs for venomous activities have occurred independently in mammals and reptiles during the evolution of their venoms (Warren et al., 2008). It is also interesting to note that many of the venom genes, including DLP-B, DLP-C, OvCNP and NGF, are also expressed in non-venom tissues, whereas DLP-A is the only peptide exclusive of venom tissues. OvCNP and NGF are also expressed in female tissues. Such expression pattern suggests broader roles for these peptides (Whittington and Belov, 2009).

4. Chiroptera venom

4.1. Introduction

Following our working definition of venom (Fry et al., 2009), which regards the feeding secretions of hematophagous specialists as a particular subtype of venom, the subfamily Desmodontinae represents the venomous mammals from the order Chiroptera. This group comprises the most well known venomous bat, *Desmodus rotundus* (common vampire bat) (Fig. 7), and two other rare species, *Diphylla ecaudata* (hairy-legged vampire bat) and *Diaemus youngi* (white-winged vampire bat); all of them generally found from Mexico to southern Argentina (Tellgren-Roth et al., 2009). These bats produce toxic saliva with anticoagulant properties and have a series of anatomical and physiological modifications to allow nourishment based solely on blood (Schondube et al., 2001). Except for chickens, that may die of hemorrhage after being bitten by vampire bats, the majority of their prey do not perish from the attack or contact with the venom. Thus, one can argue that the vampire bat saliva is not a true venom, since it causes only minor discomfort for the prey. As observed by Delpietro and Russo (2009), the feeding behavior of vampire bats resembles that of parasites. Their venom, which facilitates feeding by disruption of normal physiological processes of the prey, ensures survival of the latter for continuous supply of nutrients for the bats.

4.2. Historical background

The term "vampire" is much older than the European knowledge of hematophagous bats, but their discovery inspired much elaboration upon the traditional use of the word. Of Slavonic origin, the term *Vampyre* is widely associated to supernatural blood-sucking beings in Eastern Europe (Ditmars and Greenhall, 1935). Such supernatural vampire is considered to be the soul of a dead person that leaves the dead body at night, assuming any of many forms to suck blood of sleeping people and animals. It is interesting to point out that the original tale did not mention bats as possible shapes of such entity. The vampire tradition of Eastern Europe met the hematophagous bats and incorporated them sometime after the arrival of Columbus in America. When he discovered the Isle of Trinidad in 1498, he also discovered the hematophagous bats of the New World (Málaga-Alba, 1954). Later reports from the XVI century onwards somewhat exaggerated the feeding habits of the animal. Francisco Montejo, when disembarking in the peninsula of Yucatan in 1527, was victim of a "great plague of bats" that attacked men and animals "sucking their blood when they were asleep" (Málaga-Alba, 1954). Cortez became aware of Camazotz, the cruel vampire bat god of the Maya mythology, later reporting it to Spain (Ditmars and Greenhall, 1935). After the return of the early explorers of the New World tropics, a "vampire epidemics" broke out in Europe about 1730 (Klinger, 2008). In 1801 Felix Azara, a Spanish naturalist, described "the biter" in Paraguay, and Charles Darwin, in 1832, captured a vampire bat on the back of a horse in Chile (Málaga-Alba, 1954). The association of bats and vampirism grew, summiting in the classic portrait of Dracula, despite the fact that bats are only mentioned twice in the novel (Klinger, 2008).

Many of the discoveries regarding the habits of hematophagous bats derive from studies on their role as vectors of trypanosomiasis and rabies (Dunn, 1932; Málaga-Alba, 1954). Based on the observation of freely bleeding wounds inflicted by vampire bats, many researchers proposed that some anticoagulant agent should be present in the bats' saliva. In 1966, a plasminogen activator was identified (Hawkey, 1966). In 1991, four of these activators were characterized at the molecular level (Krätzschmar et al., 1991), and later an inhibitor of fIXa and fXa was also identified (Apitz-Castro et al., 1995).

4.3. Venom effects on humans and other animals

Feeding bites produced by vampire bats have characteristic shape and location. The large, sharp upper incisor teeth (Fig. 8) produce a crater-like, sharply circumscribed wound, with a diameter of approx. 4 mm. The bite is inflicted on the bare skin with the victim's hair being either combed or parted by the bat (Málaga-Alba, 1954; Delpietro

Fig. 7. General aspect of a common vampire bat (*Desmodus rotundus*) (photograph by Trisha Shears).

Fig. 8. Skeleton of *D. rotundus* face showing the sharp upper incisors (photograph by Antonio Sebben).

and Russo, 2009). The blood is licked up by a piston-like motion of the tongue for about 30 min (Ditmars and Greenhall, 1935). The wound they produce not only freely bleeds when first inflicted, but also continues to bleed for several hours after the bat has ceased to eat (Hawkey, 1966, 1967). In a comparison made by DiSanto (1960), a wound produced by *D. rotundus* bleeds for 3–8 h after infliction, whereas a similar knife wound ceases bleeding after 15 min.

D. rotundus feeds mainly on domesticated cattle, horses, goats, pigs and sheep. To a lesser extent they feed on poultry, wild prey and humans (Delpietro and Russo, 2009). *D. ecaudata* feeds on birds, and *D. youngi* feeds on mammals but prefers birds (Tellgren-Roth et al., 2009). The prey is preferably asleep when attacked, minimizing the bat exposure to risks. The attack is slow and delicate in nature, with the bat avoiding any chance of being perceived by the prey. The bites are frequently described as painless, with the victim hardly waking up when being attacked (Ditmars and Greenhall, 1935). Besides the residual hemorrhaging, the vampire bat venom also elicits immune response on preys frequently fed upon (Delpietro and Russo, 2009).

4.4. Toxic components

Two distinct classes of anticoagulants are found in the saliva of vampire bats i.e. plasminogen activators and inhibitors of proteinases (Zavalova et al., 2002; Basanova et al., 2002). Plasminogen activators act producing localized proteolysis in tissue remodeling, wound healing and neuronal plasticity (Schleuning, 2001). The tissue type plasminogen activators (t-PA) are serine proteases that cleave the plasmin proenzyme to its active form, which in turn is responsible for degradation of blood clots (Tellgren-Roth et al., 2009). After long speculations and preliminary results (as reviewed by DiSanto, 1960), a plasminogen activator was identified in the saliva of *D. rotundus* in 1966 (Hawkey, 1966) and extensively studied ever since (Hawkey, 1967; Cartwright, 1974). The later cloning and expression of this activator revealed high similarity to t-PA. It was also found that this activator is not a single protein, but a family of four *D. rotundus* salivary plasminogen activators (DSPAs) (Krätzschmar et al., 1991). This is not the case for the other vampire bats, which have only one t-PA in their genomes (Tellgren-Roth et al., 2009). The t-PA molecule has five domains: finger (F), epidermal growth factor (EGF), kringle 1 (K1), kringle 2 (K2) and serine protease (P). F and K2 domains are both involved in the competitive binding of fibrin as well as its inhibitor plasminogen inhibitor 1 and DD(E), a product of degraded cross-linked fibrin (Stewart et al., 1998). PA shows activity as a single chain, but this activity is highly enhanced as a cleaved two chain form (Tellgren-Roth et al., 2009). With the exception of *D. ecaudata* PA, the plasminogen activators from vampire bats have smaller chains with domain deletions when compared to t-PA. *D. youngi* PA lacks K2 domain. The four variants of DSPAs (α1, α2, β, γ) lack the K2 domain and a plasmin-sensitive activation site. DSPA-β additionally lacks F domain, while DSPA-γ lacks F and EGF domains (Fig. 9). These differences alter many binding properties of vampire bat PAs when compared to t-PA. The loss of K2 domain leads to decreased sensitivity to fibrinogen, plasminogen activator inhibitor, and DD(E). Absence of K2 domain also leads to strong fibrin specificity. The loss of plasmin-activation site allowed DSPAs to be active as single chains (Gardell et al., 1989; Krätzschmar et al., 1991; Bode and Renatus, 1997; Renatus et al., 1997; Tellgren-Roth et al., 2009). Based on its extraordinary specificity for fibrin, DSPA-α1 became a candidate for thrombolytic therapy, being renamed as desmoteplase in medical literature (Schleuning, 2001). Glycosylation studies revealed two N-glycosylation sites (more than 30 different oligosaccharides linked to Asn117 and Asn362) and one O-glycosylation site (L-fucose linked to Thr61) in DSPA-α1. Despite equivalent to glycosylation sites in t-PA, there seem to be differences between the glycosydic structures linked to these sites, especially for Asn117. N-glycans in this position of DSPA-α1 are more processed during biosynthesis that the high-mannose counterpart in t-PA (Gohlke et al., 1996, 1997). This difference is pointed as a possible reason for the fourfold lowered clearance rate observed for DSPA-α1 when compared to t-PA (Witt et al., 1994).

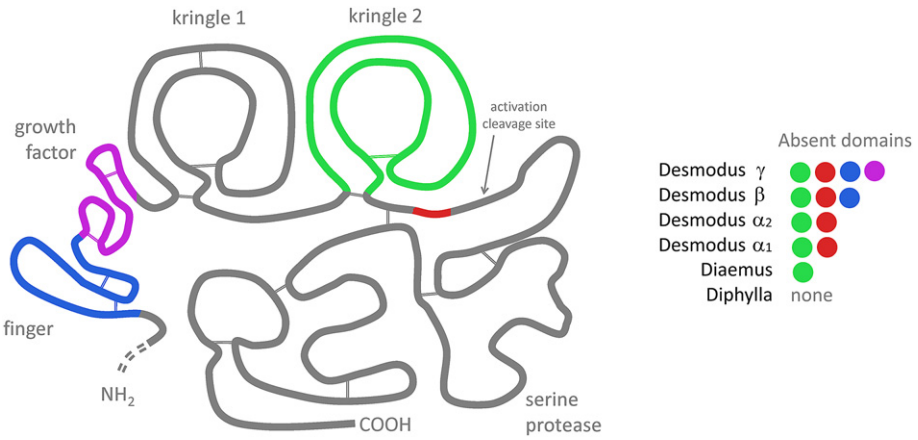


Fig. 9. Topological scheme of the plasminogen activators from vampire bats. Absent domains in each protein are color-coded. Disulphide bridges are depicted as thin double lines. Drawn based on Tellgren-Roth et al. (2009). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The other anticoagulant found in the saliva of vampire bats (so far only studied in *D. rotundus*) is draculin, an inhibitor of activated factors X (fXa) and IX (fIXa) (Apitz-Castro et al., 1995). fXa is the only enzyme in the coagulation cascade to convert prothrombin into thrombin, the key enzyme in this process, and the activation of factor X is considered a common point between intrinsic and extrinsic pathways of blood coagulation (Hemker and Béguin, 1995). Draculin is unique among compounds known to inhibit this enzyme, since it acts as a noncompetitive, tight-binding inhibitor of fXa (Fernandez et al., 1999). The biological activity of draculin is dependent on adequate N- and O-glycosylation and stress inflicted to the bat (e.g. frequent induced salivation) impairs this modification (for a detailed description of the glycosyl composition of draculin see Fernandez et al., 1998). The protein loses activity at rates proportional to the deglycosylation degree. It is suggested that native draculin is secreted as a mixture of glycoforms that modulate its final anticoagulant activity. The differential glycosylation associated to the noncompetitive inhibition of fXa are proposed to take part in the fIXa inhibition, a mechanism that is less understood (Fernandez et al., 1999).

The feeding bites of vampire bats are regarded as painless, different from their defensive bites. Although this might indicate an additional anesthetic in the saliva, Ditmars and Greenhall (1935) suggest that young bats learn to inflict these bites, and that it seems to involve trial and error. No study on the putative anesthetic has been performed so far.

4.5. Biological and evolutionary aspects

Vampire bats are highly specialized mammals, with their entire physiology modified to use blood as their only source of food and water. They have a T-shaped gastroesophageal-duodenal junction with a tubular stomach, contrasting with regular mammalian gastrointestinal tracts. This allows the ingested blood to first enter the intestine and then overflow into the stomach (Mitchell and Tigner, 1970; Machado-Santos et al., 2009). The stomach functions as a site for blood storage and water absorption. After ingesting 50% of its weight in blood, most of the water is eliminated by "instant diuresis", and the highly nitrogenous remains of blood are then processed with very little water. For this, vampire bats have a high capacity for urea concentration in their urine, being considered physiologically equivalent to desert mammals (Breidenstein, 1982). They have reduced metabolism during digestion of the highly proteinaceous meal, upon which they depend entirely. Maltase and sucrase are absent in their gastrointestinal tracts, and dietary carbohydrates seem to be almost unused by these bats (Schondube et al., 2001).

The vampire bats also lack storage fat tissue, requiring to daily feed on blood. To do so, the bats have modified sharp teeth, anticoagulants in their saliva and a specialized tongue. The blood is not sucked as normally depicted, but rather lapped up in a piston-like movement of the tongue. For this method to work, the blood must be liquid and flowing. The evolution of anticoagulants in the saliva of the three different vampire bat species reveals transitions in their preferred preys. As discussed by Tellgren-Roth et al. (2009), gene duplication, domain loss and sequence evolution altered the fibrin specificity and susceptibility to plasminogen activator inhibitor 1. The monophyletic Phillostomidae family, where Desmodontinae is included, holds the greatest diversity in feeding habits of all mammalian families, including frugivory, carnivory and hematophagy, the latter being derived from insectivory (Schondube et al., 2001). In the subfamily, *D. ecaudata* has a single copy of plasminogen activator similar to those in other mammals. This vampire bat feeds only upon birds. The more generalist *D. youngi*, which feeds upon birds and mammals, lost the K2 domain, increasing fibrin specificity. In *D. rotundus* rounds of gene duplication and domain loss led to an active form of plasminogen activator with decreased sensitivity to plasminogen activator inhibitor 1, enhancing the ability to feed solely on mammalian blood.

It is interesting to point out that vampire bats are still considered pests, especially for cattle (Delpietro and Russo, 2009). Since they can carry rabies viruses and also transmit equine trypanosomiasis (Dunn, 1932; Málaga-Alba, 1954), vampire bats have been exterminated by many methods, including shooting, gassing, poisoning, netting, trapping and dynamiting (Thompson et al., 1972). There is now great awareness about controlling the vampire bat population while minimizing the impact on other species, including nonhematophagous bats that share the same habitats.

5. Arguably venomous mammals

5.1. Introduction

The previous three sections presented mammals from different orders which satisfy (even if partially, as is the case of vampire bats) the criteria to be considered venomous. Nonetheless, there are some intriguing representatives of the Primates order that have a different way of toxin delivery. In these animals, there is no venom apparatus (as defined by Mebs, 2002) since their venom is produced and injected by unrelated body parts. These primates, the slow (*Nycticebus coucang*, *Nycticebus bengalensis*) and the pygmy slow (*Nycticebus pygmaeus*) lorises (Fig. 10), are nocturnal arboreal prosimians that inhabit the forests of Southeast Asia and Western Indonesia. By not complying with most of the definitions of venomous animals, they are kept in the "twilight zone" of toxinology labeling. These mammals are considered indirectly or secondarily venomous, since their venom is produced in the brachial gland, licked, and injected into the victim by the specialized tooth comb (Hagey et al., 2007). The brachial gland, also termed brachial organ (Krane et al., 2003) or antecubital organ (Ahmed and Ling, 1976), is a specialized apocrine sweat gland located in the ventral side of the elbow (Hagey et al., 2007), and the tooth comb is a set of spaced incisors on the lower jaw, primarily used for grooming (Fitch-Snyder and Schulze, 2001). As discussed below, these animals can be considered "actively venomous" (to employ an out-fashioned but still informative term from Bücherl (1968), to differentiate them from poisonous animals), but the absolute lack of physical

Fig. 10. General aspect of a slow loris (*Nycticebus coucang*) (photograph by David Haring, Duke Lemur Center).

connection between the brachial gland and the tooth comb makes it hard for lorises to fit in most of the definitions of a venomous animal. Additionally, this "venom" is toxic only for some incidentally susceptible species (Hagey et al., 2007).

5.2. Historical background

Slow lorises are regarded as venomous animals by the northern Thai folklore (Wilde, 1972). In spite of that, there are only two reports in the medical literature regarding loris attacks. The first report was published in the 1970s, presenting a case of anaphylaxis on an adult man bitten by a pet slow loris (*N. coucang*) (Wilde, 1972). The second report was published in the late 2000s, presenting a pygmy slow loris (*N. pygmaeus*) bite inflicted to a pregnant woman (Kalimullah et al., 2008). Some additional cases are reported on recent manuals for zookeepers (Fitch-Snyder and Schulze, 2001).

The venom of a loris reaches the target animal by means of prolonged bites. It was originally thought that the animal's saliva was responsible for the observed symptoms, causing anaphylaxis on already sensitized individuals (Wilde, 1972). Observations of captive lorises revealed that whenever disturbed by handling or capture their brachial glands secrete about 10 μL of a clear, strong-smelling liquid (Hagey et al., 2007). This secretion is licked by the lorises and frequently wiped against their heads. Alterman (1990) established a connection between this habit and the painful loris bites, proposing that toxins from the brachial gland exudate (BGE) would be responsible for the observable effects on humans. A decade later, Krane et al. (2003) isolated the major component of the BGE of *N. coucang*, which exhibits high sequence similarity to Fel d 1, the major allergen from the domestic cat (*Felis catus*).

5.3. Venom effects on humans and other animals

Nycticebus bites have a wide variety of effects upon humans, from none to death, with most of the reported cases resembling allergic reactions. In his report of a pet slow loris bite on an adult man, Wilde (1972) described a series of symptoms. The animal bit the wrist of the patient, clinging there for several seconds. Five minutes after the bite, the man noted burning pain on hands and feet, followed by pulsating backache. Thirty minutes after the bite he presented anaphylaxis symptoms, characterized by hypotension, cyanosis of the extremities, and hematuria. The patient recovered completely after two weeks. Thai tribes consider that lorises have a poisonous bite which causes excruciating pain and frequently death (Wilde, 1972). Researchers studying lorises readily develop allergies to the glandular secretions (Hagey et al., 2007), adding to the picture of bite-induced anaphylactic shock. On the other hand, Kalimullah et al. (2008) reported no ill effects from a slow pygmy loris (*N. pygmaeus*) bite to a pregnant zookeeper who was allergic to several substances and who had being exposed to the animal before.

Little is known about predation of lorises and the putative use of biting as a defense mechanism against predators (Hagey et al., 2007). Captive lorises, on the other hand, commonly inflict bites onto enclosure mates. The resulting wounds are severe, affecting a large area with loss of fur, prolonged edema, are slow-healing, and often life-threatening (Fitch-Snyder and Schulze, 2001; Hagey et al., 2007). Partially purified extracts from *N. coucang* BGE were lethal to mice upon subcutaneous injection, even when diluted 228,000-fold (Alterman, 1990). However, there is no other report of exudate testing on animals so far.

5.4. Toxic components

The BGE is a complex mixture, comprising volatile low molecular weight metabolites and non-volatile high molecular weight protein fractions (for a detailed description of the secretion components see Hagey et al., 2007). An unidentified steroid (Alterman, 1989) and polypeptides generated by mixing BGE and saliva (Alterman and Hale, 1991) were suggested to be the active toxic component of the venom.

The major component of the BGE was isolated by Krane et al. (2003) and extensively characterized by Hagey et al. (2007). This new member of the secretoglobin family is a heterodimeric protein with 17.6 kDa. The α-chain (7.8 kDa) and β-chain (9.8 kDa) are linked by two disulfide bridges. These two chains exhibit high sequence similarity with the two chains of Fel 1 d, the major allergen from domestic cat (*F. catus*) (Leitermann and Ohman, 1984; Morgenstern et al., 1991). The BGE protein also shares the disulfide-bridged dimeric structure with Fel d 1, the only

Table 1
Summary of venomous and arguably venomous mammals and their venoms.

Mammalian order	Venomous representatives	Venom source	Known toxic components	Most common effects
Insectivora	Shrews and solenodons	Salivary gland	Blarina toxin	Breathing disturbance, paralysis and convulsions
Monotremata	Platypus	Crural gland	C-type natriuretic peptides, defensin-like peptides, nerve growth factors	Acute pain and swelling
Chiroptera	Vampire bats	Salivary gland	Plasminogen activators, draculin	Prolonged bleeding
Primates	Slow and pygmy slow lorises	Brachial gland	BGE protein	Allergic reactions

other allergen with this structure discovered so far (Krane et al., 2003; Kaiser et al., 2003). All three venomous loris species (*N. coucang*, *N. bengalensis*, and *N. pygmaeus*) have two protein isoforms (Krane et al., 2003; Hagey et al., 2007). The β-chain of Fel 1 d also occurs in two forms as a result of alternative splicing, alternative initiation or expression of different alleles (Griffith et al., 1992). The epitope specificity of the slow loris allergen and its comparison to those of Fel 1 d remain to be examined, but sequence similarity suggests immunogenic cross-reactivity between these proteins (Valenta et al., 1996; Krane et al., 2003). The variable sensitivity to loris bites and the onset of anaphylaxis support the hypothesis that *Nycticebus* BGE protein may act as an allergen (Krane et al., 2003).

5.5. Biological and evolutionary aspects

Lorises have a unique mechanism of toxin use, loading their modified teeth with the exudate of a gland that is not related to the oral cavity, in a manner that is unrelated to any other venomous animals. The use of such sophisticated method of toxin delivery seems incompatible with the feeding habits of lorises or their defense against predators (Hagey et al., 2007). The diet of *Nycticebus* consists of fruits, invertebrates and small vertebrates (Wilde, 1972; Fitch-Snyder and Schulze, 2001), none of which require special killing or immobilizing methods. Orangutans, pythons, and hawk-eagles are known for capturing and killing lorises, indicating that biting is not effective in fending off predators (Utami & van Hooff, 1997; Wiens and Zitzmann, 1999; Hagey et al., 2007). The toxin thus may take part in other behaviors of these animals.

The proposition of the brachial gland as the source of loris toxins has been constantly associated to a putative dual role for its secretion. According to this hypothesis, the secretion would constitute an interspecific defense mechanism and an intraspecific communication system (Alterman, 1990). This hypothesis is strongly supported by Hagey et al. (2007). Lorises (and other nocturnal prosimians) are specialized in olfactory signals. The vigorous urine marking and sniffing behavior, and the extensive smearing of the BGE on head and neck hair, which dries to amber-colored crystals, indicate that urine and BGE make up most of their scent repertoire. The characteristic *Nycticebus* odor is not noticeable to researchers trying to catch the animal, and lorises only stop licking the gland after the encounter is over (Hagey et al., 2007). The use of BGE thus seems inadequate as a quick response to predators, and may either act repelling the predator or warning other lorises of the danger, or perhaps both. The complex composition of BGE associated to individual and social grooming behaviors that employ the tooth comb reinforce the use of such exudate as a signaling device (Ehrlich and Musicant, 1977). The species-specificity of BGE composition also point to communicatory use of this substance. Predictions of the BGE protein structure indicate that this protein may act as box, entrapping important signaling molecules from BGE and saliva (Hagey et al., 2007). Despite the constant focus on the brachial gland and its secretion, the salivary glands of lorises may also take part in the proposed communication system. Unlike other primates, the parotid and submandibular glands of slow lorises have giant secretory granules of unknown function (Tandler et al., 1996). It is reasonable to consider the participation of this secretion as an additional source of informative molecules employed in loris communication.

It is rather tragic that lorises are threatened by their illegal trade as pets and for using in traditional medicine. When sold alive, most of these animals have their tooth combs broken or removed with nail clippers, frequently leading to stress, infection, and death. Major efforts are being made to avoid the illegal trade of lorises, along with all wildlife trade that occurs in their home range (Fitch-Snyder and Schulze, 2001; Starr et al., 2010).

6. Concluding discussion

In the mid 1900s, Pearson argued that reports on venomous mammals would most probably be due to bacterial infections related to the animal bites rather than to toxic compounds produced by the animals (Pearson, 1942). For this reason he also speculated that antibiotics alone would take care of all supposed effects of the venom. It took fifty years for this picture to change, despite many pioneering studies in the field. In his extensive paper, Dufton (1992) proposed that venom delivery, despite rare in extant mammals, would be widespread in primitive mammals and that such feature was still poorly understood due to the lack of fossil evidences. A decade later this foresight seemed fulfilled. In 2005, Fox & Scott described canine grooves in the fossil *Bisonalveus browni*, probably taking part in a venom delivery. It soon became debatable whether canine grooves alone could be considered as evidence of a venom delivery apparatus. Orr et al. (2007) and Folinsbee et al. (2007) argued that many extant mammals (including coatis, lemurs, anthropoid primates, hippopotamuses, and peccaries) have canine grooves but no venom, with the grooves serving mainly for structural purposes. These authors concluded that the comparative method alone could not ascribe venom delivery to any

fossil. However, Cuenca-Bescós and Rofes (2007) found definite evidences for a venom delivery apparatus on extinct shrews, thus validating the hypothesis of venom in primitive mammals. Hurum et al. (2006) and Kielan-Jaworowska and Hurum (2006) made similar discoveries on extinct mammals with crural spurs, which could be involved in venom delivery as seen in the platypus. It is still difficult to ascertain whether venom delivery was common in primitive mammals, but it is now accepted that it was present in some of them. It was recently proposed that venom delivery had multiple origins in West Indian insectivores. Turvey (2010) found evidences for canine grooves in the extinct *Nesophontes* sp., which may point to different kind of venom channel in this mammal when compared to other insectivores. This finding could corroborate the hypothesis of venom delivery as a typical trait in primitive mammals.

In general, humans are generally not the prey and are not perceived as predators of venomous animals (Hodgson, 1997) and the rarity of cases and variety of effects in mammalian-induced envenomation do not justify the development of antivenoms, which would be the most cost-effective treatment for envenomation (Brown and Landon, 2010). Despite not posing a threat to public health, this special group of venoms has helped in the understanding of molecular processes and provided many scaffolds for biotechnological applications. The platypus venom is promising both as a tool to understand mechanisms of pain perception, and as a model to design new pain relievers, especially for long lasting, treatment-unresponsive pains (Koh et al., 2009; Whittington et al., 2009). Desmoteplase, one of the vampire bat anticoagulants, is in the final evaluation stages for the treatment of myocardial infarction, pulmonary thromboembolism, and stroke (Baruah et al., 2006; Paciaroni et al., 2009). Toxin-based drug discovery is now growing hand in hand with venomics (as reviewed by Calvete, 2009), and mammalian venoms thus offer great opportunities to readily employ new knowledge to people's benefit.

It is interesting to note that toxic components of the mammalian venoms are examples of convergent evolution, sharing many characteristics with venoms from other animal groups. One of the key features of venom is that its efficacy is dependent on its capacity to produce rapid effect (e.g. fast prey immobilization or quick pain). This is one of the proposed constraints that limit the groups of proteins that can be enrolled as toxin candidates (for an in-depth discussion on convergent evolution of toxins, see Fry et al., 2009).

The study of these very unusual venoms (summarized in Table 1) also led to unexpected discoveries regarding other aspects of mammalian physiology. Once considered very restricted in occurrence, L-to-D-amino-acid isomerases were found on the platypus venom and later in heart muscles from mice (Koh et al., 2010), indicating that many more occurrences of such enzymes in mammals may have been overlooked. Remnants of the brachial gland exudate protein from lorises are still present in human and chimp genomes in the form of pseudogenes (Hagey et al., 2007), pointing to a putative venomous past for apes. Male mice were shown to produce salivary kallikrein that is toxic to rats and guinea-pigs but harmless to mice (Hiramatsu et al., 1980). As observed by Mebs (2002), kallikrein is rather common among mammals, being present in human saliva, for example. Only tiny variations (such as small residue insertions and differential glycosylation) separate the regular, physiological kallikrein from its toxic counterpart. It is a remarkable example of a physiologically important enzyme being recruited for a toxic role. As has been shown in this review article, there are many unanswered questions regarding mammalian venoms. "Classic" assays to ascertain toxicity are still to be performed for many of these compounds, for which only anecdotal evidence of toxicity is available.

All these challenging findings show that many features of mammalian behavior and physiology are still overlooked, mainly because we are too focused on the expected, ignoring alternative possibilities. Even humans might be considered venomous, since they master many different toxins and their applications (Mebs, 2002). The very statement that mammals are venomous would be taken as heretic some years ago. Considering all the possible applications of the mammalian venoms, it is now time to re-evaluate what we think as a venomous animal and what danger it can pose to us. Not surprisingly, we have much more to learn from than to fear these creatures.

Acknowledgments

The authors thank Guido Lenz, Ph.D., for the careful proof reading of the original manuscript; the photographers who authorized the use of their work in this article; and the anonymous reviewer for many provocative inquisitions and insightful suggestions which much improved the final version of this review.

Ethical statement

The authors declare to have agreed upon the content and form of the manuscript, which has not been published previously nor is it being submitted, entire or parts of it, to any other journal. No experiments with live organisms were conducted to provide data for this work.

Funding source

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior), FAPERGS (Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Ahmed, M.M., Ling, E.A., 1976. The antecubital organ of the primate slow loris (*Nycticebus coucang coucang*). Tissue Cell 8, 335–344.
Alterman, L., 1989. Analysis of organic extracts of brachial gland exudate from *Nycticebus coucang*. Am. J. Primatol. 18, 132.
Alterman, L., 1990. Isolation of toxins from brachial gland exudates from *Nycticebus coucang*. Am. J. Phys. Anthropol. 81, 187.

Alterman, L., Hale, M.E., 1991. Comparisons of toxins from brachial gland exudates of *Nycticebus coucang* and *N. pygmaeus*. Am. J. Phys. Anthropol. 12 (Suppl.), 43.

Aminetzach, Y.T., Srouji, J.R., Kong, C.Y., Hoekstra, H.E., 2009. Convergent evolution of novel protein function in shrew and lizard venom. Curr. Biol. 19, 1925–1931.

Apitz-Castro, R., Béguin, S., Tablante, A., Bartoli, F., Holt, J.C., Hemker, H.C., 1995. Purification and partial characterization of draculin, the anticoagulant factor present in the saliva of vampire bats (*Desmodus rotundus*). Thromb. Haemost. 73, 94–100.

Baruah, D.B., Dash, R.N., Chaudhari, M.R., Kadam, S.S., 2006. Plasminogen activators: a comparison. Vascul. Pharmacol. 44, 1–9.

Basanova, A.V., Baskova, I.P., Zavalova, L.L., 2002. Vascular-platelet and plasma hemostasis regulators from bloodsucking animals. Biochemistry (Mosc) 67, 143–150.

Bode, W., Renatus, M., 1997. Tissue-type plasminogen activator: variants and crystal/solution structures demarcate structural determinants of function. Curr. Opin. Struct. Biol. 7, 865–872.

Breidenstein, C.P., 1982. Digestion and assimilation of bovine blood by a vampire bat (*Desmodus rotundus*). J. Mammal. 63, 482–484.

Brodie III, E.D., 2010. Convergent evolution: pick your poison carefully. Curr. Biol. 20, R152–R154.

Brown, N., Landon, J., 2010. Antivenom: the most cost-effective treatment in the world? Toxicon 55, 1405–1407.

Bücherl, W., 1968. Introduction. In: Bücherl, W., Buckley, E.E., Deulofeu, V. (Eds.), Venomous Animals and Their Venoms. Academic Press, New York, pp. ix–xii.

Calaby, J.H., 1968. The platypus (*Ornithorhynchus anatinus*) and its venomous characteristics. In: Bücherl, W., Buckley, E.E., Deulofeu, V. (Eds.), Venomous Animals and Their Venoms. Academic Press, New York, pp. 15–30.

Calvete, J.J., 2009. Venomics: digging into the evolution of venomous systems and learning to twist nature to fight pathology. J. Proteomics 72, 121–126.

Cartwright, T., 1974. The plasminogen activator of vampire bat saliva. Blood 43, 317–326.

Chadwick, J.B., 1969. New England's venomous mammal. N. Engl. J. Med. 281, 274.

Cuenca-Bescós, G., Rofes, J., 2007. First evidence of poisonous shrews with an envenomation apparatus. Naturwissenschaften 94, 113–116.

Datta, G., Tu, A.T., 1997. Structure and other chemical characterizations of gila toxin, a lethal toxin from lizard venom. J. Pept. Res. 50, 443–450.

de Plater, G., Martin, R.L., Milburn, P.J., 1995. A pharmacological and biochemical investigation of the venom from the platypus (*Ornithorhynchus anatinus*). Toxicon 33, 157–169.

de Plater, G.M., Martin, R.L., Milburn, P.J., 1998a. A C-type natriuretic peptide from the venom of the platypus (*Ornithorhynchus anatinus*): structure and pharmacology. Comp. Biochem. Physiol. C Pharmacol. Toxicol. Endocrinol. 120, 99–110.

de Plater, G.M., Martin, R.L., Milburn, P.J., 1998b. The natriuretic peptide (ovCNP-39) from platypus (*Ornithorhynchus anatinus*) venom relaxes the isolated rat uterus and promotes oedema and mast cell histamine release. Toxicon 36, 847–857.

de Plater, G.M., Milburn, P.J., Martin, R.L., 2001. Venom from the platypus, *Ornithorhynchus anatinus*, induces a calcium-dependent current in cultured dorsal root ganglion cells. J. Neurophysiol. 85, 1340–1345.

Delpietro, H.A., Russo, R.G., 2009. Acquired resistance to saliva anticoagulants by prey previously fed upon by vampire bats (*Desmodus rotundus*): evidence for immune response. J. Mammal. 90, 1132–1138.

DiSanto, P.E., 1960. Anatomy and histochemistry of the salivary glands of the vampire bat, *Desmodus rotundus murinus*. J. Morphol. 106, 301–335.

Ditmars, R.L., Greenhall, A.M., 1935. The vampire bats: presentation of undescribed habits and review of its history. Zoologica 19, 53–76.

Dufton, M.J., 1992. Venomous mammals. Pharmacol. Ther. 53, 199–215.

Dunn, L.H., 1932. Experiments in the transmission of *Trypanosoma hippicum* Darling with the vampire bat, *Desmodus rotundus* murinus Wagner, as a vector in Panama. J. Preventative Med. 6, 415–424.

Ehrlich, A., Musicant, A., 1977. Social and individual behaviors in captive slow lorises. Behaviour 60, 195–220.

Fenner, P.J., Williamson, J.A., Myers, D., 1992. Platypus envenomation: a painful learning experience. Med. J. Aust. 157, 829–832.

Fernandez, A.Z., Tablante, A., Bartoli, F., Beguin, S., Hemker, H.C., Apitz-Castro, R., 1998. Expression of biological activity of draculin, the anticoagulant factor from vampire bat saliva, is strictly dependent on the appropriate glycosylation of the native molecule. Biochim. Biophys. Acta 1425, 291–299.

Fernandez, A.Z., Tablante, A., Beguín, S., Hemker, H.C., Apitz-Castro, R., 1999. Draculin, the anticoagulant factor in vampire bat saliva, is a tight-binding, noncompetitive inhibitor of activated factor X. Biochim. Biophys. Acta 1434, 135–142.

Fitch-Snyder, H., Schulze, H. (Eds.), 2001. Management of Lorises in Captivity: a Husbandry Manual for Asian Lorisines (*Nycticebus* & Loris ssp.). Center for Reproduction of Endangered Species (CRES), Zoological Society of San Diego, San Diego.

Folinsbee, K.E., Müller, J., Reisz, R.R., 2007. Canine grooves: morphology, function, and relevance to venom. J. Vertebr. Paleontol. 27, 547–551.

Fox, R.C., Scott, C.S., 2005. First evidence of a venom delivery apparatus in extinct mammals. Nature 435, 1091–1093.

Fry, B.G., Roelants, K., Champagne, D.E., Scheib, H., Tyndall, J.D., King, G.F., Nevalainen, T.J., Norman, J.A., Lewis, R.J., Norton, R.S., Renjifo, C., de la Vega, R.C., 2009. The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. Annu. Rev. Genomics Hum. Genet. 10, 483–511.

Furió, M., Agustí, J., Mouskhelishvili, A., Sanisidro, Ó., Santos-Cubedo, A., 2010. The paleobiology of the extinct venomous shrew *Beremendia* (Soricidae, Insectivora, Mammalia) in relation to the geology and paleoenvironment of Dmanisi (Early Pleistocene, Georgia). J. Vertebr. Paleontol. 30, 928–942.

Gardell, S.J., Duong, L.T., Diehl, R.E., York, J.D., Hare, T.R., Register, R.B., Jacobs, J.W., Dixon, R.A., Friedman, P.A., 1989. Isolation, characterization, and cDNA cloning of a vampire bat salivary plasminogen activator. J. Biol. Chem. 264, 17947–17952.

Gohlke, M., Baude, G., Nuck, R., Grunow, D., Kannicht, C., Bringmann, P., Donner, P., Reutter, W., 1996. O-linked L-fucose is present in *Desmodus rotundus* salivary plasminogen activator. J. Biol. Chem. 271, 7381–7386.

Gohlke, M., Nuck, R., Kannicht, C., Grunow, D., Baude, G., Donner, P., Reutter, W., 1997. Analysis of site-specific N-glycosylation of recombinant *Desmodus rotundus* salivary plasminogen activator rDSPAα1 expressed in Chinese hamster ovary cells. Glycobiology 7, 67–77.

Grant, T.R., Temple-Smith, P.D., 1998. Field biology of the platypus (*Ornithorhynchus anatinus*): historical and current perspectives. Phil. Trans. R. Soc. Lond. B 353, 1081–1091.

Griffith, I.J., Craig, S., Pollock, J., Yu, X.B., Morgenstern, J.P., Rogers, B.L., 1992. Expression and genomic structure of the genes encoding FdI, the major allergen from the domestic cat. Gene 113, 263–268.

Gundlach, J., 1877. Contribucion a la mamalogia cubana. Impr. de G. Montiel y Comp., Habana, 39–44.

Hagey, L.R., Fry, B.G., Fitch-Snyder, H., 2007. Talking defensively: a dual use for the brachial gland exudate of slow and pygmy lorises. In: Gurski, S., Nekaris, K.A.I. (Eds.), 2007. Primate Anti-predatory Strategies, vol. 2. Springer, New York, pp. 253–272.

Hawkey, C., 1966. Plasminogen activator in saliva of the vampire bat *Desmodus rotundus*. Nature 211, 434–435.

Hawkey, C., 1967. Inhibitor of platelet aggregation present in saliva of the vampire bat *Desmodus rotundus*. Br. J. Haematol. 13, 1014–1020.

Hemker, H.C., Béguin, S., 1995. Thrombin generation in plasma: its assessment via the endogenous thrombin potential. Thromb. Haemost. 74, 134–138.

Hiramatsu, M., Hatakeyama, K., Minami, N., 1980. Male mouse submaxillary gland secretes highly toxic proteins. Experientia 36, 940–942.

Hodgson, W.C., 1997. Pharmacological action of Australian animal venoms. Clin. Exp. Pharmacol. Physiol. 24, 10–17.

Hurum, J.H., Luo, Z.-X., Kielan-Jaworowska, Z., 2006. Were mammals originally venomous? Acta Paleontol. Pol. 51, 1–11.

Jamison, J., 1818. Extracts from the minute-book of the Society Mar. 18, 1817. Trans. Linn. Soc. XII, 584–585.

Kaiser, L., Grönlund, H., Sandalova, T., Ljunggren, H.G., van Hage-Hamsten, M., Achour, A., Schneider, G., 2003. J. Biol. Chem. 278, 37730–37735.

Kalimullah, E.A., Schmidt, S.M., Schmidt, M.J., Lu, J.J., 2008. Beware the pygmy slow loris? Clin. Toxicol. 46, 602.

Kellaway, C.H., LeMessurier, D.H., 1935. The venom of the platypus (*Ornithorhynchus anatinus*). Aust. Exp. Biol. Med. Sci. 13, 205–221.

Kielan-Jaworowska, Z., Hurum, J.H., 2006. Limb posture in early mammals: sprawling or parasagittal. Acta Paleontol. Pol. 51, 393–406.

Kita, M., Nakamura, Y., Okumura, Y., Ohdachi, S.D., Oba, Y., Yoshikuni, M., Kido, H., Uemura, D., 2004. Blarina toxin, a mammalian lethal venom from the short-tailed shrew *Blarina brevicauda*: isolation and characterization. Proc. Natl. Acad. Sci. U. S. A. 101, 7542–7547.

Kita, M., Okumura, Y., Ohdachi, S.D., Oba, Y., Yoshikuni, M., Nakamura, Y., Kido, H., Uemura, D., 2005. Purification and characterisation of blarinasin, a new tissue kallikrein-like protease from the short-tailed shrew *Blarina brevicauda*: comparative studies with blarina toxin. Biol. Chem. 386, 177–182.

Kita, M., Black, D.S.C., Ohno, O., Yamada, K., Kigoshi, H., Uemura, D., 2009. Duck-billed platypus venom peptides induce Ca^{2+} influx in neuroblastoma cells. J. Am. Chem. Soc. 131, 18038–18039.

Klinger, L.S., 2008. The context of Dracula. In: Stoker, B. (Ed.), The New Annotated Dracula. W. W. Norton & Co, New York, pp. XIX–L.
Koh, J.M.S., Bansal, P.S., Torres, A.L., Kuchel, P.W., 2009. Platypus venom: source of novel compounds. Aust. J. Zool. 57, 203–210.
Koh, J.M., Chow, S.J., Crossett, B., Kuchel, P.W., 2010. Mammalian peptide isomerase: platypus-type activity is present in mouse heart. Chem. Biodivers. 7, 1603–1611.
Kourie, J.I., 1999a. A component of platypus (*Ornithorhynchus anatinus*) venom forms slow-kinetics cation channels. J. Membr. Biol. 172, 37–45.
Kourie, J.I., 1999b. Characterization of a C-type natriuretic peptide (CNP-39)-formed cation-selective channel from platypus (*Ornithorhynchus anatinus*) venom. J. Physiol. 518, 359–369.
Krane, S., Itagaki, Y., Nakanishi, K., Weldon, P.J., 2003. "Venom" of the slow loris: sequence similarity of prosimian skin gland protein and Fel d 1 cat allergen. Naturwissenschaften 90, 60–62.
Krätzschmar, J., Haendler, B., Langer, G., Boidol, W., Bringmann, P., Alagon, A., Donner, P., Schleuning, W.D., 1991. The plasminogen activator family from the salivary gland of the vampire bat *Desmodus rotundus*: cloning and expression. Gene 105, 229–237.
Krause, W.J., 2009. Morphological and histochemical observations on the crural gland-spur apparatus of the echidna (*Tachyglossus aculeatus*) together with comparative observations on the femoral gland-spur apparatus of the duckbilled platypus (*Ornithorhyncus anatinus*). Cells Tissues Organs 191, 336–354.
Leitermann, K., Ohman, J.L., 1984. Cat allergen 1: biochemical, antigenic, and allergenic properties. J. Allergy Clin. Immunol. 74, 147–153.
Lawrence, B., 1945. Brief comparison of short-tailed shrew and reptile poisons. J. Mammal. 26, 393–396.
Lopez-Jurado, L.F., Mateo, J.A., 1996. Evidence of venom in the Canarian shrew (*Crocidura canariensis*): immobilizing effects on the Atlantic lizard (*Gallotia atlantica*). J. Zool. 239, 394–395.
Machado-Santos, C., Nascimento, A.A., Peracchi, A.L., Mikalauskas, J.S., Rocha, P.A., Sales, A., 2009. Distributions of the endocrine cells in the gastrointestinal tract of nectarivorous and sanguivorous bats: a comparative immunocytochemical study. Tissue Cell 41, 222–229.
Málaga-Alba, A., 1954. Vampire bat as a carrier of rabies. Am. J. Public Health Nations Health 44, 909–918.
Martin, I.G., 1981. Venom of the short-tailed shrew (*Blarina brevicauda*) as an insect immobilizing agent. J. Mammal. 62, 189–192.
Martin, C.J., Tidswell, F., 1895. Observations on the femoral gland of *Ornithorhynchus* and its secretion; together with an experimental enquiry concerning its supposed toxic action. Proc. Linn. Soc. N. S. W. 9, 471–500.
Maynard, C.J., 1889. Singular effects produced by the bite of a short-tailed shrew, *Blarina brevicauda*. Cont. Sci. 1, 57–59.
Mebs, D., 1999. Studies on biological and enzymatic activities of salivary glands from the European hedgehog (*Erinaceus europaeus*). Toxicon 37, 1635–1638.
Mebs, D., 2002. Venomous and Poisonous Animals. Medpharm, Stuttgart, 1–31, 322–327 pp.
Ménez, R., Michel, S., Muller, B.H., Bossus, M., Ducancel, F., Jolivet-Reynaud, C., Stura, E.A., 2008. Crystal structure of a ternary complex between human prostate-specific antigen, its substrate acyl intermediate and an activating antibody. J. Mol. Biol. 376, 1021–1033.
Merrit, J.F., 1986. Winter survival adaptations of the short-tailed shrew (*Blarina brevicauda*) in an Appalachian montane forest. J. Mammal. 67, 450–464.
Mitchell, G.C., Tigner, J.R., 1970. The route of ingested blood in the vampire bat (*Desmodus rotundus*). J. Mammal. 51, 814–817.
Morgenstern, J.P., Griffith, I.J., Brauer, A.W., Rogers, B.L., Bond, J.F., Chapman, M.D., Kuo, M.C., 1991. Amino acid sequence of Fel dI, the major allergen of the domestic cat: protein sequence analysis and cDNA cloning. Proc. Natl. Acad. Sci. U. S. A. 88, 9690–9694.
Orr, C.M., Delezene, L.K., Scott, J.M., Tocheri, M.W., Schwartz, G.T., 2007. The comparative method and the inference of venom-delivery systems in fossil mammals. J. Vertebr. Paleontol. 27, 541–546.
Paciaroni, M., Medeiros, E., Bogousslavsky, J., 2009. Desmoteplase. Expert Opin. Biol. Ther. 9, 773–778.
Pearson, O.P., 1942. On the cause and nature of a poisonous action produced by the bite of a shrew (*Blarina brevicauda*). J. Mammal. 23, 159–166.
Petrie, A.F., Pearn, J.H., 1987. Platypus envenomation: two case reports. Vet. Hum. Toxicol. 29, 493.
Pournelle, G.H., 1968. Classification, biology, and description of the venom apparatus of insectivores of the genera *Solenodon, Neomys*, and Blarina. In: Bücherl, W., Buckley, E.E., Deulofeu, V. (Eds.), Venomous Animals and Their Venoms. Academic Press, New York, pp. 31–42.
Pucek, M., 1968. Chemistry and pharmacology of Insectivore venoms. In: Bücherl, W., Buckley, E.E., Deulofeu, V. (Eds.), Venomous Animals and Their Venoms. Academic Press, New York, pp. 43–50.
Rabb, G.B., 1959. Toxic salivary glands in the primitive insectivore Solenodon. Nat. Hist. Misc. 190, 1–3.
Renatus, M., Stubbs, M.T., Huber, R., Bringmann, P., Donner, P., Schleuning, W.D., Bode, W., 1997. Catalytic domain structure of vampire bat plasminogen activator: a molecular paradigm for proteolysis without activation cleavage. Biochemistry 36, 13483–13493.
Rofes, J., Cuenca-Bescós, G., 2009. First record of *Beremendia fissidens* (Mammalia, Soricidae) in the Pleistocene of the Iberian Peninsula, with a review of the biostratigraphy, biogeography and palaeoecology of the species. C. R. Palevol. 8, 21–37.
Schleuning, W.D., 2001. Vampire bat plasminogen activator DSPA-alpha-1 (desmoteplase): a thrombolytic drug optimized by natural selection. Haemostasis 31, 118–122.
Schondube, J.E., Herrera-M, L.G., Martínez del Rio, C., 2001. Diet and the evolution of digestion and renal function in phyllostomid bats. Zoology (Jena) 104, 59–73.
Soy, J., Mancina, C.A., 2008. Solenodon cubanus. In: IUCN 2010. IUCN Red List of Threatened Species. Version 2010.4. <www.iucnredlist.org>.
Spicer, W.W., 1876. On the effects of wounds on the human subject inflicted by the spurs of the platypus – (*Ornythorhynchus anatinus*). Pap. Proc. R. Soc., Tas., 162–167.
Starr, C., Nekaris, K.A.I., Streicher, U., Leung, L., 2010. Traditional use of slow lorises *Nycticebus bengalensis* and *N. pygmaeus* in Cambodia: an impediment to their conservation. Endang. Species Res. 12, 17–23.
Stewart, R.J., Fredenburgh, J.C., Weitz, J.I., 1998. Characterization of the interactions of plasminogen and tissue and vampire bat plasminogen activators with fibrinogen, fibrin, and the complex of D-dimer non-covalently linked to fragment E. J. Biol. Chem. 273, 18292–18299.
Tandler, B., Pinkstaff, C.A., Nagato, T., Phillips, C.J., 1996. Giant secretory granules in the ducts of the parotid and submandibular glands of the slow loris. Tissue Cell 28, 321–329.
Tellgren-Roth, A., Dittmar, K., Massey, S.E., Kemi, C., Tellgren-Roth, C., Savolainen, P., Lyons, L.A., Liberles, D.A., 2009. Keeping the blood flowing-plasminogen activator genes and feeding behavior in vampire bats. Naturwissenschaften 96, 39–47.
Thompson, R.D., Mitchell, G.C., Burns, R.J., 1972. Vampire bat control by systemic treatment of livestock with an anticoagulant. Science 177, 806–808.
Tomasi, T.E., 1978. Function of venom in the short-tailed shrew, *Blarina brevicauda*. J. Mammal. 59, 852–854.
Tonkin, M.A., Negrine, J., 1994. Wild platypus attack in the antipodes: a case report. J. Hand Surg. 19B, 162–164.
Topsell, R.E., 1607. Historie of Foure-Footed Beasts. Jaggard, London.
Torres, A.M., Wang, X., Fletcher, J.I., Alewood, D., Alewood, P.F., Smith, R., Simpson, R.J., Nicholson, G.M., Sutherland, S.K., Gallagher, C.H., King, G.F., Kuchel, P.W., 1999. Solution structure of a defensin-like peptide from platypus venom. Biochem. J. 341, 785–794.
Torres, A.M., de Plater, G.M., Doverskog, M., Birinyi-Strachan, L.C., Nicholson, G.M., Gallagher, C.H., Kuchel, P.W., 2000. Defensin-like peptide-2 from platypus venom: member of a class of peptides with a distinct structural fold. Biochem. J. 348, 649–656.
Torres, A.M., Menz, I., Alewood, P.F., Bansal, P., Lahnstein, J., Gallagher, C.H., Kuchel, P.W., 2002. D-Amino acid residue in the C-type natriuretic peptide from the venom of the mammal, Ornithorhynchus anatinus, the Australian platypus. FEBS Lett. 524, 172–176.
Torres, A.M., Kuchel, P.W., 2004. The β-defensin-fold family of polypeptides. Toxicon 44, 581–588.
Torres, A.M., Tsampazi, C., Geraghty, D.P., Bansal, P.S., Alewood, P.F., Kuchel, P.W., 2005. D-amino acid residue in a defensin-like peptide from platypus venom: effect on structure and chromatographic properties. Biochem. J. 391, 215–220.
Torres, A.M., Tsampazi, M., Tsampazi, C., Kennett, E.C., Belov, K., Geraghty, D.P., Bansal, P.S., Alewood, P.F., Kuchel, P.W., 2006. Mammalian L-to-D-amino-acid-residue isomerase from platypus venom. FEBS Lett. 580, 1587–1591.
Turvey, S.T., 2010. Evolution of non-homologous venom delivery systems in West Indian insectivores? J. Vertebr. Paleontol. 30, 1294–1299.
Turvey, S., Incháustegui, S., 2008. Solenodon paradoxus. In: IUCN 2010. IUCN Red List of Threatened Species. Version 2010.4. <www.iucnredlist.org>.
Utaisincharoen, P., Mackessy, S.P., Miller, R.A., Tu, A.T., 1993. Complete primary structure and biochemical properties of gilatoxin, a serine protease with kallikrein-like and angiotensin-degrading activities. J. Biol. Chem. 268, 21975–21983.
Utami, S.S., van Hooff, J.A.R.A.M., 1997. Meat-eating by adult female Sumatran orangutans (*Pongo pygmaeus abelii*). Am. J. Primatol. 43, 159–165.
Valenta, R., Steinberger, P., Duchêne, M., Kraft, D., 1996. Immunological and structural similarities among allergens: prerequisite for a specific and component-based therapy of allergy. Immunol. Cell Biol. 74, 187–194.

Warren, W.C., et al., 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453, 175–184.
Whittington, C.M., Belov, K., 2007. Platypus venom: a review. Aust. Mammal. 29, 57–62.
Whittington, C.M., Belov, K., 2009. Platypus venom genes expressed in non-venom tissues. Aust. J. Zool. 57, 199–202.
Whittington, C.M., Papenfuss, A.T., Bansal, P., Torres, A.M., Wong, E.S., Deakin, J.E., Graves, T., Alsop, A., Schatzkamer, K., Kremitzki, C., Ponting, C.P., Temple-Smith, P., Warren, W.C., Kuchel, P.W., Belov, K., 2008. Defensins and the convergent evolution of platypus and reptile venom genes. Genome Res. 18, 986–994.
Whittington, C.M., Koh, J.M., Warren, W.C., Papenfuss, A.T., Torres, A.M., Kuchel, P.W., Belov, K., 2009. Understanding and utilising mammalian venom via a platypus venom transcriptome. J. Proteomics 72, 155–164.
Whittington, C.M., Papenfuss, A.T., Locke, D.P., Mardis, E.R., Wilson, R.K., Abubucker, S., Mitreva, M., Wong, E.S., Hsu, A.L., Kuchel, P.W., Belov, K., Warren, W.C., 2010. Novel venom gene discovery in the platypus. Genome Biol. 11, R95.
Wiens, F., Zitzmann, A., 1999. Predation on a wild slow loris (*Nycticebus coucang*) by a reticulated python (*Python reticulatus*). Folia Primatol. 70, 362–364.
Wilde, H., 1972. Anaphylactic shock following bite by a 'Slow loris', *Nycticebus coucang*. Am. J. Trop. Med. Hyg. 21, 592–594.
Witt, W., Maass, B., Baldus, B., Hildebrand, M., Donner, P., Schleuning, W.D., 1994. Coronary thrombolysis with *Desmodus* salivary plasminogen activator in dogs. Fast and persistent recanalization by intravenous bolus administration. Circulation 90, 421–426.
Zavalova, L.L., Basanova, A.V., Baskova, I.P., 2002. Fibrinogen-fibrin system regulators from bloodsuckers. Biochemistry (Mosc) 67, 135–142.

Apêndice C

Estrutura de receptores-alvo de fármacos
através de modelagem comparativa

Sachett LG, Ligabue-Braun R, Verli H

In Trossini G, Castilho MS (Eds.)*, Práticas de Modelagem Molecular*

O capítulo a seguir, apresentado em sua formatação preliminar, é parte do livro-texto “Práticas de Modelagem Molecular” que encontra-se em fase final de preparação. Algumas seções do capítulo (como práticas e exercícios avançados) foram omitidas da presente tese por solicitação dos editores.

ESTRUTURA DE RECEPTORES-ALVO DE FÁRMACOS ATRAVÉS DE MODELAGEM COMPARATIVA

Liana G. Sachett, Rodrigo L. Braun, Hugo Verli.

Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia

Uma das principais dificuldades associadas ao planejamento de novos fármacos está na ausência de informações sobre a estrutura 3D de seus receptores-alvo. Uma das técnicas computacionais disponíveis para contornar esta limitação é denominada modelagem comparativa. Esta técnica consiste na construção de um modelo 3D para uma dada proteína usando, como referência (ou molde), a estrutura de uma proteína semelhante. O modelo assim obtido poderá ser empregado em outros estudos, como ancoramento (docking). Os passos necessários para a realização da modelagem comparativa serão descrito em detalhes neste capítulo, incluindo-se exemplos, limitações e dificuldades mais comuns, bem como interpretação de resultados.

CAPITULO 10

INTRODUÇÃO

Diversas estratégias de desenvolvimento de novos fármacos estão baseadas no conhecimento da estrutura 3D do receptor-alvo, tais como o planejamento baseado na estrutura do receptor (structure-based drug design), estudos de ancoramento (docking) e simulações por dinâmica molecular. Adicionalmente, métodos como QSAR e SAR também podem empregar informações sobre a estrutura 3D da proteína-alvo. Infelizmente, em torno de 3% das proteínas humanas tiveram suas estruturas 3D determinadas experimentalmente, o que demanda o emprego de técnicas adicionais.

Uma das principais técnicas computacionais empregada na obtenção de estruturas 3D de proteínas, tais como receptores-alvo de fármacos, é denominada modelagem comparativa de proteínas ou modelagem por homologia. Esta técnica é baseada no pressuposto de que proteínas com sequências de aminoácidos semelhantes possuem estruturas 3D semelhantes. Esta semelhança nas sequências representa, usualmente, uma origem evolutiva comum, o que é denominado homologia. Assim, a proteína de interesse do Químico Medicinal, para a qual não há estrutura 3D disponível (denominada de sequência-alvo), deve ser usualmente homóloga a uma proteína que possua estrutura 3D já estabelecida (denominada de molde) para que a técnica de modelagem por homologia possa ser empregada.

Contudo, o conceito de homologia implica em ser ou não ser homólogo, nos impedindo de quantificar o "grau de homologia". Isto é importante, porque as sequências de aminoácidos de duas proteínas podem ser muito ou pouco parecidas. E quanto mais "parecidas" forem estas sequências (da sequência-alvo e do molde), maior será a confiabilidade do modelo obtido. Uma das formas de comparar duas sequências de aminoácidos é através do grau de identidade entre as mesmas, de 0 a 100%. Em outras palavras, a identidade mede o número de resíduos de aminoácidos que são idênticos e ocupam as mesmas posições nas sequências das proteínas em estudo. A identidade mínima entre a sequência-alvo e o molde deverá ser de 30% para obtenção de modelos confiáveis por modelagem comparativa. Quanto maior o grau de identidade, mais confiável será o modelo.

Usualmente, a modelagem comparativa compreende os seguintes passos: busca por moldes, seleção de um ou mais moldes, construção do modelo e avaliação do modelo (Figura 1). Adicionalmente, embora diversas ferramentas computacionais estejam disponíveis para a realização de estudos de modelagem comparativa, empregaremos aqui o servidor denominado SWISS-MODEL (http://swissmodel.expasy.org/). Considere que algumas pequenas diferenças nos procedimentos apresentados podem ser encontradas em outros programas, embora o roteiro geral seja bastante semelhante.

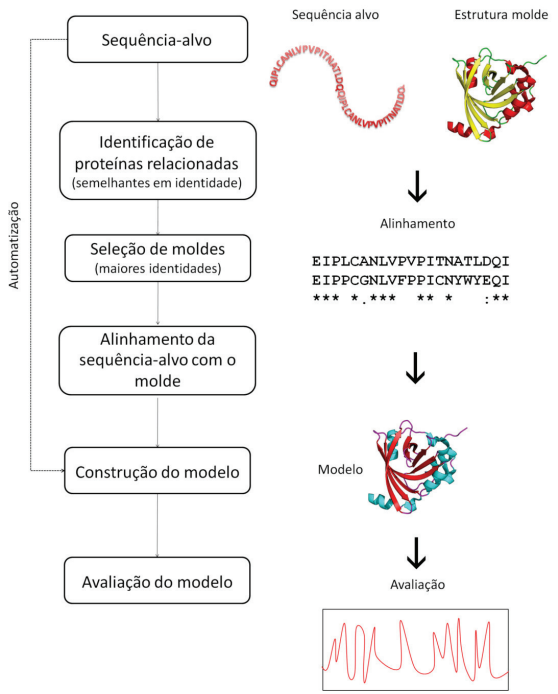


Figura 1. Passos da modelagem comparativa (baseado em Sánchez & Šali, 2000).

A identificação de estruturas relacionadas (Figura 1) é realizada comparando (ou alinhando) a sequência-alvo à sequências de aminoácidos de proteínas que já tenham estruturas 3D determinadas. Esta etapa pode ser realizada, por exemplo, utilizando-se a ferramenta BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), selecionando o menu "protein blast". Gera-se, assim, uma lista de diferentes proteínas, relacionadas à sequência-alvo, candidatos a moldes para a modelagem comparativa. Adicionalmente, são apresentadas as identidades entre a sequência-alvo e cada possível molde, além de seus respectivos alinhamentos.

O alinhamento realizado entre a sequência-alvo e seu molde é uma das partes mais importantes da modelagem comparativa de proteínas. A definição de quais resíduos de aminoácidos correspondem entre as duas macromoléculas torna-se determinante da forma 3D a ser adotada pela sequência-alvo. Por outro lado, a seleção de um molde, a

partir dos candidatos obtidos pelo alinhamento, tem como regra mais simples a seleção da proteína com a maior identidade em relação à sequência-alvo. Adicionalmente, deve-se considerar a qualidade da estrutura 3D experimental, por exemplo pela resolução cristalográfica, medida em angstrons (Å). Quanto menor o valor da resolução, melhor (idealmente em torno de 2,5 Å ou menor).

A etapa final da modelagem comparativa consiste na validação do modelo, onde diversas características, principalmente estereoquímicas e funcionais, são avaliadas. Neste processo, a proteína modelada deve tanto apresentar uma estrutura 3D satisfatória, seguindo alguns critérios previamente estabelecidos pela comunidade científica, quanto ser capaz de explicar aspectos funcionais já conhecidos, como aqueles relacionados à catálise (como a composição e geometria da tríade catalítica), à ligação de cofatores estruturais (como metais) ou de moduladores (neurotransmissores ou fármacos, por exemplo). A qualidade do modelo dependerá, principalmente, da qualidade do molde escolhido e do alinhamento. A quantidade de alças na proteína, alinhamentos contendo erros ou gaps muito longos são alguns fatores que também podem influenciar na qualidade do modelo obtido.

Diversas estratégias estão disponíveis para realizar a validação estereoquímica dos modelos obtidos. Uma das principais envolve o Mapa de Ramachandran. Resumidamente, este mapa apresenta geometrias (ou zonas) proibidas e permitidas para o esqueleto peptídico da proteína. Para a obtenção deste dado pode ser utilizado, por exemplo, o programa PROCHEK através do servidor PDBSUM (www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html). Outras estratégias comuns para validação estereoquímica podem ser encontradas na plataforma SWISS-MODEL e incluem o QMEAN, que pontua a confiabilidade do modelo a partir de comparações à estruturas cristalográficas de alta resolução, e a Análise de Erros Locais, que consiste em estimar a qualidade estereoquímica resíduo a resíduo e entre resíduos vizinhos. Este dado é apresentado na forma de um gráfico contendo pontuações para cada resíduo que refletem a probabilidade de erro em cada um deles. Para facilitar a interpretação dos resultados, uma representação do modelo proteico colorido por esta pontuação de erros é apresentada simultaneamente.

A validação funcional dos modelos, ao contrário da estereoquímica, é baseada mais na análise da literatura científica do que em resultados de programas ou servidores na internet. Em geral, busca-se o maior volume possível de informações bioquímicas, estruturais, funcionais e terapêuticas disponíveis, e avalia-se a capacidade do modelo em responder ou explicar estas questões.

A modelagem comparativa pode ser realizada tanto de uma forma automatizada, frequentemente disponível em servidores na internet, quanto de uma forma manual, usualmente baseada em programas que funcionam localmente, no computador do usuário. Nos casos mais difíceis, o emprego de estratégias manuais pode ser necessária, maximizando a qualidade dos modelos. Para estes casos mais delicados, ou quando o usuário deseja

um controle total de todas as etapas realizadas, o programa MODELLER é um dos mais utilizados (www.salilab.org/modeller). Na página do programa encontram-se instruções para instalação e tutoriais para o seu uso.

Discussão e Conclusão

Ao realizar os passos essenciais na construção de um modelo tridimensional de proteínas, a atividade proposta no roteiro de aula prática evidencia tanto a praticidade quando a aplicabilidade do método. Como observado ao longo do capítulo, a maioria das etapas não apresenta grandes dificuldades em sua execução, podendo ser seguida à maneira de um protocolo. Em casos de modelagem manual, a escolha da estrutura molde, o alinhamento das sequências alvo e molde e a seleção de parâmetros para construção do modelo são realizados pelo usuário. No procedimento automatizado, uma vez de posse da sequência polipeptídica, o próprio servidor de modelagem seleciona o melhor molde, o alinha à sequência-alvo e executa as rotinas de construção tridimensional. Ao seguir uma abordagem automatizada para modelagem, o ponto crítico passa a ser a avaliação do modelo obtido. A qualidade e a validade do modelo poderão ser definidas apenas por verificação manual dos resultados de sua avaliação, caracterizando o modelo obtido como realístico e confiável. De fato, esta verificação é essencial para que estudos avaliando a interação entre ligante e receptor possam ser realizados de forma realística. De maneira geral, modelos ruins podem levar a conclusões ruins.

Entre as avaliações essenciais está o gráfico de Ramachandran. De interpretação objetiva, este gráfico permite tanto avaliar a qualidade global do modelo, quanto identificar resíduos específicos que possam apresentar falhas estereoquímicas. Adicionalmente, do ponto de vista funcional, é de grande importância avaliar a integridade do sítio catalítico na estrutura modelada, no caso de enzimas, como a DOPAd, enzima exemplificada no roteiro de aula prática. Isso é geralmente feito por comparação com dados da literatura. Proteínas não-enzimáticas (como carreadores de fármacos, canais iônicos ou receptores acoplados à proteína G) também apresentam resíduos importantes para sua atividade, e estes devem ser inspecionados no modelo obtido.

Após avaliação e validação, pode-se verificar a alta qualidade da estrutura da DOPA descarboxilase humana obtida nesta atividade. Tal estrutura pode ser empregada em estudos de planejamento racional de novos fármacos, por exemplo, através de ancoramento (docking) molecular. Ainda, a estrutura obtida por modelagem comparativa pode ser visualizada em softwares como PyMol (Capítulo XX), VMD, UCSF Chimera e Swiss PDB Viewer.

Ao realizar a atividade de modelagem comparativa baseada na DOPAd, são desenvolvidos os passos essenciais na construção de um modelo tridimensional de uma proteína. Este é, contudo, um exemplo geral, existindo casos especiais na construção de estruturas tridimensionais. Alguns deles são discutidos nos exercícios adicionais (proteínas transmembrana e acopladas a membranas). Em outros casos, entretanto, não é possível aplicar a modelagem

comparativa. Tais casos geralmente envolvem proteínas para as quais não existem moldes disponíveis. Uma discussão sobre estes casos está além do objetivo deste capítulo, sendo o leitor encaminhado ao artigo de Jones (2000) como um guia inicial para estes casos. Apesar de poder ser empregada isoladamente, em busca da estrutura de uma proteína específica, a modelagem comparativa constitui uma etapa intermediária em estudos de desenvolvimento de fármacos, auxiliando na compreensão das relações estrutura-atividade de compostos bioativos, utilizando ferramentas gratuitas e dados disponíveis publicamente.

REFERÊNCIAS

CHOTHIA, C.; LESK, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J., 1986, 5, 823-826.

DE BREVERN, A.G. 3D structural models of transmembrane proteins. Methods Mol. Biol., 2010, 654, 387–401.

FISER, A.; DO, R.K.G.; ŠALI, A. Modeling of loops in protein structures. Protein Sci., 2000, 9, 1753-1773.

FORSTER, M.J. Molecular modelling in structural biology. Micron, 2002, 33, 365-384.

JONES, D.T. A practical guide to protein structure prediction. Methods Mol. Biol., 2000, 143, 131-154.

LESK, A.M. Estrutura de proteínas e descoberta de fármacos. In: _____. Introdução à Bioinformática. Porto Alegre: Artmed, 2008.

MARTÍ-RENOM, M.A.; et al. Comparative protein structure modeling of genes and genomes. Annu. Rev.. Biophys. Biomol. Struct., 2000, 29, 291-325.

REECK, G.R. et al. “Homology” in protein and nucleic acids: a terminology muddle and a way out of it. Cell, 1987, 50, 667.

ŠALI, A; BLUNDELL, T.L. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol., 1993, 234, 779-815.

SÁNCHEZ, R.; ŠALI, A. Comparative protein structure modeling. Introduction and practical examples with Modeller. Methods Mol. Biol., 2000, 143, 97-129.

SANTOS FILHO, O.A.; ALENCASTRO, R.B. Modelagem de proteínas por homologia. Quim. Nova, 2003, 26, 253-259.

TRAMONTANO, A. The role of molecular modelling in biomedical research. FEBS Lett., 2006, 580, 2928-2934.

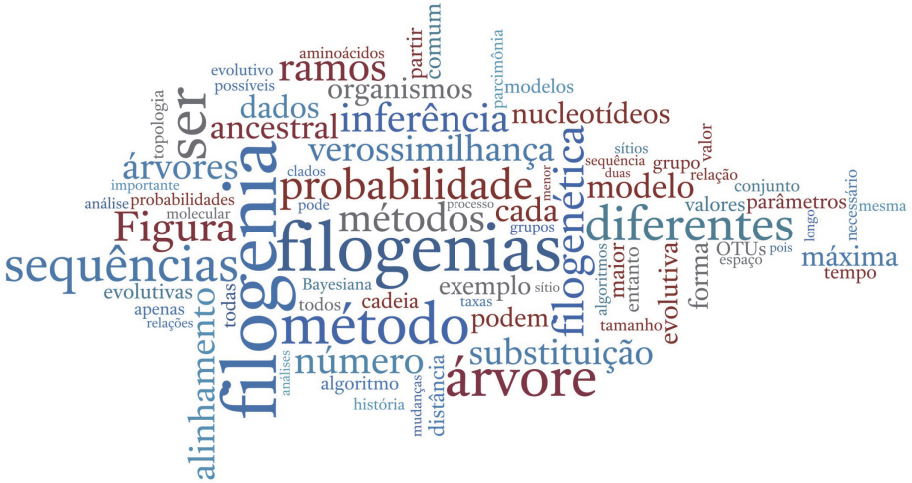
Apêndice D

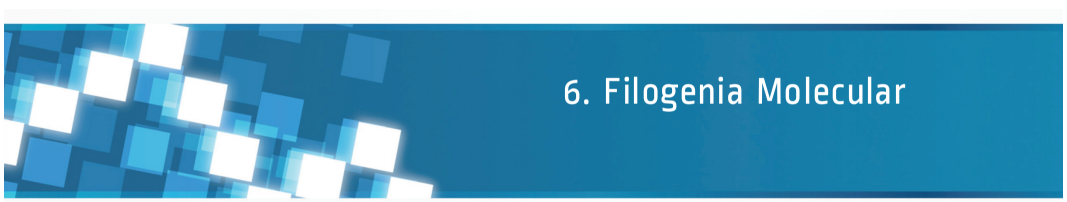
Filogenia Molecular

Ligabue-Braun R, Junqueira DM, Verli H

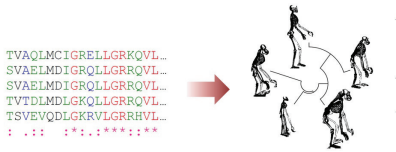
In Verli H (Ed.)*, Bioinformática: da Biologia à Flexibilidade Molecular*

O capítulo a seguir, apresentado em sua formatação preliminar, é parte do livro-texto “Bioinformática: da Biologia à Flexibilidade Molecular” que encontra-se em fase final de preparação. Com distribuição eletrônica gratuita, seu lançamento está previsto para maio de 2014.





6. Filogenia Molecular



Estabelecimento de relações evolutivas a partir de sequências de aminoácidos ou nucleotídeos.

Rodrigo Ligabue Braun
Dennis Maletich Junqueira
Hugo Verli

6.1. Introdução

Desde seus primórdios, a humanidade se mostrou inclinada a organizar e classificar o mundo à sua volta com o objetivo de facilitar o entendimento e a comunicação. Em relação ao mundo natural, diferentes sistemas foram empregados para compor métodos de organização e classificar os organismos, utilizando critérios naturais ou artificiais.

Um dos sistemas de maior influência no período pré-Darwiniano foi a Escala Natural de Platão. Neste sistema, do fogo ao ser humano, diferentes níveis eram organizados à maneira de uma escada. A ideia de ascensão estava associada à perfeição, representada em sua forma plena pelo homem. O sistema classificatório de Lineu, por sua vez, se baseava em características visíveis, arbitrariamente selecionadas para classificar os seres vivos (por exemplo, número de patas ou de pétalas), sendo o ser humano o organismo do topo da cadeia. Sistemas como este são considerados sistemas artificiais, pois estão sujeitos à tendência de seu autor em considerar um caractere em detrimento de outro(s), conforme sua vontade ou necessidade. Entretanto, como o próprio Lineu reconheceu, tais sistemas foram absolutamente necessários para a fase inicial (descritiva) da biologia, servindo de base para o sistema natural de classificação e para as hipóteses de similaridade que surgiriam a seguir.

Ao final do século XVIII e início do século XIX, surgem os sistemas naturais de classificação. Estes buscavam refletir sobre a ordem natural dos seres vivos através de poucas características intrínsecas, geralmente associadas à forma. No entanto, com o objetivo de tornar a classificação mais racional, tomaram lugar debates sobre a real necessidade de haver um sistema hierárquico de organização dos organismos. Opositores da ideia consideravam que a classificação era, muitas vezes, inadequada e desnecessária, e que não deveria ser um fim em si mesma, senão um método para o levantamento de novas perguntas à Biologia.

Em 1818, a introdução do conceito de ***homologia*** por E.G. Saint-Hillaire causa uma

revolução nas ciências biológicas. Para ele e seus colegas, partes homólogas correspondiam às partes de animais diferentes com uma estrutura essencialmente semelhante, mesmo com forma ou função distintas. Por exemplo, as asas de um morcego, as nadadeiras de uma baleia e os braços de um macaco, segundo esta lógica, são considerados órgãos homólogos e podem servir como critério para agrupar morcegos, baleias e macacos em um mesmo grupo. Assim, a homologia serviria como critério principal para uma classificação natural dos organismos.

A primeira árvore filogenética moderna (esboço de Darwin no manuscrito de A Origem das Espécies)

A partir da famosa publicação de Darwin, "A Origem das Espécies", em 1859, a classificação dos organismos passou a ser não apenas natural, mas também a apresentar uma condição essencial de ancestralidade comum. Segundo este pensamento, os organismos são derivados uns dos outros, desde o surgimento da vida na terra. Darwin representou este padrão através de um esquema de ramificação, onde os galhos representam o tempo entre o organismo ancestral e o novo organismo, e os nós representam os próprios organismos. Mais tarde, esta viria a ser a primeira árvore filogenética utilizada para representar processos evolutivos.

Com influência direta da teoria evolutiva de Darwin (e colaborações de Wallace e Lamarck), desenvolve-se a ***Taxonomia Evolutiva***. Este sistema de classificação incorporou o vetor tempo (caráter temporal normalmente inferido por meio de fósseis) e, além disto, adicionou uma quantificação da divergência estrutural entre os grupos (a chamada ***distância patrística***). Já em meados do século XX, inicia-se a Fenética (taxonomia numérica ou neodansoniana). Esta escola buscava incluir na classificação dos organismos o máximo possível de características, atribuindo-lhes o mesmo peso na tentativa de eliminar qualquer subjetividade ou arbitrariedade. Seu impacto, entretanto, foi limitado devido às dificuldades em traduzir os índices (valores) obtidos em informações relevantes do ponto de vista biológico (como a separação de espécies, por exemplo). Na mesma época, surge a ***Cladística*** (ou sistemática filogenética), liderada pelo entomólogo alemão Willi Hennig. Na proposta de Hennig (1950), organismos que compartilhassem características derivadas (apomórficas) poderiam ser considerados descendentes do organismo ancestral, na qual a característica em seu estado primitivo (ou plesiomórfico) passou para o estado derivado.

Desde a origem dos sistemas de classificação até a Cladística, os métodos baseavam-se essencialmente no fenótipo dos organismos, ou seja, em suas características físicas claramente discerníveis. Entretanto, com o advento dos métodos de sequenciamento, tanto protéico quanto genômico, cada vez mais os dados moleculares foram se tornando importantes nas análises evolutivas de ancestralidade. Neste sentido, a ciência passa de um ponto de vista macroscópico a um ponto de vista molecular de análise.

O método de sequenciamento de aminoácidos, iniciado por Sanger em 1954, abriu caminho para que proteínas de uma mesma classe, em diferentes organismos, pudessem ser comparadas quanto às suas origens evolutivas. Da mesma forma, ao decodificar a primeira longa sequência de DNA, em 1977, Sanger deu início à explosão do sequenciamento de ácidos nucleicos, permitindo a comparação de genes em larga escala. É importante destacar que as sequências moleculares podem tanto ser comparadas entre si, buscando conhecer a história evolutiva de um gene ou proteína (por exemplo, relações entre hemoglobinas de diferentes mamíferos), quanto podem ser

associadas a outros dados na reconstrução da história evolutiva de organismos (por exemplo, associando as relações obtidas por comparação de DNA ribossomal de aves com datação de fósseis, buscando estabelecer relações de ancestralidade).

No entanto, ao lidar com sequências moleculares, diferentes questões podem surgir. Por exemplo, o conceito de gene é dinâmico e mudou muito desde sua primeira definição. Além disso, genes podem sofrer diferentes processos evolutivos que alteram sua estrutura e/ou função, como mutações e rearranjos, ou ainda duplicações e perdas de função. Esses fatores fazem com que a relação 1:1 entre gene e organismo seja perdida. Por exemplo, uma mesma leguminosa pode possuir duas cópias do gene para a proteína leghemoglobina (***genes parálogos***). Além disso, muitas sequências do genoma não chegam à etapa de tradução, podendo conter elementos regulatórios ou transponíveis. Tais variações aumentam a complexidade e dificultam a interpretação das relações de descendência.

6.2. Aplicações

Ao classificarmos os organismos, atribuímo-lhes uma história evolutiva. Essa história, entretanto, é frequentemente desconhecida. Sendo assim, é necessário inferir a sequencia de mudanças que levaram ao surgimento de um novo organismo ou proteína. Contudo, existe apenas uma história verdadeira, que talvez jamais seja conhecida. Assim, ao empregarmos as técnicas filogenéticas, o objetivo é coletar e analisar dados capazes de fornecer a melhor estimativa para chegarmos à filogenia verdadeira. De certa forma, a obtenção de filogenias lembra a atuação de um historiador. Baseando-se em dados disponíveis no presente (tais como organismos vivos, fósseis e sequências moleculares), tenta-se obter uma imagem de como teria sido o passado.

Quando analisamos sequências de nucleotídeos ou aminoácidos para inferir uma filogenia, utilizamos informações derivadas das taxas evolutivas para determinar a sequência de eventos que levaram ao surgimento de novos organismos. A ***taxa de evolução*** molecular refere-se à velocidade na qual os organismos acumulam diferenças genéticas ao longo do tempo. Essa taxa é frequentemente definida pelo número de substituições por sítio (ou posição no alinhamento de sequências) por unidade de tempo e, portanto, são usadas para descrever a dinâmica das mudanças em uma linhagem ao longo de várias gerações.

As taxas evolutivas são empregadas quando se buscam estimativas temporais para datação de eventos evolutivos. Normalmente, se assume que as mudanças nas sequências se acumulam a uma taxa mais ou menos constante ao longo do tempo. Esse conceito é chamado de ***Hipótese do Relógio Molecular***. Entretanto, é conhecido que as taxas evolutivas são dependentes de vários fatores, tais como o tempo de geração, o tamanho da população e do próprio metabolismo, o que normalmente viola o modelo estrito de relógio molecular. Com base nestas informações, diversos modelos foram propostos para lidar com desvios no comportamento temporal de diferentes linhagens moleculares e, hoje em dia, são referidos como relógios moleculares relaxados.

Atualmente, a inferência filogenética é um campo de pesquisa à parte das outras ciências. Tornou-se uma ferramenta complementar para diversas áreas e indispensável para outras. Apesar de ter sido idealizada para desvendar apenas as relações evolutivas entre organismos, atualmente a filogenética molecular é aplicada a problemas muito mais diversos que este. Com o advento do relógio molecular estrito, foi possível aplicar a estimativa de tempo às filogenias e datar surgimento de espécies, disseminação de organismos e, até mesmo, entender grandes eventos biológicos que ocorreram no passado. Com a abordagem relaxada do relógio molecular, iniciou-se a utilização de modelos de dinâmica populacional que

comportam os eventos coletivos de grupos específicos. Ainda, com o avanço da capacidade de processamento computacional, vem sendo possível criar algoritmos capazes de reconstruir genomas ancestrais. Também a partir da filogenética molecular desenvolveu-se o campo da filogeografia. Segundo esta área do conhecimento, as filogenias podem ser utilizadas para verificar a distribuição geográfica de indivíduos. Neste contexto, outras técnicas, além das filogenias, são incorporadas às análises, incluindo a estruturação de genes, as análises de redes e as análises de haplótipos.

A filogenia molecular busca inferir a história evolutiva de organismos ou outras entidades biológicas (como proteínas e genes) a partir de sequencias de ácidos nucleicos ou aminoácidos. Ao investigar as relações entre diferentes espécies, análises de genes ribossomais são comumente empregadas, pois independentemente da espécie ou do organismo, os indivíduos possuirão genes codificantes de RNA ribossômico. Em contrapartida, quando se busca compreender as relações entre diferentes enzimas de uma mesma família é necessário utilizar sequências de aminoácidos, e não de nucleotídeos. Em determinadas situações, o genoma completo pode ainda ser utilizado para inferir a filogenia. Este é o caso de diversos vírus, especialmente quando se busca compreender a origem de novas variantes ou a disseminação de uma cepa. O alvo de estudo (isto é, sequência de nucleotídeos ou aminoácidos, gene ou genoma) depende, exclusivamente, do objetivo da análise e é um dos principais fatores a ser definido primariamente pelo pesquisador.

Atualmente, as filogenias funcionam como importantes ferramentas para diferentes áreas do conhecimento, incluindo as áreas de evolução, genética, epidemiologia, microbiologia, virologia, parasitologia, botânica e zoologia, dentre outras. Adicionalmente, de maneira inédita, a inferência filogenética foi utilizada como evidência para a resolução de crime e principal prova durante um impasse internacional envolvendo diferentes países. Em resumo, dependendo do objetivo, os métodos de construção de filogenias (inferência filogenética) são a base para diversas áreas e importantes objetos para o avanço computacional na análise de dados biológicos.

6.3. Representação de árvores

A Filogenética (termo obtido por união dos termos gregos para tribo e origem) é a ciência que busca reconstruir a história evolutiva dos organismos, levando em conta as sequências de nucleotídeos ou aminoácidos. As hipóteses sobre a história evolutiva são o resultado dos estudos filogenéticos e se chamam ***Filogenia***.

As filogenias ou árvores filogenéticas representam o contexto evolutivo dos organismos de forma gráfica. São formadas por nós (pontos) ligados por diversos ramos (linhas) (Figura 1-6). Os nós terminais, mais externos na filogenia, identificam os indivíduos, genes ou proteínas que foram amostrados e incluídos na análise filogenética. Geralmente representam o alvo de estudo do pesquisador e estão ligados aos nós mais internos na filogenia através de traços horizontais, chamados de ramos terminais (Figura 1-6).

Os nós internos, pelo contrário, representam indivíduos não amostrados. Eles identificam uma inferência evolutiva do ancestral comum mais recente dos ramos derivados daquele nó e se ligam a nós cada vez mais internos, através dos ramos internos. Por exemplo, na Figura 1-6, os grupos de nós terminais representados em verde possuem como ancestral comum o nó laranja, mais interno, enquanto os nós terminais azuis possuem como ancestral comum o nó lilás. Da mesma forma, o nó vermelho é a representação do indivíduo, gene ou proteína mais ancestral da filogenia que, através de processos evolutivos, deu origem aos nós laranja e lilás.

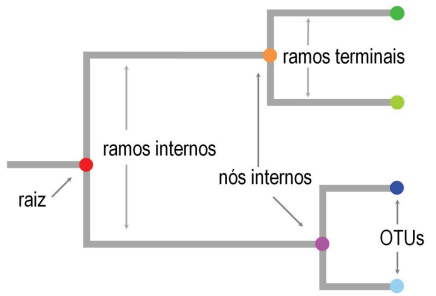


Figura 1-6: Nomenclatura associada a árvores filogenéticas.

O tamanho dos ramos horizontais pode ter diferentes significados, dependendo do método para inferência da filogenia, conforme veremos a seguir. No entanto, os ramos representados na vertical (Figura 1-6) não expressam qualquer significado, e seu tamanho não altera em nada a idéia filogenética. Como a análise pode ser feita em diferentes níveis, utilizando dados moleculares de genes, proteínas, indivíduos, espécies, gêneros, famílias, ou qualquer outro *taxon*, os nós terminais são amplamente denominados ***OTUs*** (*operational taxonomical units*), ou unidades taxonômicas operacionais (também chamados de folhas, Figura 2-6). A ordem e disposição exata das OTUs em uma filogenia é denominada ***topologia***.

Além da forma gráfica, as árvores filogenéticas podem também ser descritas na forma textual. Em vez do diagrama com linhas e pontos, as relações evolutivas são representadas por notações com parênteses. A estrutura da árvore da Figura 2-6, por exemplo, pode ser descrita linearmente como (Peixes pulmonados, (Anfíbios, (Mamíferos, (Tartarugas, (Lagartos, (Crocodilos, Aves)))))) ou (Peixes pulmonados + (Anfíbios + (Mamíferos + (Tartarugas + (Lagartos + (Crocodilos + Aves)))))). Estas notações foram desenvolvidas para utilização computacional da informação filogenética. Algoritmos e programas que realizam análises moleculares necessitam da informação na forma textual e, quando necessário, fornecem a saída para o usuário na forma gráfica.

Partindo do princípio de derivação

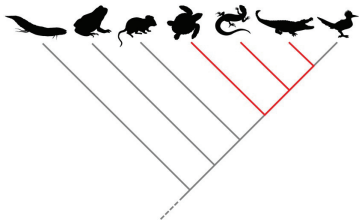


Figura 2-6: Árvore dicotômica dos grupos de vertebrados. As OTUs (nós terminais) estão representadas por ícones (peixes pulmonados, anfíbios, mamíferos, tartarugas, lagartos e serpentes, crocodilos e aves). Observe que o grupo dos répteis é parafilético (destacado em vermelho). O grupo seria considerado monofilético se incluísse as aves.

evolutiva, onde um organismo dá origem a outro (ou outros), podemos reconhecer dois principais processos na representação de filogenias: ***derivação dicotômica*** e ***derivação politômica***. No primeiro caso, cada nó interno dá origem a apenas dois ramos. Para espécies, por exemplo, a ramificação de um ancestral comum em dois ramos evidencia o processo de especiação. No segundo caso, três ou mais ramos surgem de um mesmo nó interno.

Apesar de árvores dicotômicas serem mais comuns e normalmente esperadas, em alguns casos, como a dispersão explosiva do HIV e do HCV, árvores politômicas representam melhor o processo evolutivo. Casos como estes, onde um ancestral comum origina simultaneamente várias linhagens descendentes, são chamadas de politomias verdadeiras (*hard polytomies*). Por outro lado, as politomias falsas (*soft polytomies*) são casos onde a topologia não foi bem resolvida por não haver certeza do padrão de ancestralidade, tornando múltipla uma divisão que se esperaria ser formada por uma série de divisões dicotômicas.

Assim, ao agruparmos as OTUs segundo a sua ancestralidade, podemos reconhecer diferentes padrões: grupos monofiléticos, parafiléticos e polifiléticos (Figura 2-6). Os grupos monofiléticos incluem todos os membros descendentes de um único

ancestral, assim como o próprio ancestral. Na Figura 2-6, por exemplo, as aves e os crocodilos são considerados um grupo monofilético, pois compartilham o mesmo ancestral comum. Da mesma forma, as aves, os crocodilos e os lagartos também podem ser considerados um grupo monofilético, pois se originaram de um mesmo ancestral. A análise das relações entre os grupos, neste caso, dependerá do objetivo do pesquisador. Adicionalmente, os grupos monofiléticos podem ser denominados ***clados*** por agruparem duas ou mais sequências que são descendentes de um mesmo ancestral (Figura 3-6a e b). A organização da topologia em que um clado está contido em outro é comumente chamada de clados aninhados ou clados embutidos (Figura 3-6c).

Os grupos parafiléticos, por sua vez, se originam de um único ancestral, mas nem todos os organismos derivados deste

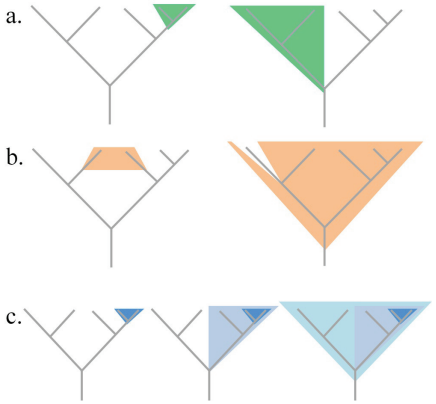


Figura 3-6: (a) exemplos de clados destacados em verde. (b) exemplos de organizações da topologia que não caracterizam a existência de um clado, destacados em laranja. (c) diferentes níveis de clados que podem estar embutidos em um clado de maior ordem. Observe que os clados de diferentes ordens, quando embutidos, formam clados monofiléticos.

ancestral fazem parte do grupo. Na Figura 2-6, os répteis são um grupo formado pelas tartarugas, lagartos e crocodilos, e seu ancestral comum esta na base do ramo que dá origem às tartarugas. No entanto, este ancestral comum também deu origem às aves e, por isso, os répteis não podem ser considerados um grupo monofilético, mas um grupo parafilético.

Finalmente, os grupos polifiléticos provêm de dois ou mais ancestrais diferentes. Nestas relações se encontram OTUs que apresentam características comuns, mas que possuem diferentes ancestrais comuns. Por exemplo, a condição endotérmica (animais que mantém a sua temperatura corporal constante) é apenas apresentada por aves e mamíferos. Por este critério, poderíamos agrupar estes dois grandes grupos sem, no entanto, compartilharem o mesmo ancestral comum direto (Figura 2-6). A organização destes grupos permite descrever características resultantes de convergência evolutiva, pois uma mesma característica se desenvolveu independentemente em diferentes grupos.

Sabendo das relações evolutivas entre os táxons e da existência de ancestrais comuns, as árvores podem ser representadas de maneira a evidenciar o ancestral mais antigo (árvore com raiz ou enraizada), ou apenas destacar as relações evolutivas entre os táxons, sem destacar qual a OTU mais ancestral (árvore sem raiz ou não enraizada) (Figura 4-6).

A ***raiz*** da filogenia é a espécie ou sequência ancestral a todo o grupo que está sob análise. Quando presente, a raiz aplica uma direção temporal à árvore, permitindo observar o sentido das mudanças evolutivas da raiz (mais antigo) aos ramos terminais (mais modernos). Uma árvore não enraizada, pelo contrário, reflete apenas a topologia estabelecida entre as OTUs, sem indicar o ancestral do grupo. Árvores não enraizadas podem ser confusas, e sua interpretação requer mais cuidado devido à facilidade em cometer erros de análise (Figura 4-6).

A identificação de uma raiz nas

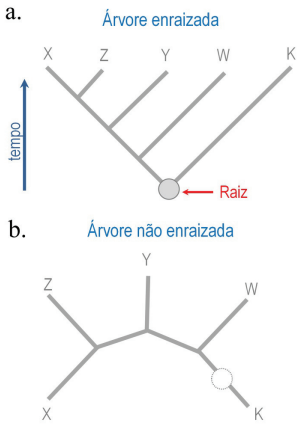


Figura 4-6: Comparação de árvores (a) enraizadas e (b) não enraizadas. No primeiro caso, é possível definir a direção das mudanças evolutivas, devido à presença do vetor tempo dado pela presença da raiz.

filogenias geralmente requer a inclusão de uma ou diversas OTUs que representem ***grupos externos***. Os grupos externos devem ser ancestrais comuns das OTUs em estudo, já conhecidos, que indicarão caracteres presentes em organismos mais próximos aos ancestrais, provendo um direcionamento para a interpretação dos processos evolutivos. Para o caso do estudo de HIV, por exemplo, é comum que os vírus da imunodeficiência de símios (SIV) sejam utilizados como grupo externo nas filogenias, pois sabidamente estes vírus deram origem ao HIV.

A adição de grupos externos aumenta o número de topologias diferentes que uma filogenia pode assumir. O número de árvores possíveis varia com o número de OTUs e com a presença ou ausência de raiz. Para mais de duas OTUs, a quantidade de possíveis árvores com raiz é sempre maior que o número de árvores sem raiz (observe abaixo as equações para obtenção do número de árvores com e sem raiz, onde n representa o número de OTUs). A possibilidade de inferência de diferentes topologias para os mesmos dados moleculares ressalta a extrema variabilidade de cenários possíveis na busca do verdadeiro evento evolutivo. É importante também ressaltar que, assim como a complexidade, o tempo computacional envolvido na construção das filogenias aumenta exponencialmente com o aumento de OTUs.

Em relação à topologia das árvores, a inversão de ramos derivados de um mesmo nó não altera a relação evolutiva apresentada pela árvore (Figura 5-6). Nesse sentido, a árvore filogenética pode ser comparada a um móbile: cada peça suspensa é livre para girar em seu eixo, ficando mais próxima ou mais distante espacialmente das outras peças, sem alterar a estrutura geral do objeto. Independentemente da posição destas OTUs, após o giro dos ramos, o mesmo ancestral comum será identificado e, por isso, não há qualquer alteração no significado da filogenia.

Quanto à nomenclatura de árvores filogenéticas, diferentes termos são empregados, tais como cladogramas, filogramas e dendrogramas (Figura 6-6). Um ***cladograma*** é uma árvore simples, que retrata as relações entre os nós terminais. Pelo contrário, uma árvore aditiva (árvore métrica ou ***filograma***) apresenta informações adicionais, pois o comprimento dos ramos é proporcional a algum atributo, como quantidade de mudança. Por sua vez, uma árvore ultramétrica (ou ***dendrograma***) constitui um tipo especial de filogenia devido aos seus ramos serem equidistantes da raiz. Os dendrogramas podem, desta forma, retratar o tempo evolutivo. É importante ressaltar que alguns autores denominam qualquer filogenia como cladograma, o que pode ser confuso.

O tipo de dado molecular a ser empregado nas análises também deve ser levado em conta. Sequências de aminoácidos são mais conservadas que sequências de ácidos nucleotídeos em decorrência da degeneração do código genético. São, portanto, úteis em análises de produtos de genes ou espécies que visam entender fenômenos que aconteceram há amplos períodos de tempo evolutivo. Além disso, por formarem um conjunto de pelo menos 20 membros (contra quatro membros presentes em DNA ou RNA), sua variação pode ser mais significativa.

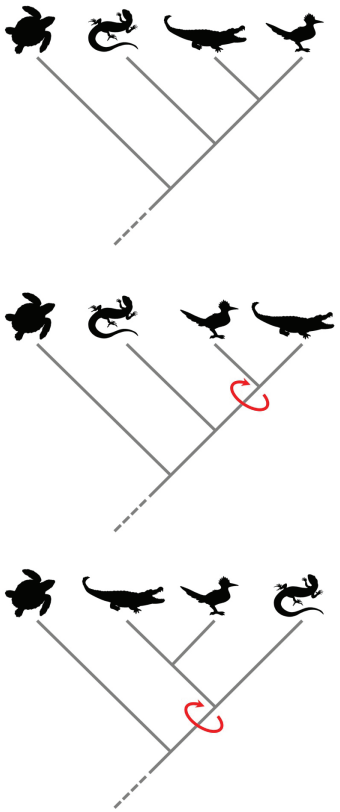


Figura 5-6: A porção terminal da árvore dos vertebrados (representada na Figura 2-6) foi rearranjada de diferentes maneiras (as setas indicam o ponto de rotação). Conforme a analogia de um móbile, todas elas representam a mesma relação evolutiva.

A despeito desta diferença no volume de informação, com a popularização do sequenciamento de ácidos nucleicos, especialmente DNA, sequências de nucleotídeos passaram a ser as mais empregadas em estudos de filogenia. Ácidos nucleicos são mais propensos a alterações, podendo sofrer transições (quando ocorre a troca de uma purina por outra purina, ou de uma pirimidina por outra pirimidina) e

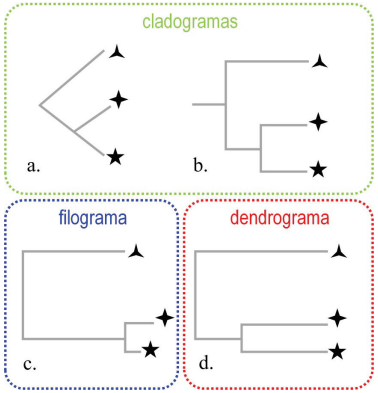


Figura 6-6: Nomenclatura de árvores filogenéticas. Observe que os cladogramas *a* e *b* são equivalentes, mas o filograma *c* e o dendrograma *d* não o são.

transversões (quando ocorre a troca de uma purina por uma pirimidina ou vice-versa), além de inserções ou deleções de pares de base que interferem no quadro de leitura. Essa variabilidade pode ser interessante no estudo de eventos mais recentes do ponto de vista evolutivo.

É preciso, assim, conhecer o caso de estudo e o tipo de pergunta que se busca responder com cada filogenia. Ao lidarmos com genes de diferentes espécies, por exemplo, é importante saber da existência e disposição de íntrons, da necessidade de lidar com o gene inteiro ou apenas parte dele ou da necessidade de incluir regiões regulatórias para a análise.

Um exemplo recente da aplicação de análises filogenéticas está no caso da identificação da origem da linhagem do vírus influenza H1N1, envolvido no surto de gripe de 2009. Para tanto, Smith e colaboradores empregaram genomas completos de influenza isolados de diferentes localidades e hospedeiros, e construíram árvores filogenéticas para cada uma das oito regiões do genoma buscando identificar a fonte de cada rearranjo presente no vírus envolvido no surto. Por meio das árvores obtidas, foi

possível rastrear a contribuição genética dos vírus isolados de aves, suínos e humanos (Figura 7-6). Assim, o emprego da filogenia neste trabalho permitiu não apenas caracterizar o vírus do ponto de vista molecular, como também reconstruir a história evolutiva do agente etiológico de uma pandemia.

6.4. Distância genética

A formulação de modelos evolutivos é uma maneira de descrever matematicamente os processos que moldam as mudanças nas sequências de nucleotídeos ou aminoácidos dos organismos ao longo do tempo. Do ponto de vista molecular, estas mudanças podem ser resultado de diferentes forças evolutivas que reorganizam a sequência e a própria estrutura dos genes.

Um modelo geral para descrever de maneira eficaz estas alterações evolutivas deveria considerar os processos de substituição, inserção, deleção e duplicação, bem como ocorrência de transposição ou até mesmo de retrotransposição. Contudo, apesar de estes fenômenos serem claros agentes na modelagem dos genomas, matematicamente ainda não é factível colocá-los como componentes de modelos que

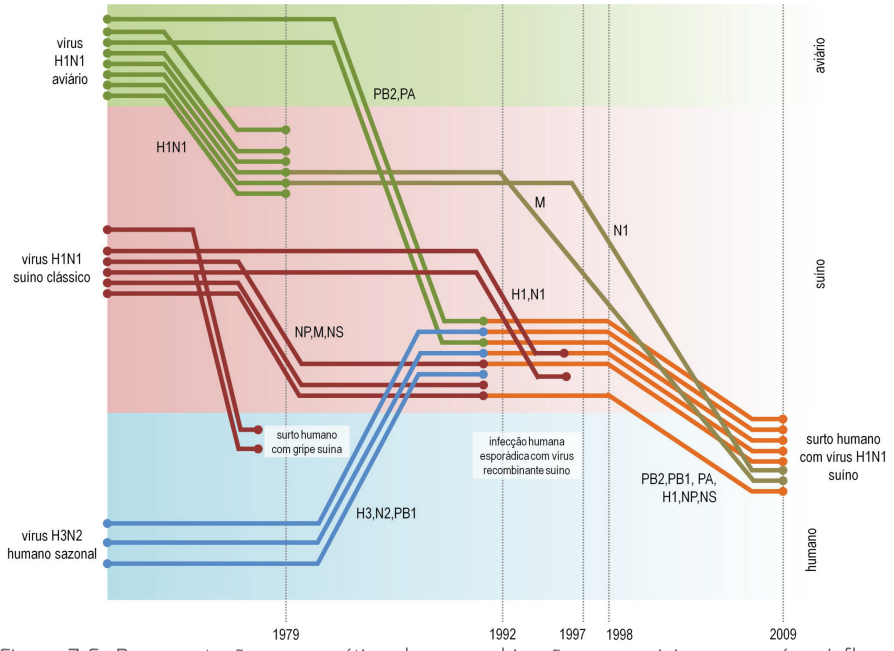


Figura 7-6: Representação esquemática das recombinações que originaram o vírus Influenza envolvido no surto de gripe suína em 2009. Diferentes linhas representam diferentes regiões do genoma do vírus. Observe a interação entre vírus de origens aviária, suína e humana em eventos que datam, pelo menos, desde 1990. Os eventos de recombinação e as análises temporais foram baseadas em análises filogenéticas (Adaptado de Smith e colaboradores, *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic*. *Nature*, *459*, 1122-1125, **2009**).

expliquem inteiramente o processo evolutivo.

Assim, devido à grande relevância dos mecanismos de substituição para a evolução dos genomas em diferentes organismos e da disponibilidade de modelos de probabilidade estatística que expliquem este processo, as trocas têm sido o principal alvo para o desenvolvimento de modelos matemáticos e compõem a base de diversos métodos de inferência filogenética.

Após a divergência de duas sequências a partir de seu ancestral comum, de forma dicotômica, fenômenos evolutivos garantirão as mudanças nas sequências de nucleotídeos de forma independente (Figura 8-6). Uma medida tradicional para expressar o número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência é chamada de ***distância genética***. Esta informação é uma medida quantitativa da dissimilaridade genética entre diferentes OTUs, e permite estabelecer uma estimativa relativa da quantidade de mudanças que ocorreram desde a divergência.

É também um importante conceito na construção de filogenias, pois está diretamente relacionada com a relação evolutiva entre duas OTUs: uma menor distância genética indica uma relação evolutiva mais próxima, enquanto que um valor maior sugere uma derivação evolutiva proporcionalmente maior. Tipicamente, a informação da distância genética é incorporada à inferência filogenética na definição do tamanho dos ramos. No entanto, além desta informação é necessária uma escala de distância que especifique o número de mudanças que ocorreram ao longo do ramo.

O método mais simplista para avaliar a distância genética entre duas sequências é conhecido como ***distância p***. Este método é baseado na contagem das diferenças dividida pelo número total de sítios do alinhamento. Se oito sítios são diferentes entre duas sequências homólogas com tamanhode 100pb, a distância p obtida será 0,08. Este resultado reflete a porcentagem de sítios diferentes em relação ao tamanho total da

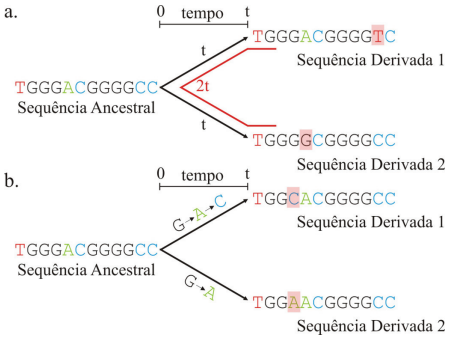


Figura 8-6: Após a divergência de dois organismos a partir de seu ancestral comum, seus genomas acumularão diferenças independentemente. (a) A medida da dissimilaridade genética entre duas sequências homólogas ao longo do tempo é chamada de distância genética, e a relação temporal entre duas sequências divergentes é dada por $2t$. (b) A ocorrência de múltiplas substituições ao longo do tempo na divergência de sequências homólogas pode mascarar as verdadeiras diferenças entre as sequências. Apesar de ocorrerem dois eventos de mutação na sequência derivada 1, apenas o último evento é observado, pois ocorreram no mesmo sítio. Os quadrados em vermelho evidenciam as diferenças em relação às sequências ancestrais.

sequência, e geralmente é utilizado na especificação da escala de distância das filogenias (Figura 18-6).

A variação genética em um determinado sítio pode decorrer de diferentes processos e resultar em mais de uma substituição. As múltiplas substituições, ou *multiple hits*, ocorrem naturalmente e podem subestimar o verdadeiro número de mudanças no cálculo da distância p, já que "escondem" as diversas trocas de nucleotídeos ou aminoácidos. Na figura 8-6b, por exemplo, apesar de ocorrerem duas substituições no mesmo sítio ao longo de um dos ramos, aparentemente a sequência derivada parece ter sofrido somente um evento evolutivo. Sendo assim, a

relação entre as diferenças nas sequências e o tempo decorrido da divergência nem sempre é linear, especialmente devido à ocorrência das múltiplas substituições em um mesmo sítio.

Devido à ineficácia da distância p em efetivamente estimar a distância genética entre duas sequências, diferentes modelos probabilísticos foram desenvolvidos para descrever as mudanças entre os nucleotídeos e corrigir a distância observada. Tais modelos implicam no uso de diversas suposições simples a respeito das probabilidades de substituição de um nucleotídeo por outro, mas garantem uma aproximação da realidade quando sustentadas por uma taxa de mutação fidedigna.

Estas técnicas de correção são comumente conhecidas por ***modelos de substituição*** (ou matrizes de susbtituição), e garantem a conversão da distância observada em medidas de distâncias evolutivas próximas da realidade, permitindo reconstruir a história evolutiva dos organismos.

Diversos modelos de substituição foram propostos para explicar as trocas de nucleotídeos em sequências de DNA, reduzindo a complexidade do processo evolutivo a um padrão de mudança simples que consegue ser explicado através de poucos parâmetros. Todos estes modelos, no entanto, de alguma forma são inter-relacionados, diferindo principalmente no número de parâmetros utilizados para explicar estas substituições. Devido à influência do modelo de substituição na inferência de filogenias, a escolha de um método particular deve ser justificada. A estratégia mais simples é utilizar os modelos que comportam o maior número de variáveis, embora a complexidade não esteja diretamente relacionada à melhor qualidade de análise das sequências. Com o aumento de parâmetros, o sistema se torna mais complexo, aumentando a probabilidade de erro e exigindo um maior processamento computacional. Assim, é necessário verificar os alinhamentos caso-a-caso para atribuir o melhor modelo de substituição na inferência filogenética.

A substituição de nucleotídeos ou aminoácidos em uma sequência é usualmente modelada sob a forma de um processo quase aleatório. Devido ao caráter dinâmico desta aleatoriedade, é necessário enquadrar as substituições, seguindo certos pressupostos. Assim, as substituições são descritas por um processo de Markov homogêneo, onde a probabilidade de substituição de um nucleotídeo X pelo Y não depende do estado prévio do nucleotídeo X.

As probabilidades de mudança de um nucleotídeo para outro (ou de um aminoácido para outro) são especificadas através de uma matriz 4x4 das taxas de substituição (ou 20x20 no caso dos aminoácidos) que especificam com qual taxa cada um dos nucleotídeos ou aminoácidos poderá mudar para outro. É necessário assumir também que os eventos de substituição sejam independentes ao longo dos sítios das sequências, e ainda, possuam um caráter reversível. Além disso, devem especificar a frequência estacionária dos nucleotídeos, ou ***frequência de equilíbrio***, onde será atribuída a provável proporção de cada um dos caracteres na sequência.

Para sequências de nucleotídeos, o modelo de substituição mais simples foi proposto por Jukes e Cantor em 1969 (JC69). Segundo este modelo, as mudanças entre os nucleotídeos podem ocorrer com a mesma probabilidade, assumindo uma frequência estacionária igual para todos (cada nucleotídeo tem 25% de chance de ocorrer na sequência).

Com o advento da publicação das primeiras sequências de genoma mitocondrial, na década de 1980, se observou que as transições eram muito mais comuns que as transversões. Devido à uniformidade do método proposto por Jukes e Cantor, foi necessário criar um modelo que acomodasse essas diferenças.

Assim, o modelo proposto por Kimura (K80 ou K2P) cria as variáveis α e β para representar, respectivamente, as taxas de transição e de transversão. Apesar da inclusão de dois parâmetros, as frequências de equilíbrio se mantêm constantes em ¼ para cada nucleotídeo. Em 1981, Kimura adiciona um terceiro parâmetro (γ) ao modelo já proposto, passando a ser identificado como K3P. A atualização do modelo permitiu dividir as taxas de transversão em duas variáveis.

Alguns genomas apresentam uma grande quantidade de guaninas e citosinas em relação a timinas e adeninas. Se algumas bases são mais frequentes que outras, será esperado que algumas substituições ocorram com mais frequência que outras. O modelo criado por Felseinstein (F81) acomoda essas observações e permite que as proporções individuais de cada nucleotídeo (frequência

estacionária) sejam diferentes de ¼. É importante ressaltar que este modelo considerará a mesma proporção de bases em todas as sequências envolvidas no alinhamento. Se diferentes sequências possuem diferente composição de bases, a pressuposição principal do modelo será violada.

O modelo HKY85, proposto por Hasegawa, Kishino e Yano, essencialmente mistura os modelos K2P e F81. Além de supor que a frequência das bases é variável, este modelo permite que transições e transversões ocorram com taxas diferentes.

Posteriormente, o modelo GTR (*generalised time-reversible*), o mais complexo dos modelos aqui apresentados, foi desenvolvido a partir do HKY85 com o intuito de acomodar diferentes taxas de substituição e diferentes frequências de bases. Este modelo requer seis parâmetros para taxa de substituição e quatro parâmetros para a frequência das bases, misturando todos os modelos aqui descritos.

Atualmente, além destes mais de 200 modelos de substituição podem ser aplicados a alinhamentos de nucleotídeos. Alguns programas, como Modeltest e Jmodeltest, são capazes de selecionar o modelo de substituição que melhor se ajusta a um dado alinhamento.

Uma importante extensão desses modelos de substituição incorpora a possibilidade de variação nas taxas evolutivas entre os sítios, permitindo ao modelo mais realismo. Assim, para cada sítio no DNA será atribuída uma probabilidade de evolução a uma taxa contida em um intervalo discreto de probabilidades. O método que garante a heterogeneidade de taxas evolutivas é modelado através de uma distribuição gama (Γ), que considera um número específico de taxas de evolução para os sítios do DNA.

A aplicabilidade deste modelo nas inferências filogenéticas é facilitada pela simplicidade do método, já que apenas um único parâmetro (α) controla a forma da distribuição gama. Quando $\alpha < 1$, existe um grande número de taxas de evolução entre os sítios das sequências em análise, ou seja, quanto maior α, menor a heterogeneidade. Algumas vezes, uma proporção de sítios invariáveis (I), no qual uma determinada proporção de sítios é assumida como incapaz de sofrer substituição, pode também ser usada para modelar a heterogeneidade entre os sítios.

Ao contrário dos modelos de substituição de nucleotídeos, os modelos que explicam as trocas de aminoácidos são tradicionalmente empíricos. A partir da análise de alinhamentos de proteínas com identidade mínima de 85% Dayhoff, em 1970, desenvolveu uma série de matrizes de probabilidade que explicavam as mudanças de aminoácidos ao longo do tempo.

As matrizes PAM, como ficaram conhecidas, correspondem a modelos de evolução nos quais os aminoácidos são substituídos aleatoriamente e independentemente, de acordo com uma probabilidade pré-definida que depende do próprio aminoácido.

Em 1992, um novo modelo de substituição de aminoácidos é criado por Henikoff e Henikoff. A análise de sequências de proteínas distantes evolutivamente, possibilitada pelo modelo de Henikoff-Henikoff, estabeleceu as bases para a criação das matrizes BLOSUM. As matrizes desta série foram identificadas por números (por exemplo, BLOSUM62) que se referem à porcentagem mínima de identidade dos blocos dos aminoácidos utilizados para construir o alinhamento. Matrizes similares, como GONNET e JTT, surgiram na mesma época.

Em 1996, foi proposto um modelo de substituição específico para proteínas codificadas pelo DNA mitocondrial, onde foi observado desvio de transições entre aminoácidos em relação às proteínas codificadas pelo material genético nuclear. Essa matriz, criada por Adachi e Hasegawa, foi chamada de mtREV.

Finalmente, em 2001, Whelan e Goldman propõem a matriz WAG, baseada em combinação e ampliação de vários modelos de substituição anteriores. Tal matriz é considerada superior às suas antecessoras para descrever filogenias de proteínas globulares.

6.5. Inferência filogenética

A reconstrução filogenética, ou seja, a reconstrução da história evolutiva de organismos, é um complexo processo que envolve uma série de etapas. O alinhamento, além de ser o primeiro passo, é um importante ponto para a inferência de filogenias (ver capítulo 4). Um alinhamento preciso, além de garantir maior confiabilidade nas análises posteriores, é requerido por todos os métodos de inferência filogenética para construção da árvore.

Depois que o alinhamento foi proposto, diversos métodos podem ser usados para estimar a filogenia das sequências estudadas.

Podemos dividir estes métodos em dois principais grupos: métodos quantitativos e métodos qualitativos (Tabela 1-6). Estes grupos diferem na forma como os dados são tratados, refletindo diretamente como os dados do alinhamento serão inicialmente processados.

Os ***métodos quantitativos*** se baseiam na quantidade de diferenças entre as sequências do alinhamento para calcular uma árvore final. Já os ***métodos qualitativos*** constroem diversas filogenias que são classificadas seguindo uma determinada qualidade (critério). A filogenia que obtiver o maior valor associado à tal qualidade será a filogenia resultante.

Os métodos quantitativos compreendem os ***métodos de distância***. Estes métodos convertem o alinhamento em matrizes de distância par-a-par para todas as sequências incluídas. Dentro destes algoritmos destacam-se dois métodos principais: UPGMA e aproximação dos vizinhos. Devido à grande eficiência computacional, estes métodos geralmente são utilizados para construção de uma filogenia inicial, que posteriormente é submetida a algum método do grupo qualitativo. Como principal ponto negativo, estes métodos apresentam apenas uma filogenia como resultado final (ver adiante).

Idealmente, todas as possíveis árvores para um dado alinhamento deveriam ser analisadas para garantir a escolha da melhor filogenia. Para isso, é necessário atribuir certos parâmetros que avaliem, dentre todas as árvores, aquela que explica as relações evolutivas de forma mais precisa.

Assim, os métodos qualitativos envolvem algoritmos que atribuem um critério de otimização para escolher a melhor filogenia. Nestes métodos, diversas filogenias são construídas e, seguindo um critério definido pelo algoritmo utilizado, uma filogenia será identificada como a que melhor explica a relação evolutiva entre os OTUs. O critério é utilizado para atribuir um valor a cada filogenia e ordená-las segundo este valor.

Estes métodos têm a vantagem de requerer uma função explícita para escolha das filogenias, sendo portanto independente da escolha do operador. No entanto, devido ao caráter de sua análise, são métodos mais refinados e intrinsecamente mais demorados computacionalmente. Três critérios de otimização são tradicionalmente empregados na inferência de filogenias: (a) Máxima

Tabela 1-6: Comparação entre os tipos de métodos para inferência de filogenias.

Tipo	Método	Princípio	Programa
Métodos Quantitativos	UPGMA	Agrupa sequencialmente as OTUs com menor distância evolutiva entre si	Geneious MEGA
	Aproximação dos vizinhos	Busca a árvore com a menor soma total de ramos	MEGA Geneious HyPhy
Métodos Qualitativos	Máxima Parcimônia	Busca a filogenia com menor número de eventos evolutivos	PAUP MEGA Mesquite
	Máxima Verossimilhança	Busca a árvore com o valor de maior verossimilhança entre todas as filogenias construídas	PAUP PAML phyML MEGA
	Estatística Bayesiana	Amostra um número representativo de filogenias a partir do espaço amostral total de árvores e busca a mais provável	Mr. Bayes BEAST BAMBE

Parcimônia, (b) Máxima Verossimilhança e (c) Inferência Bayesiana.

Por se tratarem de métodos que buscam uma única filogenia entre diversas árvores, os métodos qualitativos exigem algoritmos que vasculhem o maior número possível de filogenias em busca da melhor árvore. Dois grupos de algoritmos são destacados: os algoritmos exatos e os algoritmos heurísticos. Atualmente, devido ao tempo e à exigência computacional, os métodos heurísticos são preferidos aos exatos. No entanto, qualquer um deles pode ser aplicado aos métodos qualitativos de inferência filogenética. Como desvantagem dos métodos qualitativos, repetidos processos de procura em um mesmo conjunto de sequências podem levar a resultados diferentes, dependendo da árvore que é construída inicialmente pelo algoritmo.

Os métodos exatos buscam todas as filogenias possíveis para um grupo de sequências. O funcionamento destes métodos geralmente envolve a seleção aleatória inicial de três OTUs para a construção de uma árvore filogenética não enraizada. Por tentativa, um a um, novas OTUs, também tomadas aleatoriamente do alinhamento, são inseridas em diferentes posições na árvore. Esse procedimento é repetido até todos os táxons serem inseridos, garantindo que todas as filogenias possíveis para o alinhamento dado sejam geradas.

A partir da aplicação de um critério de otimização (dado pelo método qualitativo) para classificar as filogenias e ordená-las segundo este valor, é possível organizar um espaço virtual que contém todas as filogenias possíveis para o alinhamento empregado. É importante lembrar que, tomando poucas sequências, milhões de árvores podem ser geradas. Este conjunto total de filogenias é comumente chamado de ***espaço amostral***. Como exemplo, podemos organizar o espaço amostral de filogenias originadas a partir de um alinhamento de dez sequências em um gráfico bidimensional baseado no valor atribuído pelo critério de otimização a cada árvore (Figura 9-6). Nestas condições, será possível observar que algumas árvores possuem valores maiores que outras, formando picos que agrupam as melhores filogenias. Da mesma forma, entre diferentes picos existem vales representados por árvores com valores menores e, portanto, menos consistentes.

Os métodos de busca exaustiva construirão um espaço amostral de árvores através de métodos específicos de modificação das filogenias. Por acumularem um grande número de resultados, estes métodos exigem um tempo computacional muito elevado, por vezes tornando-se proibitivos.

Os algoritmos de busca heurística procuram pela melhor filogenia em um subconjunto de todas as filogenias possíveis. Apesar de serem muito mais rápidos computacionalmente, estes métodos não garantem que a filogenia correta seja encontrada, pois apenas algumas árvores do espaço amostral total serão consideradas. Ainda assim, estes métodos tem mostrado grande eficiência.

Atualmente, os principais métodos qualitativos de inferência filogenética incorporam algoritmos de busca heurística para amostrar as filogenias do espaço amostral virtual. Usualmente, estes algoritmos de busca são executados em dois passos. Primeiramente,

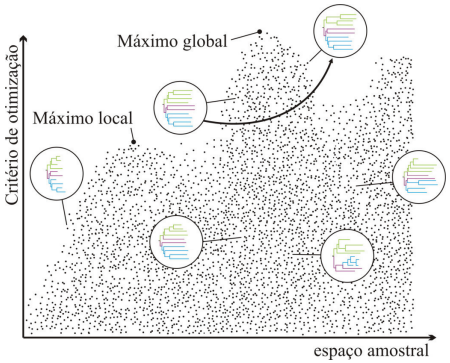


Figura 9-6: Descrição de parte do espaço amostral das possíveis filogenias para um determinado sistema, ordenadas segundo um valor atribuído pelo critério de otimização. Cada ponto no gráfico representa uma topologia diferente inferida a partir de um conjunto de dez sequências homólogas. O espaço amostral, neste caso, é definido por 2.027.025 filogenias e apresenta, segundo o critério de otimização, dois máximos locais e um máximo global, que contém as melhores filogenias. Em destaque, algumas filogenias exemplificando as possibilidades de arranjo dos ramos. A seta indica a mudança de topologia da filogenia e o consequente aumento de seu valor dado pelo critério de otimização.

diferentes árvores são construídas e, após encontrar a melhor árvore guiada por um critério de otimização, aplica-se um algoritmo para modificar aleatoriamente o arranjo dos ramos. Este método permite testar se outros arranjos são ou não mais consistentes.

Devido ao grande número de métodos para inferência filogenética, a decisão quanto ao uso de cada um é de grande importância para a interpretação do resultado final: a filogenia.

Ao escolher um método, é fundamental verificar o poder (tamanho e quantidade de sequências necessária para resolver a filogenia), a eficiência (habilidade de estimar a filogenia correta com um número limitado de dados), a consistência (habilidade de estimar a filogenia correta com um número de dados ilimitado) e a robustez (habilidade de estimar a filogenia correta quando certos pressupostos da análise são violados).

Até o momento, não existe um método que apresente todas estas características simultaneamente e garanta a reconstrução filogenética correta. É importante, sobretudo, conhecer a biologia do organismo (ou dos organismos) em questão para que a escolha do método tenha, além de tudo, uma justificativa biológica.

6.6. Abordagens quantitativas

UPGMA

O método baseado em distâncias ***UPGMA*** (*unweighted pair-group method using arithmetic averages*, ou método de agrupamento par a par usando médias aritméticas não ponderadas) foi proposto por Sneath e Sokal, em 1973, e é o método mais simples para reconstrução filogenética. O UPGMA parte do pressuposto de que todas as linhagens evoluem a uma taxa constante (hipótese do relógio molecular).

No UPGMA, uma medida de distância evolutiva é computada para todos os pares de sequências utilizando um modelo evolutivo. Após, estas distâncias são organizadas na forma de uma matriz, conforme ilustrado abaixo:

O agrupamento das sequências é iniciado pelo par com menor distância. Supondo que $d_{1,2}$ seja a menor distância no exemplo acima, as sequências *1* e *2* são agrupadas com um ponto de ramificação na metade dessa distância ($d_{1,2/2}$). As sequências *1* e *2* são então combinadas em uma entidade composta, agora denominada y, e a distância entre esta entidade y e as outras sequências é computada (observe abaixo).

Supondo que $d_{y,3}$ seja a menor distância, y e *3* são combinados em uma nova entidade composta, digamos, z. Seu ponto de

Sequências	*1*	*2*	*3*	*4*
2	$d_{1,2}$			
3	$d_{1,3}$	$d_{2,3}$		
4	$d_{1,4}$	$d_{2,4}$	$d_{3,4}$	
5	$d_{1,5}$	$d_{2,5}$	$d_{3,5}$	$d_{4,5}$

ramificação é calculado levando em conta a distância de cada membro de y (*1* e *2*) em relação a *3* e dividindo por *2*, ou seja, $(d_{1,3}+d_{2,3})/2$. O mesmo procedimento se repete, calculando a menor distância entre z e outra sequencia (suponhamos que seja a sequência *4*). Calculam-se a distância de cada membro de z até *4*, divide-se o somatório das distâncias por dois e cria-se uma nova sequência composta. O mesmo procedimento

Sequências	$y_{(1,2)}$	*3*	*4*
3	$d_{y,3}$		
4	$d_{y,4}$	$d_{3,4}$	
5	$d_{y,5}$	$d_{3,5}$	$d_{4,5}$

é repetido até que existam apenas duas sequências a serem agrupadas (comumente, uma sequência simples e uma entidade composta).

Ao empregar sequências de DNA ou

proteína proximamente relacionadas, o UPGMA pode construir duas ou mais "árvores empatadas" (*tie trees*). Essas árvores surgem quando dois ou mais valores de distância na matriz se mostram idênticos. É possível representar todas as árvores empatadas, mas essa abordagem é pouco útil, uma vez que tais árvores são muito semelhantes e surgem por erros de estimativa das distâncias. Para tais casos, sugere-se apresentar uma única árvore, geralmente a árvore consenso do *bootstrap* (ver seção 6.8).

Por se basear na hipótese do relógio molecular, o UPGMA pode levar à obtenção de topologias falsas quando tal hipótese não for satisfeita pelos dados. Sabe-se que o método é muito sensível a variações nas taxas evolutivas entre linhagens, fato este que levou a proposição de métodos onde as variações são ajustadas para a obtenção de sequências que satisfaçam o relógio molecular. Apesar disso, devido ao surgimento de métodos mais robustos e mais eficientes em lidar com dados não uniformes, o UPGMA encontra-se praticamente abandonado como alternativa para reconstrução filogenética.

Aproximação dos Vizinhos

O método de ***aproximação dos vizinhos*** (*neighbor joining* ou NJ) foi proposto por Saitou e Nei em 1987. Este método se baseia em um aceleramento dos algoritmos de evolução mínima que existiam até então. Em sua versão original, estes algoritmos buscavam a árvore com menor soma total de ramos, de maneira que todas as árvores possíveis precisavam ser construídas para que se verificasse qual delas apresentava a menor soma. O algoritmo de NJ facilitou esse processo, tendo o princípio de evolução mínima implícito no processo e produzindo apenas uma árvore final.

Para construir a filogenia, o NJ começa por uma árvore totalmente não resolvida (topologia em estrela) (Figura 10-6). Tendo como base uma matriz de distâncias (semelhante à matriz inicial construída pelo método de UPGMA) entre todos os pares de sequências, construída a partir da aplicação de um modelo de substituição (conforme descrito na seção 6.4), o par que apresentar a menor distância é identificado, unido por um nó (que representará o ancestral comum deste par de sequências) e incorporado na árvore (na Figura 10-6, f e g são unidos pelo nó u). As distâncias de cada sequência do par são recalculadas em relação ao novo nó u, assim como as distâncias de todas as outras sequências são recalculadas em relação ao novo nó u. O algoritmo reinicia, substituindo o par de vizinhos unidos pelo novo nó e usando as distâncias calculadas no passo anterior.

Quando duas somatórias de ramos são iguais, a decisão sobre quais ramos unir depende do programa empregado. Alguns optam pela primeira sequência apresentada no arquivo de dados, enquanto outros escolhem aleatoriamente qual dos pares deve ser unido primeiro. Árvores empatadas (*tie trees*) são raras com o uso de NJ, e recomenda-se o emprego da árvore consenso do *bootstrap* (ver seção 6.8) para evitá-las. Uma variação do algoritmo NJ, o BIONJ tem se mostrado ligeiramente melhor que o NJ em casos pontuais; no entanto, conserva o mesmo princípio do algoritmo.

6.7. Abordagens qualitativas

Parcimônia

O princípio de parcimônia foi proposto por Guilherme de Occam (ou *William of Ockham*) no século XVII. Occam defendia que a natureza é por si só econômica e opta por caminhos mais simples. O pensamento se espalhou por diversas áreas do conhecimento e, atualmente, seu princípio é conhecido como Navalha de Occam.

Historicamente, a parcimônia teve um papel muito importante no estabelecimento da disciplina de filogenética molecular. Desde 1970, foi o critério de otimização mais utilizado para inferência de filogenias.

Contudo, atualmente a ***máxima parcimônia*** foi substituída por outros métodos, como máxima verossimilhança e inferência Bayesiana devido, principalmente, às simplificações nos processos evolutivos assumidas pelo método e, sobretudo, nas limitações de seu uso. Apesar disso, a máxima parcimônia ainda está integrada ao campo da inferência filogenética por ser um método rápido e, em alguns casos, muito efetivo.

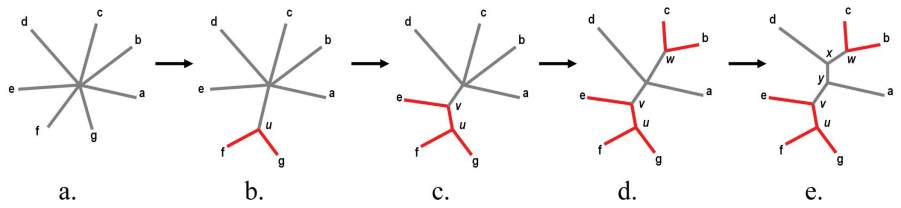


Figura 10-6: Começando com uma árvore em estrela (a), a matriz de distâncias é calculada para identificar o par de nós a ser unido (nesse caso, f e g). Estes são unidos ao novo nó u (b). A porção em vermelho é fixada e não será mais alterada. As distâncias do nó u até os nós a-e são calculadas e usadas para unir o próximo vizinho. No caso, u e e são unidos ao recém criado nó v (c). Mais duas etapas de cálculo levam à árvore em (d) e então à árvore em (e), que está totalmente resolvida, encerrando o algoritmo.

A aplicação do princípio de máxima parcimônia nas reconstruções filogenéticas é conceitualmente simples: dentro de um conjunto de filogenias, aquela filogenia que apresentar o menor número de eventos evolutivos (substituições) deve ser a mais provável para explicar os dados do alinhamento.

Metodologicamente, o critério de parcimônia deve determinar a quantidade total de mudanças na filogenia, descrevendo o tamanho dos ramos. Adicionalmente, a parcimônia guia a busca, entre todas as árvores possíveis, daquela filogenia que minimiza os passos evolutivos de forma máxima sendo, portanto, a filogenia de máxima parcimônia.

Assim que uma determinada filogenia é proposta, o método calculará as probabilidades de mudanças dos nucleotídeos desde os ramos terminais até os ramos mais ancestrais da árvore. Por se tratar de um método qualitativo, a parcimônia considera cada sítio do alinhamento individualmente e calcula as probabilidades de ocorrência dos quatro nucleotídeos nos táxons ancestrais.

Devido ao caráter probabilístico do método, é necessário que certas pressuposições sejam estabelecidas para especificar o custo de substituição dos nucleotídeos. A forma mais simples do método (Parcimônia de Wagner) assume que as substituições de nucleotídeos tem custo 1, enquanto que a não alteração não é penalizada (Figura 11-6a). No entanto, esquemas um pouco mais complexos que levam em consideração as questões biológicas envolvidas no processo evolutivo foram propostas. Um esquema comum de matriz com custo desigual, proposto para especificar as transições e as transversões, leva em consideração a diferença na probabilidade de mudança entre purinas e pirimidinas (Figura 11-6b). Comumente, a matriz é especificada sem que constem os respectivos nucleotídeos, no entanto, por convenção são atribuídos nas linhas e colunas em ordem alfabética (A, C, G e T).

Para o método de parcimônia, apenas sítios variáveis são considerados informativos. Estes sítios devem apresentar dois caracteres diferentes presentes em, no mínimo, dois indivíduos (Figura 12-6b). Aqueles sítios que não apresentam variação ou apresentam autapomorfias (caracter diferente presente em apenas um indivíduo) serão descartados automaticamente das análises.

Devido ao tamanho dos alinhamentos e ao número de OTUs incluídas para a inferência de filogenias, foi necessário que algoritmos fossem desenvolvidos para acelerar os cálculos na busca pela árvore de máxima

a.

Matriz de custo igual

	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

b.

Matriz de custo desigual

	A	C	G	T
A	0	4	1	4
C	4	0	4	1
G	1	4	0	4
T	4	1	4	0

Figura 11-6: Matrizes de custo aplicadas ao método de máxima parcimônia para penalizar as substituições de um nucleotídeo por outro. (a) Matriz de custos iguais para todas as mudanças entre nucleotídeos. (b) Matriz de custo desigual, considerando a maior probabilidade de ocorrência de transições em relação às transversões ao longo do processo evolutivo.

parcimônia. Algoritmos de programação dinâmica são capazes de lidar com a atribuição de custos e realizar os devidos cálculos para escolha da filogenia com o menor custo. Diversos algoritmos foram desenvolvidos, embora a parcimônia de Sankoff, desenvolvida em 1975, tenha se tornado uma das mais populares.

Após a atribuição de uma matriz de custo e a proposição de uma filogenia, o algoritmo utilizará cada um dos sítios informativos do alinhamento independentemente para cálculo dos custos (Figura 11-6).

Considere a matriz desigual da Figura 11-6b e a filogenia inicialmente proposta na Figura 12-6a. O esquema demonstra que para cada sítio informativo será construída uma filogenia com a mesma topologia da árvore proposta em 12-6a (ver adiante).

Tomando, por exemplo, o sítio 28, identificamos a presença de três ancestrais não amostrados que, no entanto, para o cálculo dos custos, terão que ter seus caracteres inferidos. Segundo o algoritmo de Sankoff, os cálculos devem iniciar tomando os clados mais derivados (isto é, mais recentes). Em 12-6c, a posição "Y" da filogenia necessariamente foi ocupada por um dos quatro nucleotídeos. Em cada uma das proposições (A, C, G ou T), o custo associado à substituição é consultado na matriz. No primeiro caso, a hipótese para ocupação da posição "Y" é A. O custo da substituição em cada um dos ramos deve ser verificado e somado. Por exemplo, a substituição de A por T possui custo 4. Como a mesma substituição ocorreu em dois ramos diferentes, somamos o custo total, que totaliza 8. O mesmo procedimento será repetido considerando os outros três nucleotídeos na posição "Y".

Após o cálculo dos custos para as posições "Y" e "Z", é necessário verificar os custos de substituição de "X" para "Y" e "X" para "Z". A figura 12-6d apresenta a primeira hipótese para ocupação da posição "X": o nucleotídeo A. Aqui, o algoritmo somará os custos de substituição de todos os ramos, novamente considerando cada um dos quatro nucleotídeos na posição "X", mas também considerando a variação nas posições "Y" e "Z". A Figura 12-6e identifica a filogenia com o menor custo para o sítio 28. Note que o caractere mais ancestral pode ser tanto o nucleotídeo T quanto C. Os mesmos cálculos serão realizados para todos os sítios do alinhamento, tomando a topologia dada em 12-6a e, ao final, os menores custos para cada sítio serão somados para encontrar o tamanho dos ramos da árvore. A árvore que possuir os ramos mais parcimoniosos será tomada como a árvore de máxima parcimônia.

Computacionalmente, o cálculo dos tamanhos de ramos mais parcimoniosos não é um problema. O desafio da maioria dos métodos de reconstrução filogenética está na inferência da topologia. Assim como no método de máxima verossimilhança, discutido a seguir, o método de máxima parcimônia contará com algoritmos heurísticos para arranjo das topologias. A filogenia é então

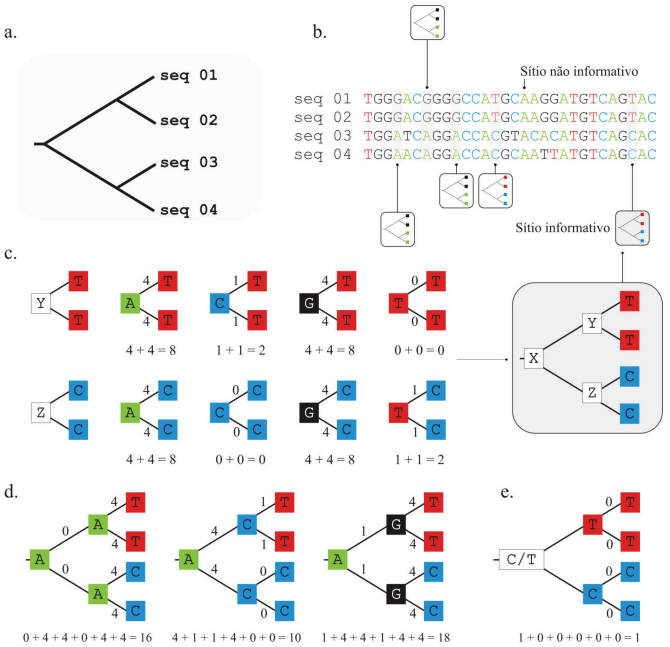


Figura 12-6: Determinação dos custos de substituição pelo método de parcimônia para um sítio do alinhamento de nucleotídeos. (a) Topologia da filogenia proposta para quatro táxons (ver adiante). (b) Alinhamento de nucleotídeos de quatro sequências homólogas. Destacados em cinza estão os sítios informativos para o método de parcimônia. Os demais sítios são considerados não informativos e serão descartados durante os cálculos. (c) Cálculo dos custos para os dois clados presentes na filogenia proposta em "a". O método supõe que a posição "Y" possa ser ocupada por qualquer um dos quatro nucleotídeos. (d) Exemplo do procedimento adotado pelo método, supondo que a posição "X" na filogenia foi ocupada pelo nucleotídeo A. É necessário considerar todas as possibilidades de caracteres nos sítios ancestrais e calcular os respectivos custos. (e) Arranjo de menor custo para a posição 28 do alinhamento de nucleotídeos.

proposta pelo algoritmo, e o critério de parcimônia avalia a árvore. A partir de perturbações realizadas nesta topologia, uma nova topologia é proposta e novamente o critério qualifica a filogenia.

Apesar de velozes, os métodos de parcimônia falham ao estimar a relação evolutiva entre um grande número de táxons, especialmente se diferentes linhagens possuem taxas evolutivas variáveis ou taxas evolutivas muito rápidas. Nestes casos, é comum que o método agrupe incorretamente os táxons com maiores taxas de evolução, levando à inferência da filogenia errada (atração de ramos longos).

Ainda, por não ter um modelo de substituição especificado, o método de parcimônia é incapaz de considerar mutações reversas ou múltiplas substituições. Métodos que geram diferentes hipóteses a partir do alinhamento, considerando as observações biológicas na seleção do modo de substituição dos nucleotídeos e, assim, lidam com eventos aleatórios de probabilidade, substituíram o

uso da máxima parcimônia e, atualmente, são os principais métodos utilizados para a inferência de filogenias.

Máxima Verossimilhança

Idealmente, os métodos de inferência filogenética devem resgatar o máximo de informações contidas em um dado conjunto de sequências homólogas, buscando desvendar a verdadeira história evolutiva dos organismos.

Quando um grande número de mudanças evolutivas em diferentes linhagens é demasiadamente desigual, o método de máxima parcimônia tende a inferir filogenias inconsistentes, proporcionalmente convergindo à árvore errada quanto maior o número de sequências no alinhamento. Assim, abre-se espaço para uma técnica de inferência filogenética mais robusta, que alie as informações do alinhamento a um modelo estatístico capaz de lidar com a probabilidade de mudança de um nucleotídeo para outro de maneira mais completa.

Dentro do campo da filogenética computacional, o método de ***máxima verossimilhança*** primeiramente ocupou este espaço e, desde então, tem sido amplamente utilizado devido à qualidade da abordagem estatística empregada.

A implementação de uma concepção estatística para a máxima verossimilhança, originalmente desenvolvida para estimar parâmetros desconhecidos em modelos probabilísticos, se deu entre 1912 e 1922 através dos trabalhos de A. R. Fisher.

Apesar de utilizado para dados moleculares na década de 70, o método de máxima verossimilhança só se tornou popular na área da filogenética a partir de 1981, com o desenvolvimento de um algoritmo para estimar filogenias baseadas no alinhamento de nucleotídeos. Atualmente, diversos programas implementam este método para realizar a inferência filogenética, incluindo PAUP, MEGA, PHYLIP, fastDNAml, IQPNNI e METAPIGA, dentre outros (Tabela 1-6).

O objetivo principal do método da máxima verossimilhança é inferir a história evolutiva mais consistente com relação aos dados fornecidos pelo conjunto de sequências. Neste modelo, a hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) é avaliada pela capacidade de predizer os dados observados (alinhamento de sequências homólogas). Sendo assim, a verossimilhança de uma árvore é proporcional à probabilidade de explicar os dados do alinhamento. Aquela árvore que com maior probabilidade, entre as outras árvores possíveis, produz o conjunto de sequências do alinhamento, é a árvore que reflete a história evolutiva mais próxima da realidade, mais verossímil e, por isso, de máxima verossimilhança.

É importante ressaltar que diferentes filogenias podem explicar um determinado conjunto de sequências, algumas com maior probabilidade e, outras, com menor probabilidade. No entanto, a soma das verossimilhanças de todas as árvores possíveis para um determinado conjunto de sequências nunca resultará em 1, pois não estamos lidando com as probabilidades de que estas filogenias estejam corretas, mas avaliando a probabilidade de explicarem o alinhamento que foi fornecido.

Se, por exemplo, aplicássemos o método de máxima verossimilhança para inferir a árvore filogenética de um grupo de sequências homólogas que incluem porções recombinantes, encontraríamos uma árvore filogenética com um determinado valor de verossimilhança. A utilização do método, por si só, garantiria como resultado a inferência de uma filogenia. No entanto, sabemos que esta árvore, apesar de ser a mais plausível para explicar o alinhamento dado, não tem qualquer relação com a realidade evolutiva do organismo, já que eventos de recombinação aconteceram no decorrer do tempo e impedem a explicação sob a forma dicotômica de uma filogenia.

A aplicação do método de máxima verossimilhança exige a construção de uma filogenia inicial, geralmente obtida por

métodos quantitativos. Como exemplo, considere a árvore filogenética proposta inicialmente e o respectivo alinhamento de nucleotídeos da Figura 13-6. Para calcularmos a verossimilhança desta filogenia será necessário utilizar um modelo evolutivo, que será importante para atribuir valores e parâmetros às substituições e ajudará no cálculo da probabilidade de que uma sequência X mude para uma sequência Y ao longo de um segmento da árvore.

Dado um determinado modelo evolutivo (JC69, K2P, F81, HKY ou GTR, por exemplo), e assumindo que cada sítio do alinhamento evolui de maneira independente dos demais, podemos calcular o valor de verossimilhança para cada um destes sítios e, posteriormente, multiplicar os valores de cada sítio para encontrar a verossimilhança da árvore dada (Figura 13-6 e a Figura 14-6). Sítios que apresentam deleções serão eliminados da análise.

Como os nós internos destas árvores, geradas a partir de cada sítio do alinhamento, são a representação de OTUs não amostrados (isto é, ancestrais) e, por

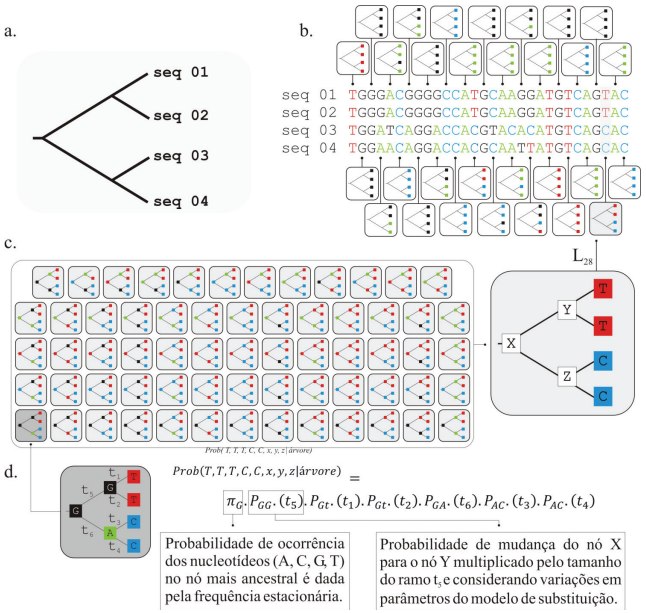


Figura 13-6: Esquema do cálculo da verossimilhança para uma filogenia e seu respectivo alinhamento de nucleotídeos. (a) Árvore filogenética proposta inicialmente para o alinhamento em "b". (b) Para cada posição do alinhamento é destacada a organização dos quatro sítios do alinhamento na árvore proposta em "a". Como exemplo, apenas o sítio do alinhamento destacado em cinza será considerado para o cálculo da verossimilhança. Os quadrados pretos, azuis, verdes e vermelhos nos ramos terminais das filogenias representam, respectivamente, os nucleotídeos guanina, citosina, adenina e timina. (c) Probabilidade de cada uma das 64 possíveis combinações de nucleotídeos nos nós internos da árvore, já que estes representam os sítios de táxons ancestrais não amostrados (P_{XY}, P_{YT}, P_{XZ}, P_{ZC}). (d) O esquema para o cálculo da máxima verossimilhança leva em conta a multiplicação do tamanho dos ramos (t_1, t_2, t_3, t_4, t_5 e t_6) pelas respectivas probabilidades de transição (P_{GG}, P_{GT}, P_{GA} e P_{AC}), além da frequência estacionária dos quatro nucleotídeos no nó mais ancestral (π_X).

conseguinte, não se conhece suas sequências de nucleotídeos, será necessário considerar a ocorrência de todos os nucleotídeos (A, T, C e G) nestas posições da árvore (Figura 13-6c).

Por certo, alguns cenários são mais prováveis que outros; no entanto, todos devem ser considerados durante os cálculos de verossimilhança, pois apresentam alguma probabilidade de terem gerado as sequências dadas no alinhamento.

Adicionalmente, além de calcular a probabilidade de todas as mudanças possíveis para cada um dos sítios do alinhamento (Figura 13-6c), a expressão matemática da verossimilhança ainda incluirá o tamanho dos ramos, dentre outros elementos do modelo de substituição, como um fator determinante para o cálculo (Figura 13-6d).

A probabilidade de ocorrência de cada um dos quatro nucleotídeos no nó mais interno da árvore será igual à respectiva frequência estacionária dada pelo modelo de substituição, já que este parâmetro especifica a proporção esperada de cada um dos quatro nucleotídeos. No modelo de Jukes e Cantor, por exemplo, assume-se que os quatro nucleotídeos ocorrem em proporções iguais de 25%.

Conforme o exemplo da Figura 13-6d, a equação utilizada para calcular a verossimilhança da filogenia proposta no sítio 28, inicialmente, leva em consideração a frequência estacionária do nucleotídeo G, já que este é o nucleotídeo que está sendo considerado como presente no nó mais ancestral da árvore. A probabilidade de este G ser substituído por um A (P_{GA}), ou permanecer G (P_{GG}) será dada pelo modelo de substituição escolhido. Da mesma forma, serão os casos P_{GT}, P_{AC} (repetido duas vezes cada pelo fato de existirem dois ramos terminais com o mesmo nucleotídeo).

O tamanho dos ramos entre dois nós será multiplicado pelas probabilidades de substituição dos nucleotídeos, levando em conta variações em parâmetros do modelo de substituição. Apesar da dificuldade de cálculo computacional, os algoritmos aplicados à inferência filogenética (baseados no princípio de Pulley) automaticamente estimarão o tamanho de cada ramo de modo que este maximize o valor da verossimilhança da árvore filogenética em construção. Nestes casos, o algoritmo atribui diversos valores de distância para um ramo e, a cada valor, verifica a verossimilhança da árvore, buscando aqueles valores que resultam na filogenia com a maior verossimilhança.

A probabilidade de observar os dados em um sítio particular é a soma das probabilidades de todos os possíveis nucleotídeos que poderiam ser observados nos nós internos da árvore (Figura 13-6c). O número de nós internos rapidamente se torna muito grande com o aumento do número de OTUs. Felizmente, através de um algoritmo criado por Felseinstein (algoritmo de "poda"), que se aproveita da própria topologia da filogenia, esses cálculos podem ser realizados de uma maneira computacionalmente eficiente.

Neste processo, propõe-se que os cálculos da verossimilhança de uma determinada árvore sejam feitos a partir de sub-árvores dos ramos terminais em direção aos nós internos, semelhante ao algoritmo usado para o cálculo da parcimônia. No entanto, quando aplicado este método à inferência por máxima verossimilhança é necessário garantir que os modelos de substituição, não presentes no método de máxima parcimônia, sejam reversíveis, ou seja, que a probabilidade de mudança de A para T (P_{AT}) seja a mesma que T para A (P_{TA}). A introdução deste método permitiu que as análises de verossimilhança pudessem ser aplicadas a grandes conjuntos de sequências, de forma mais rápida e efetiva.

Ao final, multiplicamos os valores de verossimilhança de todos os sítios e encontramos o valor de verossimilhança da árvore (Figura 14-6):

A expressão matemática acima indica que a verossimilhança (L) é igual à multiplicação (Π) das probabilidades de cada sítio i (D^i, calculado conforme Figura 13-6), dada a árvore filogenética (topologia, modelo evolutivo e tamanho dos ramos). Aquela árvore que tiver o maior valor de verossimilhança entre todas as árvores possíveis para um determinado alinhamento de sequências será a árvore que melhor explica o alinhamento e, por isso, a árvore de máxima verossimilhança. Por fim, é importante ressaltar que, apesar de estarmos avaliando nucleotídeos neste exemplo, o mesmo raciocínio poderia ser aplicado para a inferência filogenética para um alinhamento de aminoácidos.

Até o momento vimos, em linhas gerais, como realizar o cálculo de verossimilhança para uma dada filogenia (Figura 13-6). No entanto, outra função importante dos métodos computacionais de inferência

filogenética é apontar a topologia e encontrar a árvore de máxima verossimilhança entre todas as árvores possíveis para o conjunto de dados. Infelizmente, não existem algoritmos que garantam a localização da árvore real devido ao grande espaço amostral de árvores possíveis (Figura 9-6).

Após uma árvore ser construída, é necessário calcular sua verossimilhança e comparar este valor com todas as árvores já construídas. Como é impossível testar a verossimilhança para todas as filogenias possíveis, os algoritmos de máxima verossimilhança incluirão buscas heurísticas

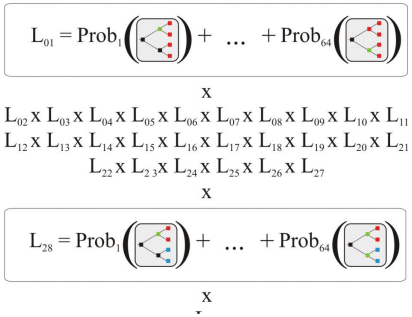


Figura 14-6: Cálculo da máxima verossimilhança de uma dada filogenia, considerando seu respectivo alinhamento de nucleotídeos contendo quatro táxons e 30 bases (Figura 13-6b). Para cada sítio (L_{01}, L_{02}, ..., L_{30}) será calculado um valor de probabilidade que envolve a consideração de todos os quatro nucleotídeos em cada um dos ramos ancestrais da filogenia. Posteriormente, os valores de verossimilhança de cada sítio serão multiplicados para encontrar a verossimilhança total da filogenia.

para solucionar este problema (estes métodos construirão diferentes filogenias a partir do mesmo conjunto de dados do alinhamento).

Na problemática das filogenias, diferentes programas têm proposto as mais diversas alternativas para avaliar o maior número de árvores do espaço amostral total e encontrar aquela com o maior valor de verossimilhança. No entanto, como regra geral, a maioria dos programas de máxima verossimilhança segue alguns passos comuns:

1. Uma filogenia preliminar com determinada topologia é construída (geralmente são utilizadas árvores construídas pelo método de aproximação de vizinhos);
2. Os parâmetros para esta árvore são modificados buscando maximizar a verossimilhança (em alguns casos, a filogenia vai sendo construída pela adição de novos táxons aleatoriamente). Para a modificação da filogenia, os algoritmos podem implementar técnicas de rearranjos de ramos, conforme descrito em 6.4;
3. O valor de máxima verossimilhança para esta árvore é armazenado;
4. Outras topologias são construídas e seus parâmetros também são avaliados;
5. Finalmente, a filogenia que possuir o valor de máxima verossimilhança será a melhor estimativa evolutiva para o dado conjunto de sequências.

Embora estes processos simplifiquem os verdadeiros fenômenos biológicos que governam a evolução de uma sequência, apresentando assim dificuldades em identificar a árvore com o maior valor de verossimilhança, eles são normalmente robustos o bastante para estimar as relações evolutivas entre táxons.

Como estes métodos implicam em encontrar a árvore com o valor máximo de verossimilhança entre todas as árvores amostradas, o resultado final sempre fornecerá apenas uma filogenia, ao contrário dos métodos Bayesianos que serão vistos a seguir. Cabe ressaltar que, devido ao uso de diferentes algoritmos, na prática, um mesmo conjunto de sequências submetido a diferentes programas para inferência filogenética por máxima verossimilhança

dificilmente resultará na mesma árvore. Por isso, é necessário ser cauteloso ao interpretar árvores geradas pelo método de máxima verossimilhança.

Análises Bayesianas

A estatística Bayesiana nasceu com a publicação de um ensaio matemático do reverendo Thomas Bayes, em 1793. Nesta publicação, o reverendo apresenta o desenvolvimento de um método formal para incorporar evidências prévias no cálculo da probabilidade de acontecimento de determinados eventos.

Inicialmente, este método foi aplicado apenas no campo da matemática e, só a partir de 1973, passa a ser incorporado no pensamento biológico e na inferência filogenética. Com o advento de diversos programas de acesso livre para realizar a inferência de filogenias por estatística Bayesiana, o método se difundiu e, atualmente, tornou-se um campo de estudo específico dentro da filogenética computacional.

A ***inferência Bayesiana*** engloba o método de máxima verossimilhança (Tabela 2-6) mas, adicionalmente, inclui o uso de informações dadas *a priori*. Estas informações refletem características a respeito da filogenia, do alinhamento ou dos táxons, que o pesquisador sabe de antemão.

Entre os principais parâmetros que podem ser conhecidos antes da reconstrução filogenética pode-se destacar a taxa evolutiva, tipo de relógio molecular, parâmetros do modelo de substituição, datas de coleta das amostras, datas para calibração da filogenia (achados fósseis, datação por carbono-14, aproximações arqueológicas, etc.), distribuição geográfica, organização monofilética de um grupo de indivíduos ou, até mesmo, parâmetros de dinâmica populacional.

Os valores atribuídos *a priori* são incorporados à estatística Bayesiana na forma de probabilidades e comporão o termo chamado de ***probabilidade anterior*** (*prior probability*). Se sabemos de antemão que um determinado grupo de organismos é ancestral em relação a outro, podemos atribuir uma maior probabilidade àquelas filogenias que relacionam estes organismos da maneira como sabemos *a priori*.

Qualquer informação útil, que é fornecida pelo pesquisador antes da própria reconstrução da filogenia, poderá ser convertida em uma probabilidade anterior para ser inserida nas análises de inferência Bayesiana. No entanto, as informações cedidas *a priori* devem ser distribuições de números prováveis (mínimo e máximo), e não números exatos. Quando estes valores não são conhecidos ou quando, por exemplo, não se quer atribuir maior probabilidade a uma determinada topologia, o parâmetro terá uma distribuição uniforme de probabilidades.

Na maioria dos aplicativos que lidam com inferência Bayesiana existem distribuições uniformes associadas às

Tabela 2-6: Comparação entre os métodos de máxima verossimilhança e inferência Bayesiana.

Método	Vantagens	Desvantagens
Máxima Verossimilhança	Captura totalmente a informação dos sítios do alinhamento para construção das filogenias	Comparativamente ao método Bayesiano, o algoritmo para reconstrução por máxima verossimilhança é mais lento
Estatística Bayesiana	Tem grande ligação com a máxima verossimilhança, sendo, no entanto, geralmente mais rápida. Modelos populacionais podem ser incluídos para inferência das filogenias	Os parâmetros para as probabilidades anteriores devem ser especificados e pode ser difícil especificar quando as análises são satisfatórias

probabilidades anteriores que assumem que todos os valores possíveis são dados pela mesma probabilidade.

Além das probabilidades anteriores, a inferência Bayesiana é baseada nas ***probabilidades posteriores*** de um parâmetro como, por exemplo, a topologia. Através da probabilidade posterior é possível verificar a probabilidade de cada uma das hipóteses (árvores filogenéticas). Sendo assim, ao final das análises, é possível estabelecer uma estimativa da probabilidade dos eventos retratados por uma determinada filogenia, ou seja, a probabilidade de cada filogenia. As probabilidades posteriores são calculadas utilizando a fórmula de Bayes:

O termo $L(H \mid D)$ é chamado de distribuição de probabilidades posteriores, e é dado pela probabilidade da hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) a partir dos dados disponíveis (alinhamento de sequências). O termo $L(D \mid H)$ descreve o cálculo de máxima verossimilhança, enquanto o multiplicador $L(H)$ é a probabilidade anterior. Para o termo que envolve a função de máxima verossimilhança, é ainda necessário considerar também todos os tópicos já discutidos na seção anterior. O denominador $L(D)$ é uma integração sobre todas as possibilidades de topologias, tamanhos de ramo e valores para os parâmetros do modelo evolutivo, o que garante que a soma da probabilidade posterior para todos eles seja 1. O denominador atuará como um normalizador para o numerador. Reescrevendo, temos:

onde o termo filogenia descreve a topologia da árvore, o modelo de substituição e o comprimento dos ramos. Assim, através da multiplicação das probabilidades anteriores pela verossimilhança, divididos pelo fator de normalização, o método busca a hipótese (topologia da árvore, o modelo de substituição e o comprimento dos ramos) em que a probabilidade posterior é máxima.

O objetivo da inferência Bayesiana é calcular a probabilidade posterior para cada filogenia proposta. No entanto, para cada árvore diversos parâmetros devem ser especificados pelo usuário, incluindo topologia, tamanho dos ramos, parâmetros do modelo de substituição, parâmetros populacionais, relógio molecular, taxa evolutiva e etc. Dada uma filogenia, todos os parâmetros terão sua probabilidade posterior calculada. Se dadas 1000 filogenias, teremos 1000 valores de probabilidade posterior para cada parâmetro.

Devido à impossibilidade de construção de todas as filogenias possíveis para a maioria dos alinhamentos, a análise Bayesiana se aproveita de técnicas de amostragem para estimar os valores esperados de cada parâmetro.

Neste sentido, os métodos de inferência Bayesiana utilizam as Cadeias de Markov Monte Carlo (MCMC, *Monte Carlo Markov Chain*) para aproximar as distribuições probabilísticas em uma grande variedade de contextos. Esta abordagem permite realizar amostragens a partir do conjunto total de filogenias, relacionando cada filogenia a um valor probabilístico. Sem a aplicação de um método que obtenha amostras do espaço de possíveis filogenias, como o modelo de MCMC, a estimativa de todos os parâmetros se tornaria analiticamente impossível nos atuais computadores.

Um dos métodos de MCMC mais usados na inferência filogenética é uma modificação do algoritmo Metropolis, chamado de Metropolis-Hastings. A ideia central deste método é causar pequenas mudanças em uma filogenia (topologia, tamanho dos braços, parâmetros do modelo de substituição, etc.) e, após a modificação, aceitar ou rejeitar a nova hipótese de acordo com o cálculo de razão das probabilidades. Este método garante que diversas árvores sejam amostradas do espaço total de filogenias, amostrando filogenias com probabilidade posterior mais alta (Figura 15-6):

1. Inicialmente, o algoritmo MCMC gera uma filogenia aleatória X, arbitrariamente escolhendo o tamanho dos ramos para dar início à cadeia;
2. O valor de probabilidade associado a esta filogenia é calculado (probabilidade posterior calculada através da fórmula de Bayes);
3. Perturbações aleatórias são realizadas nesta filogenia inicial X (mudanças na

topologia, no tamanho dos ramos, nos parâmetros do modelo de substituição, etc.) e geram uma filogenia Y;

4. A probabilidade posterior é calculada para a filogenia Y;
5. A filogenia Y é tomada ou rejeitada para o próximo passo baseado na razão R (probabilidade posterior de Y dividida pela probabilidade posterior de X). Se R é maior que 1, a filogenia Y é tomada como base para o próximo passo. Se R é menor que 1, um número entre 0 e 1 é tomado aleatoriamente. Se R é maior que o número aleatório gerado, a filogenia será tomada, no entanto se for menor, a filogenia Y é rejeitada;
6. Se a nova proposta Y for rejeitada, retorna-se ao estado X e novas modificações serão realizadas nesta filogenia;
7. Supondo que a proposta Y tenha sido aceita, ela sofrerá uma nova perturbação a fim de gerar uma nova filogenia;
8. Todas as árvores amostradas são armazenadas para posterior comparação. Os pontos visitados formam uma espécie de cadeia ao longo do espaço amostral total de filogenias.

O principal objetivo da cadeia é amostrar filogenias com probabilidades crescentes. No entanto, é importante que o algoritmo utilizado para tal permita que algumas árvores com menor probabilidade sejam amostradas para evitar que a cadeia fique "presa" em picos de máximo local (Figura 9-6).

Sendo assim, o cálculo da razão R considerando um valor aleatório entre 0 e 1 garantirá que, em determinados momentos, uma filogenia com menor probabilidade seja aceita. Por este método, é possível amostrar filogenias da região de um vale passando, por exemplo, de um pico de ótimo local para o pico de ótimo global (Figura 9-6).

A proposta de novas árvores na cadeia de Markov é uma etapa crucial para uma boa amostragem de filogenias. Na abordagem Bayesiana, uma boa amostragem inclui um grande número de filogenias, suficientemente diferentes entre si. Se filogenias muito diferentes são propostas, serão rejeitadas com muita frequência, pois é provável que tenham menor probabilidade posterior. Pelo contrário, se filogenias muito similares forem geradas, o espaço amostral não será varrido adequadamente e a cadeia deverá "correr" por muitos passos (amostrar um maior número de filogenias), aumentando o

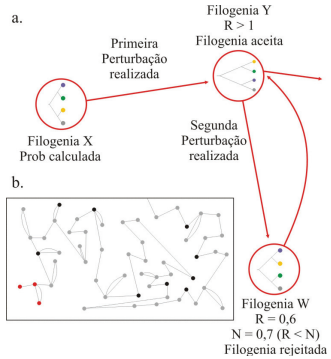


Figura 15-6: Esquema de amostragens MCMC aplicada à inferência filogenética pelo método Bayesiano utilizando o algoritmo de Metropolis-Hastings. (a) Após a proposição de uma filogenia inicial X, perturbações aleatórias são realizadas para gerar a filogenia Y. Devido à razão R>1, a nova filogenia é aceita. Nova perturbação é realizada para gerar a filogenia W e, devido a razão de probabilidades R resultar em um número menor que 1, um número aleatório N é sorteado. Sendo R<N, a nova proposição é rejeitada e a cadeia retorna à filogenia Y. (b) Andamento da cadeia na amostragem de filogenias. Cada círculo destaca uma nova filogenia que é proposta após a perturbação. As linhas conectando os círculos evidenciam a direção do andamento da cadeia. Apesar de a cadeia percorrer muitos passos, apenas alguns serão registrados para análise final (círculos pretos). Os círculos em vermelho são aqueles evidenciados em (a).

tamanho da cadeia e o tempo computacional.

Estimar o quanto a cadeia deve percorrer para amostrar um número suficiente de filogenias para as sequências dadas (espaço de árvores) é um fator fundamental para obter bons resultados em uma análise Bayesiana. Na maioria dos programas que utilizam estatística Bayesiana para inferir filogenias, o usuário deve especificar o tamanho da cadeia. Esse número é de grande subjetividade, e depende diretamente da distribuição das probabilidades anteriores, do número de táxons incluídos na filogenia e da relação evolutiva entre eles.

A Figura 16-6 exemplifica o andamento da amostragem da MCMC em um espaço de filogenias. Supondo que os quadrados em *a*, *b* e *c* representam um espaço amostral de filogenias, semelhante ao apresentado na Figura 15-6b, e que os pontos pretos sejam as filogenias que vão sendo amostradas com o desenvolvimento da MCMC vemos que, ao final do processo, depois de empregados 100 mil passos (Figura 16-6c), um grande número de filogenias foi amostrado.

Ainda, na região delimitada por um círculo, assumimos que estão as filogenias com maior probabilidade de explicar a história evolutiva de um grupo de organismos, ou seja, as filogenias reais. Note que quanto maior o número de passos percorridos pela cadeia, maior a amostragem do espaço de filogenias e maior o número de amostras dentro da região com filogenias de alta probabilidade.

Ao final, após o término da cadeia, a distribuição das probabilidades posteriores de todos os parâmetros deve ser verificada. No entanto, as amostras tomadas no início da cadeia são tipicamente descartadas, pois estão sob forte influência do local de início da cadeia. As filogenias do início da cadeia estão muito longe de pontos máximos no espaço amostral e, por isso, é provável que todas as novas filogenias sugeridas subsequentemente sejam tomadas para o próximo passo (qualquer árvore proposta será mais provável que as árvores iniciais semelhantes àquela gerada aleatoriamente).

Esta fase inicial é conhecida como período de *burn in* (Figura 17-6). Conforme a cadeia avança, espera-se que a probabilidade das árvores amostradas aumente e, quando um número suficiente de filogenias for amostrado, chegue a uma distribuição

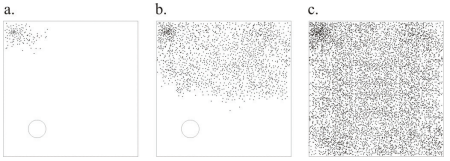


Figura 16-6: Espaço de possíveis árvores analisadas pela MCMC. Considerando que os quadrados descrevem o espaço amostral de todas as filogenias possíveis para um dado conjunto de sequencias, os pontos pretos representam as filogenias que foram amostradas ao longo da cadeia. Os círculos presentes no canto esquerdo inferior representam a região de máximo global (isto é, maior probabilidade) neste espaço amostral. O andamento da cadeia neste exemplo é o mesmo apresentado na Figura 15-6b (a) cento e trinta passos percorridos pela cadeia; (b) trinta mil passos percorridos pela cadeia; (c) cem mil passos percorridos pela cadeia. Nota-se que quanto maior o número de passos percorridos, maior a amostragem de filogenias no espaço. Da mesma forma, aumenta a probabilidade de a cadeia amostrar àquelas filogenias de máximo global.

estacionária. Em termos Bayesianos, espera-se que a cadeia atinja a convergência.

Um dos primeiros indicativos de que a cadeia convergiu para a distribuição correta está na estabilidade dos valores de probabilidade dos parâmetros da cadeia (cada parâmetro da filogenia poderá ter uma distribuição independente). Portanto, a representação gráfica dos valores das probabilidades e dos respectivos passos da cadeia (*trace plot*) é uma importante ferramenta para monitorar o desempenho da

MCMC (Figura 17-6).

Devido ao aumento brusco de probabilidade das filogenias que são visitadas pelo andamento da cadeia, os gráficos necessariamente incluirão os valores medidos em escala logarítmica (Ln L, Figura 17-6). Em estatística Bayesiana, é comum que seja atribuído um intervalo de credibilidade de 95% para os parâmetros amostrados. Estes valores são obtidos através da eliminação de 2,5% dos valores mais baixos e de 2,5% dos valores mais altos para um determinado parâmetro. Um intervalo de credibilidade contém o valor correto com 95% de probabilidade; no entanto, não se trata de um intervalo de confiança.

Adicionalmente, outros métodos são úteis para diagnosticar a convergência da cadeia, tais como o exame do tamanho amostral efetivo (ESS) e a comparação de amostras resultantes de diferentes cadeias (várias cadeias de MCMC são aplicadas para o mesmo conjunto de dados). Apesar de ser computacionalmente intensiva, a última alternativa parece ser a mais confiável para verificar a convergência. Contudo, o exame de ESS é, ainda hoje, o método mais utilizado. O tamanho amostral efetivo é uma estimativa para verificar o número de amostras independentes existentes na cadeia, ou seja, quantas amostras não similares foram tomadas. Atualmente, um ESS maior que 200 é um indicativo de que a cadeia convergiu adequadamente.

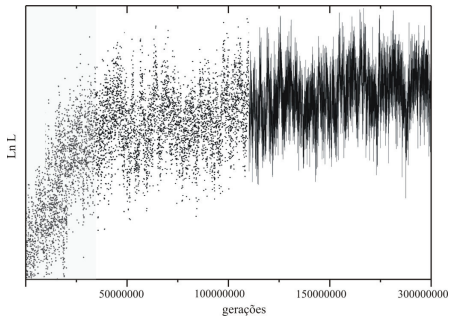


Figura 17-6: Representação gráfica das probabilidades das filogenias na cadeia ao longo de 300 milhões de amostragens. O esquema demonstra duas visualizações possíveis: à esquerda, são mostrados apenas os pontos referentes às amostras tomadas ao longo da cadeia e, à direita, as amostragens sucessivas são ligadas umas as outras para facilitar a visualização do comportamento da cadeia. Em cinza, a fase inicial de *burn in* da Cadeia de Markov Monte Carlo.

A técnica de *Metropolis Coupling*, conhecida como MCMCMC ou $(MC)^3$, através da introdução da corrida simultânea de duas cadeias, pode ajudar na amostragem de máximos globais e beneficiar na convergência da cadeia. Nesta técnica uma cadeia, chamada de quente (*hot chain*), permite aproximar os valores de máxima e mínima probabilidade das amostras para que a cadeia possa, de forma mais rápida, "saltar" entre picos de probabilidade, especialmente de máximos locais para máximos globais. O aquecimento da cadeia é dado pelo parâmetro β e visa diminuir a altura dos picos locais no espaço amostral. Uma segunda cadeia simultânea, chamada de fria (*cold chain*), utiliza as informações destes saltos da cadeia quente para melhorar a sua amostragem e garantir a convergência.

Os métodos Bayesianos de inferência filogenética ainda têm a vantagem de aplicar modelos que envolvem diferentes tipos de relógios moleculares.

As distâncias genéticas, depois de "tratadas" pelos modelos de substituição, não tem qualquer significado sozinhas quando se deseja estimar, por exemplo, a idade do ancestral comum mais recente de duas OTUs. Esta e outras questões podem ser avaliadas quando aplicamos uma medida de tempo nas inferências, a fim de calibrar as taxas evolutivas.

Sequenciamentos de amostras isoladas em diferentes épocas podem fornecer a calibração adequada para inferências temporais, pois se assume uma taxa evolutiva constante ao longo de um tempo t para todos os ramos de uma filogenia (relógio molecular estrito).

As taxas evolutivas dependem de diversos fatores e podem variar, nem sempre seguindo a constância proposta por este modelo. Após a introdução de um tipo específico de relógio molecular relaxado, as taxas de evolução podem variar ao longo da árvore para diferentes grupos e não são correlacionadas, ou seja, grupos evolutivamente próximos não necessariamente terão taxas de evolução semelhantes (relógio molecular relaxado não correlacionado).

Complexos modelos de dinâmica populacional podem ser analisados sob uma perspectiva Bayesiana. Quando o conjunto de sequências submetido às analises são isolados de uma população homogênea, os parâmetros de história demográfica podem ser usados para modelar as mudanças populacionais ao longo do tempo. Desta forma, através da estatística Bayesiana é possível, além da inferência filogenética, refinar as análises e datar filogenias e ramos específicos (Figura 18-6), inferir caracteres ancestrais e analisar a dinâmica populacional sob uma ótica evolutiva.

6.8. Confiabilidade

O papel principal das técnicas de inferência filogenética é desvendar as relações evolutivas reais através de dados moleculares, buscando garantir que esta reconstrução seja fidedigna. Além da inferência das relações evolutivas entre os táxons, é igualmente importante que a filogenia possua precisão. Esta característica está relacionada ao número de filogenias que podem ser excluídas, a partir do conjunto total de filogenias, por não serem "verdadeiras". Quanto maior o número de filogenias excluídas neste processo, mais preciso é o método.

Em geral, na maioria dos casos de reconstrução filogenética, a falta de precisão das filogenias está relacionada ao conjunto de dados que está sendo fornecido no alinhamento. O gene considerado, o tamanho das sequências, o número de indivíduos e o grupo externo são atribuições fundamentais para uma reconstrução filogenética precisa e dependem, especialmente, do objetivo do estudo e da própria disponibilidade de informação.

Em muitos casos, o pesquisador é ainda dependente do número de amostras e do sucesso de coleta em campo, sobretudo, quando seu objeto de estudo se trata de uma espécie rara ou de indivíduos de difícil amostragem. No entanto, apesar de toda a informação relacionada ao conjunto de dados, a dificuldade de amostragem de indivíduos

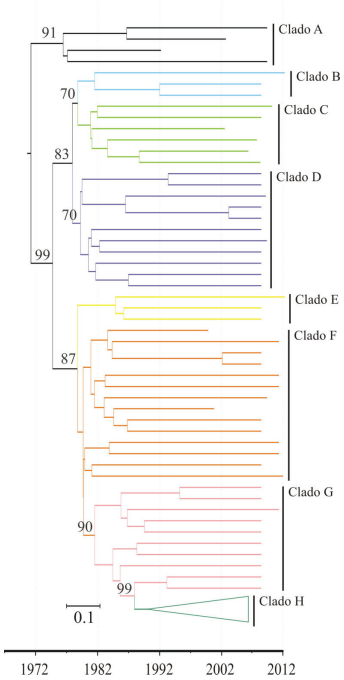


Figura 18-6: Árvore filogenética consenso gerada por inferência Bayesiana para 70 sequências de nucleotídeos. As cores nos ramos representam diferentes clados (B-H). O grupo externo está identificado como clado A. O Clado H foi agrupado para facilitar a representação. Nos nós estão especificados os valores de probabilidade posterior acima de 70. Abaixo, é apresentada a escala temporal inferida a partir da utilização de um relógio molecular relaxado.

parece ser, sem dúvida, o principal problema relacionado a precisão das filogenias, pois a falta de dados de variabilidade genética compromete a inferência de história evolutiva coerente.

Como é possível saber se a amostragem foi suficiente e a filogenia é confiável? Usualmente, a resposta para esta questão consiste na reamostragem de dados. Se novas amostras forem tomadas e a mesma

filogenia for reproduzida, a filogenia proposta tem seu valor reforçado. No entanto, na maioria dos casos, a reamostragem de dados da forma usual (coletas de novos espécimes, reamostragens em campo, achado fóssil diferente, etc) não é factível. Assim, algoritmos que produzem diferentes amostragens utilizando o mesmo conjunto de dados foram desenvolvidos para possibilitar a verificação da confiabilidade nos clados das filogenias. Destaca-se entre estes algoritmos o método de ***bootstrap***.

Bootstrap é um método de reamostragem utilizado para realizar comparações da variabilidade das hipóteses filogenéticas, oferecendo medidas de confiabilidade aos clados propostos. A reamostragem é realizada a partir do mesmo conjunto de dados, e novas amostras fictícias com o mesmo tamanho serão geradas.

Segundo este método, cada sítio do alinhamento será tratado de forma independente. Conforme a Figura 19-6, inicialmente o algoritmo reconstruirá a filogenia a partir do alinhamento dado e, posteriormente, diversas replicatas serão reconstruídas. As colunas, representando os sítios do alinhamento, serão aleatoriamente tomadas (amostradas) pelo algoritmo e, em seguida, serão agrupadas uma ao lado da outra de maneira a formar um novo alinhamento (com o mesmo número de sítios do alinhamento original, Figura 19-6).

Por este método, é possível que um mesmo sítio seja amostrado mais de uma vez e, portanto, alguns sítios não serão selecionados para o novo alinhamento. Um número fornecido pelo usuário especificará o número de pseudoreplicatas (novos alinhamentos) que serão construídas. Assim que uma pseudoreplicata for criada, o algoritmo constrói a filogenia correspondente.

É importante ressaltar que a inferência destas filogenias será realizada pelo método de construção especificado pelo usuário, seja aproximação de vizinhos, máxima parcimônia ou máxima verossimilhança (para árvores bayesianas, veja adiante). Ao final, o algoritmo analisará os clados e automaticamente verificará a presença de determinados agrupamentos em todas as filogenias construídas. Se, por exemplo, encontramos as sequências 1 e 2 formando um clado em 70% das filogenias construídas, atribuiremos a confiabilidade de 70 ao clado formado por estas duas sequências. Comumente, o valor de confiabilidade dos clados é colocado próximo ao ancestral comum do clado (Figura 18-6).

A partir dos resultados de confiabilidade dos clados é possível também construir filogenias baseando-se na árvore consenso gerada pela regra da maioria (*majority-rule consensus tree*). Neste método, o algoritmo tabulará todos os clados formados em todas as replicatas geradas. Aqueles clados que

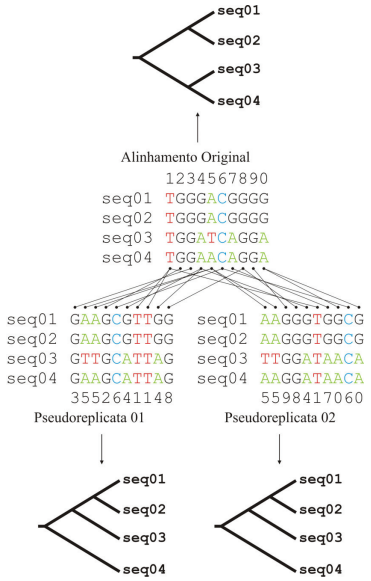


Figura 19-6: Método de *boostrap* para filogenias. A partir do alinhamento original, as colunas que representam os sítios serão aleatoriamente amostradas para construir pseudoreplicatas (um mesmo sítio pode ser sorteado diversas vezes). Estas, por sua vez, serão utilizadas para a inferência de filogenias, da mesma forma que o alinhamento original.

mais aparecerem servirão para montar a filogenia consenso.

Ao contrário dos métodos de aproximação de vizinhos, máxima parcimônia e máxima verossimilhança, a confiabilidade de filogenias construídas através de estatística Bayesiana é inerente ao processo. Como diversas filogenias são amostradas ao longo do desempenho da Cadeia de Markov, não é necessário nenhum método para simular reamostragens do mesmo conjunto de dados. As amostras serão resumidas a partir da distribuição posterior de filogenias como frequência de clados individuais e serão identificadas por um número próximo ao ancestral comum daqueles clados (Figura 18-6). Portanto, o valor de probabilidade posterior de um clado representa uma inferência a respeito da probabilidade daquele clado.

A comparação dos valores de *bootstrap* e de probabilidade posterior dos clados para filogenias construídas a partir do mesmo alinhamento utilizando máxima verossimilhança e o método Bayesiano, respectivamente, leva a conclusão de que o método Bayesiano superestima a confiança aos clados. A confiança atribuída pela probabilidade posterior é geralmente maior que aquela atribuída pelo método de *bootstrap*. Por isso, enquanto uma confiança acima de 70 é considerada sustentada para o *bootstrap*, apenas valores acima de 90 podem ser considerados relevantes para os métodos Bayesianos.

6.9. Interpretação de filogenias

Árvores filogenéticas são diagramas que denotam a história evolutiva de diferentes OTUs a partir de seu ancestral comum. Mais do que isso, as filogenias moleculares são ferramentas que ajudam no entendimento dos diversos processos evolutivos que moldam o genoma dos organismos. Desta forma, a interpretação das implicações evolutivas associadas a um, ou a um conjunto de táxons, está diretamente relacionada à disposição dos ramos internos e externos de uma árvore. Independentemente do método de inferência, ou da forma como a árvore é apresentada, a interpretação dos resultados será baseada nos mesmos pressupostos, ainda que métodos diferentes possam originar filogenias diferentes.

Inicialmente, é necessário observar a presença de uma raiz. Como já discutido, o método de enraizamento pelo grupo externo é o mais comum e utiliza organismos sabidamente relacionados ao grupo em evidência, servindo para orientar o algoritmo em relação às características mais ancestrais do grupo. O grupo externo ajudará a evidenciar o tempo evolutivo. Na Figura 20-6, por exemplo, o grupo externo é dado pelo orangotango, pois este compartilha o mesmo ancestral comum que o restante do grupo. No caso de filogenias sem raiz, é necessário ter cautela nas interpretações, pois este tipo de diagrama apenas revela a relação entre os táxons.

Depois de encontrada a raiz da filogenia, é preciso avaliar os ramos. Dependendo do método, os ramos podem ter significados diferentes. Na Figura 18-6, os ramos evidenciam o tempo real, apresentando OTUs amostradas no passado. Pelo contrário, na Figura 20-6, os ramos evidenciam apenas um tempo evolutivo representado pelo número de modificações genômicas, desde o organismo ancestral até os ramos terminais. Além disso, deve-se perceber a escala na qual os ramos foram representados, pois estes indicam o número de substituições que provavelmente ocorreram ao longo do processo evolutivo e podem ajudar na interpretação das taxas evolutivas.

Conclusões evolutivas baseadas em árvores filogenéticas devem ser sustentadas em árvore confiáveis e, por isso, a medida de confiabilidade dos ramos deve ser denotada. Inicialmente, é necessário verificar o método utilizado para reconstrução da filogenia e, quando necessário, verificar o algoritmo utilizado para gerar a confiabilidade dos clados. Ramos com maiores valores de confiabilidade gerarão conclusões mais confiáveis, enquanto que clados com baixos

valores deverão ser interpretados com maior cuidado. No entanto, não é necessário negar totalmente conclusões baseadas em

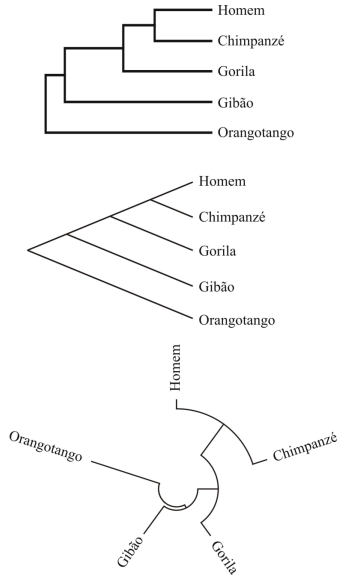


Figura 20-6: Diferentes representações da filogenia dos primatas

filogenias com baixa confiabilidade nos ramos. O tipo de método, a forma de amostragem e o número de OTUs podem ser fatores de interferência e, assim, podem prejudicar a valorização dos ramos.

O padrão de organização dos ramos de uma filogenia denota o padrão de ancestralidade. As filogenias não são escadas, onde alguns organismos são "mais evoluídos" que outros, mas uma representação da história da derivação de OTUs. Na Figura 18-6, por exemplo, é possível observar que os clados B, C, D, E, F e G possuem um ancestral comum que compartilha um outro ancestral com o clado A. Já o clado H, representado por um triângulo para evidenciar um grande número de táxons naquele ponto da filogenia, teve um ancestral comum dentro do clado G. Este padrão sugere que o clado H se originou a partir do clado G. Da mesma forma, podemos observar a disposição do clado G em relação ao F e concluir que o primeiro se originou a partir do segundo.

No caso da Figura 20-6, observamos que humanos e chimpanzés tiveram um mesmo ancestral comum. Com base nestes dados, é incorreto pensarmos que humanos são derivados de chimpanzés, ou que humanos são mais evoluídos que chimpanzés. Estes organismos estão apenas formando um mesmo clado dentro da filogenia dos primatas.

Por último, é fundamental saber o objetivo do estudo filogenético a ser realizado. Árvores filogenéticas devem ser construídas para responder uma determinada questão, que pode envolver apenas um, ou diversos organismos.

Quando possível, é importante reconstruir a filogenia utilizando diferentes métodos de inferência e compará-las entre si. A conclusão desta forma será melhor sustentada. Além disso, atualmente, a história retratada em uma filogenia não é por si só satisfatória. Outras ferramentas podem ser utilizadas para complementar e sustentar a interpretação de uma filogenia, incluindo análises de recombinação, pressão seletiva e estruturação populacional, verificação de coespeciação, construção de redes filogeográficas, comparação com dados de fósseis, eventos geológicos, dados históricos e, até mesmo, análises de dados comportamentais.

Um exemplo da combinação de análises filogenéticas com dados históricos veio na confirmação da origem e disseminação humana a partir da África. Através da utilização de dados histórico-antropológicos (como vestígios materiais de hominídeos ancestrais), fósseis de hominídeos e análises de DNA mitocondrial de representantes de diferentes etnias, os pesquisadores puderam traçar as rotas de disseminação humana a partir da África.

Outro exemplo está na solução de um enigma que perturbou zoólogos por um longo período: a posição taxônomica do panda-gigante entre os mamíferos carnívoros. Apesar de esta espécie ser fisicamente muito similar a um urso, outras características, como dentição e anatomia das patas, levaram à proposição de uma hipótese antes não imaginada.

Tal hipótese propunha que o panda-gigante (*Ailuropoda melanoleuca*) seria proximamente relacionado ao o panda-vermelho (*Ailurus fulgens*), um mamífero de pequeno porte, semelhante ao guaxinim. Com o emprego de diferentes dados, incluindo fósseis, anatomia de mamíferos atuais, distribuição geográfica, sequências de DNA de diferentes porções do genoma, sequências de aminoácidos de diferentes proteínas e mapeamento cromossômico, foi possível estabelecer uma história evolutiva plausível, capaz de descrever a origem evolutiva do panda-gigante (Figura 21-6).

Por meio dessa análise combinada de dados, se propôs que o panda-gigante, um urso, derivou do ancestral comum dos ursos há cerca de 24 milhões de anos atrás, muito antes das derivações que originaram todos os outros ursos existentes hoje. Além disso, observou-se que os ursos e os procionídeos (grupo que inclui o guaxinim e o panda-vermelho) possuem um ancestral comum que deu origem às duas linhagens há aproximadamente 30 milhões de anos.

A filogenia molecular é uma ferramenta útil quando empregada isoladamente, mas que pode se beneficiar de diferentes tipos de dados para propor uma história evolutiva. Em última análise, a decisão sobre que tipos de dados (além dos moleculares) serão empregados na análise filogenética dependerá da pergunta a ser respondida com essa técnica. Não existem regras pré-estabelecidas, e as estratégias analíticas precisam ser propostas caso a caso.

6.10. Conceitos-chave

Ancestral: organismo ou sequência que originou novo(s) organismo(s) ou sequência(s). Em alguns casos pode ser considerado o mesmo que primitivo.

Apomórfico: refere-se a um caractere novo adquirido ao longo do processo evolutivo, uma inovação. Uma apomorfia pode servir de diagnóstico para separação de clados.

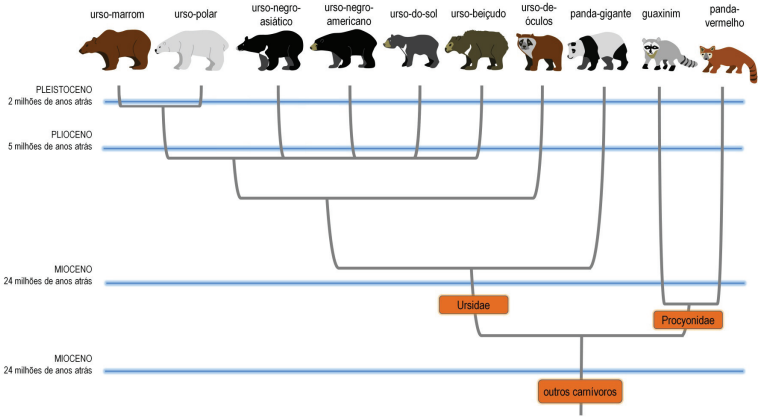


Figura 21-6: Posição filogenética do panda-gigante, baseada na combinação de diferentes tipos de dados. Baseado em BININDA-EMONDS, Olaf R.P. *Phylogenetic position of the giant panda*. Em: LINDBURG, D.G. & Baragona, K. *Giant pandas: Biology and conservation*. Berkeley: University of California Press, 2004; e em O'BRIEN, Stephen J. e colaboradores: *A molecular solution to the riddle of the giant panda's phylogeny*. *Nature*, 317, 140-144, 1985.

Autapomorfias: apomorfias específicas e restritas a um clado.

Bootstrap: método de reamostragem que permite verificar a confiabilidade dos ramos de uma filogenia.

Cadeias de Markov Monte Carlo: método utilizado pela estatística Bayesiana para amostrar as probabilidades de distribuição de diferentes parâmetros das filogenias.

Clado: grupo formado por um ancestral e todos seus descendentes, um ramo único em uma árvore filogenética.

Derivado: que se originou de um ancestral e é mais recente no tempo evolutivo (nota: deve-se evitar o termo "mais evoluído" e, em seu lugar, empregar "derivado").

Distância Genética: medida quantitativa da divergência genética entre organismos.

Espaço Amostral de Filogenias: espaço teórico que inclui todas as filogenias possíveis (com raiz ou sem raiz) para um determinado alinhamento.

Freqüência de equilíbrio: ??????

Grupos irmãos: clados que dividem um ancestral comum.

Homologia: similaridade originada por ancestralidade comum.

Inferência filogenética Bayesiana: método qualitativo de inferência filogenética baseado na estatística Bayesiana. Através da Cadeia de Markov Monte Carlos este método buscará as árvores mais prováveis dentro das filogenias amostradas.

Máxima Parcimônia: método qualitativo de inferência filogenética que busca a árvore que minimiza o número total de substituição de nucleotídeos.

Máxima Verossimilhança: método qualitativo de inferência filogenética que busca a árvore com a máxima verossimilhança.

Monofilia: associação entre o ancestral comum e todos os seus descendentes, formando um clado monofilético.

Múltiplas Substituições: eventos múltiplos de substituição de nucleotídeo localizado em um mesmo sítio do DNA.

Modelos de Substituição: modelos matemáticos utilizados para descrever o processo evolutivo ao longo do tempo, podendo ser aplicados ao alinhamento de nucleotídeos ou aminoácidos.

Aproximação dos vizinhos: *neighbor joining* (NJ), método de inferência filogenética quantitativo baseado em distância genética.

Ortólogo: genes homólogos em diferentes organismos e que mantém a mesma função.

OTU: unidade taxonômica operacional, folha ou nó terminal em uma árvore filogenética.

Parafilia: associação entre o ancestral comum e apenas parte de seus descendentes, formando um clado parafilético.

Parálogo: genes homólogos de um mesmo organismo que divergiram após duplicação.

Plesiomórfico: dotado de características do ancestral que são conservadas nos descendentes.

Polifilia: associação entre diferentes OTUs sem a necessidade de um único ancestral comum, frequentemente originada por convergência evolutiva.

Primitivo: diz-se de características ou organismos ancestrais, anteriores no tempo evolutivo a organismos ou características mais recentes.

Probabilidades Anteriores: distribuição dos valores de um parâmetro filogenético que é sabido de antemão pelo pesquisador.

Probabilidades Posteriores: conjunto da distribuição dos valores de parâmetros filogenéticos resultantes do método de inferência Bayesiana.

Sistemática: estudo da diversificação das formas vivas e suas relações ao longo do tempo.

Taxonomia: estudo que busca agrupar os organismos com base em suas características e nomear os grupos obtidos, classificando-os em alguma escala.

Taxon: grupo (de qualquer nível hierárquico) proposto pela taxonomia.

Topologia: descreve a ordem e a disposição exata das OTUs em uma filogenia.

UPGMA: *unweighted pair-group method using arithmetic average*, método de inferência filogenética quantitativo baseado em distância.

6.11. Leitura recomendada

FELSENSTEIN, Joseph. ***Inferring Phylogenies***. Sunderland: Sinauer, 2004.

LEMEY, Philippe; SALEMI, Marco; Vandamme, Anne-Mieke (Eds.). ***The Phylogenetic Handbook***. 2.ed. Cambridge: Cambridge University Press, 2009.

MATIOLI, Sergio Russo; FERNANDES, Flora M.C. (Eds.). ***Biologia Molecular e Evolução***. 2.ed. Ribeirão Preto: Holos, 2012.

NEI, Masatoshi; KUMAR, Sudhir. ***Molecular Evolution and Phylogenetics***. Nova Iorque: Oxford University Press, 2000.

PABÓN-MORA, Natalia; GONZÁLEZ, Favio. ***A classificação biológica: de espécies a genes***. In: ABRANTES, Paulo C. (Org.), Filosofia da Biologia. Porto Alegre: Artmed, 2011.

SCHNEIDER, Horacio. ***Métodos de Análise Filogenética: Um Guia Prático***. 3.ed. Ribeirão Preto: Holos, 2007.

6. Filogenia Molecular

Apêndice E

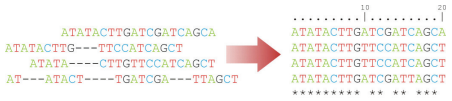
Alinhamentos

Ligabue-Braun R, Junqueira DM, Verli H

In Verli H (Ed.), *Bioinformática: da Biologia à Flexibilidade Molecular*

O capítulo a seguir, apresentado em sua formatação preliminar, é parte do livro-texto “Bioinformática: da Biologia à Flexibilidade Molecular” que encontra-se em fase final de preparação. Com distribuição eletrônica gratuita, seu lançamento está previsto para maio de 2014.

3. Alinhamentos



Alinhamento de múltiplas sequências.

Rodrigo Ligabue Braun
Dennis Maletich Junqueira
Hugo Verli

3.1. Introdução

O avanço nas técnicas de sequenciamento do DNA tem permitido um crescente aumento no número de genomas disponíveis em bancos de dados públicos. Esta maior disponibilidade exigiu um grande aumento na capacidade computacional de armazenamento e no investimento em desenvolvimento de técnicas de processamento adequadas para a análise destes dados. Algoritmos de análise tiveram de ser criados e aperfeiçoados e, dentre estes, as técnicas de alinhamento de sequências tornaram-se ferramentas essenciais e primordiais na análise de sequências biológicas. Atualmente, diversos programas on-line, ou mesmo de instalação local, são capazes de alinhar centenas de sequências em poucos minutos.

Devido à extensão de suas aplicações, o alinhamento de sequências biológicas é um processo de fundamental importância para a bioinformática. Conceitualmente, os alinhamentos são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas.

Em geral, as moléculas consideradas por estes programas, sejam elas formadas por nucleotídeos (DNA ou RNA) ou aminoácidos (peptídeos e proteínas), são polímeros representados por uma série de caracteres, e a comparação entre as moléculas depende apenas da comparação entre as respectivas letras. Apesar da facilidade e da aparente simplicidade do processo, a análise de similaridade das sequências é uma tarefa complexa e uma etapa decisiva para grande parte dos métodos de bioinformática que fazem uso de sequências biológicas.

Durante o alinhamento, as sequências são organizadas em linhas e os caracteres biológicos integram as colunas do alinhamento (Figura 1-3). Seguido à organização inicial, algoritmos específicos

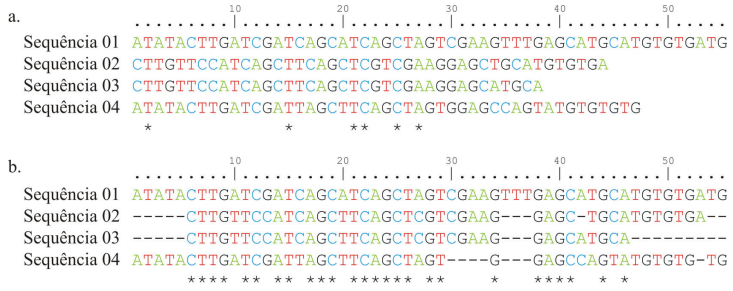


Figura 1-3: Alinhamento de quatro sequências de nucleotídeos envolvendo 55 caracteres. *a*) Grupo de sequências não alinhadas, cada sequência ocupando uma linha individual. *b*) Grupo de sequências alinhadas, onde caracteres idênticos são dispostos em uma mesma coluna e estas são identificadas por asteriscos (dispostos na parte inferior do alinhamento). Nucleotídeos ausentes em determinadas sequências são substituídos por hífens para identificar eventos de inserção/deleção.

buscarão a melhor correspondência para as sequências em questão, permitindo a criação de espaços entre estes caracteres para que, ao final, todas as sequências tenham o mesmo comprimento. Isto possibilita uma fácil visualização da similaridade, permitindo que caracteres idênticos ou similares em cada uma das sequências integrem a mesma coluna. A ideia central destes algoritmos é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. Comumente, a similaridade entre as sequências envolvidas é expressa pelo termo identidade, que quantifica a porcentagem de caracteres idênticos entre duas sequências.

A relevância e abrangência do uso do método tornam os procedimentos de alinhamento o cerne para diferentes campos dentro da grande área da bioinformática. Além de fundamentais em pesquisas de filogenética e análise evolutiva, os alinhamentos são exigidos em estudos de inferência estrutural e funcional de proteínas, análises de similaridade e identificação de sequências e em estudos aplicados ao campo da genômica.

Através dos métodos de alinhamento, é possível obter informações a respeito da relação evolutiva entre organismos, indivíduos, genes ou entre sequências diversas (Figura 2a-3). Se duas sequências distintas podem ser alinhadas com certo grau de similaridade, é possível inicialmente assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, por isso, são evolutivamente relacionadas. A partir da separação destas sequências de seu ancestral comum, individualmente cada uma delas acumulou diferentes variações ao longo do processo evolutivo. O termo homologia é utilizado frequentemente para definir estes eventos onde, através da relação de ancestralidade, dois indivíduos distintos possuem regiões em seu DNA (incluindo regiões codificantes) herdadas de um ancestral comum. Neste caso, a similaridade deve-se à descendência comum e, portanto, as sequências envolvidas na análise são ditas homólogas.

Cabe ressaltar que a homologia não requer necessariamente alta identidade de caracteres entre as sequências, uma vez que a maior ou menor identidade entre elas dependerá da taxa de evolução do organismo ou da espécie (consultar capítulo 06). Ainda, a similaridade entre sequências pode ser gerada não somente por descendência, mas por pressão seletiva de um determinado ambiente. Nestes casos, teremos regiões similares na sequência de nucleotídeos (ou aminoácidos) que surgiram de maneira independente, sem qualquer relação de

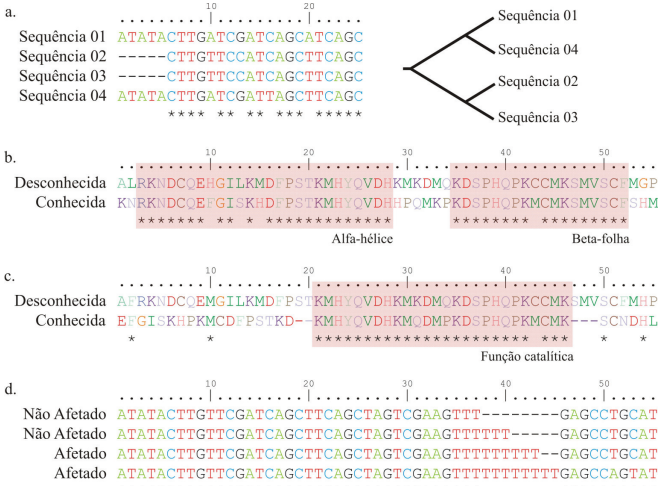


Figura 2-3: Aplicações dos métodos de alinhamento de sequências biológicas. *a*) Inferência filogenética a partir do alinhamento de quatro sequências de nucleotídeos. *b*) Inferência da estrutura de uma proteína-alvo (Desconhecida) a partir do alinhamento com uma sequência de aminoácidos cuja estrutura tridimensional é conhecida (Conhecida). *c*) Inferência da função de um domínio protéico a partir da comparação de sequências de aminoácidos. *d*) Comparação de sequências de uma porção de determinado gene de indivíduos afetados e não afetados por uma doença genética. Os asteriscos identificam colunas com total similaridade dos caracteres.

descendência, e evoluíram por convergência, não sendo portanto homólogas. Assim, não é possível quantificar a homologia entre as sequências envolvidas, somente dizer se há ou não. Quando identificamos quantos caracteres se repetem nas mesmas posições entre duas ou mais sequências estamos, de fato, verificando a identidade entre estas, e não a homologia.

As técnicas de alinhamento vêm se mostrando fundamentais na construção de algoritmos que visam comparar a informação de diversas sequências biológicas. À exemplo do programa BLAST, estes algoritmos permitem comparar uma sequência-alvo com milhares de dados disponíveis em grandes bancos de armazenamento, fornecendo um valor de significância estatística associada à esta comparação de similaridade. Devido à facilidade de acesso e rapidez no processamento de dados, estes programas vêm cada vez mais ampliando as possibilidades e opções para o tipo de comparação ou pesquisa a ser realizada.

Os métodos de alinhamento podem ainda ser necessários para fornecer informações a respeito da função e da estrutura de sequências biológicas, particularmente nos alinhamentos de ribonucleotídeos e aminoácidos (Figura 2-3). Nestes casos, a similaridade entre duas ou mais sequências (dada em porcentagem) revela padrões referentes à composição química e podem fornecer embasamento para a definição de um arranjo tridimensional semelhante, principalmente no caso de proteínas (Figura 2b-3). A mesma relação é feita para inferir a função de domínios de uma proteína recém-descoberta, ainda sem função definida. Sabendo que sua forma está diretamente relacionada à sua função, através da comparação com outras proteínas

com estrutura e função já estabelecidas, é possível inferir a função realizada por determinado domínio da proteína sob investigação (Figura 2c-3). Nestes casos, as sequências envolvidas no alinhamento não são necessariamente homólogas. Através do fenômeno da evolução convergente, diferentes regiões codificantes do DNA podem gerar produtos proteicos com funções similares, sem obrigatoriamente compartilharem um ancestral comum.

Finalmente, as técnicas de alinhamento têm grande importância para a análise de genes e genomas. Com o aumento da disponibilidade de sequências nucleotídicas de genomas completos, e mesmo com o surgimento de modernas técnicas de biologia molecular, como o microarray e deep sequencing, os métodos de comparação permitiram o entendimento a respeito da variabilidade genética de indivíduos e populações.

A comparação entre genomas de diferentes espécies, ou até mesmo de indivíduos da mesma espécie, possibilita a análise de variações (mutações ou polimorfismos) nas sequências e, em alguns casos, permite a identificação de relações entre variações no DNA e susceptibilidade a determinadas doenças, beneficiando o campo da genética e áreas relacionadas. Adicionalmente, como um recurso para a caracterização de eventos evolutivos, os alinhamentos permitem análises comparativas entre genomas. A abrangência e importância evolutiva dos eventos de quebra e reparo de DNA, ou mesmo dos eventos de recombinação, inversões e translocações, tem sido desvendados, primariamente, através dos métodos de alinhamento.

Além do alinhamento de sequências, o alinhamento de estruturas constitui outra importante ferramenta em estudos de bioinformática. A metodologia é bastante diferente daquela empregada em alinhamentos de sequências, pois passamos de um problema unidimensional para um problema tridimensional. Sua utilização passou a ser difundida a partir de 1978, com o trabalho de Rossmann e Argos, comparando os sítios ativos de enzimas cujas estruturas eram conhecidas até aquele momento. Os métodos de sobreposição simples de estruturas estão disponíveis há mais tempo, tendo sido propostos a partir da década de 1970, enquanto os métodos de comparação e alinhamento se desenvolveram posteriormente, principalmente a partir da década de 1990.

A comparação de estruturas se refere à análise de similaridades e diferenças entre duas ou mais estruturas, enquanto o alinhamento de estruturas se refere à determinação de quais aminoácidos seriam equivalentes entre tais estruturas. É importante destacar também a diferença entre alinhamento e sobreposição de estruturas. Apesar desses termos ainda serem empregados na literatura como sinônimos, eles se referem a procedimentos diferentes. Conforme mencionado acima, enquanto o alinhamento de estruturas busca identificar equivalências entre pares de aminoácidos nas estruturas a serem sobrepostas, a sobreposição necessita desse conhecimento prévio sobre as equivalências.

Sendo assim, a sobreposição estrutural busca solucionar um problema muito mais simples, ou seja, minimizar a distância entre dois resíduos já reconhecidos como equivalentes. Isso se dá por encontrar transformações que encontram ou o menor desvio médio quadrático (RMSD) ou as equivalências máximas dentro de um valor limite para o RMSD.

Considerando que a estrutura das proteínas é mais conservada que a sequência, o alinhamento de estruturas confere maior especificidade ao alinhamento de sequências quando comparado ao alinhamento de sequências independente de estrutura. A maioria dos métodos de sobreposição de estruturas é adequado para identificar similaridades entre estruturas proteicas. O alinhamento de duas ou mais estruturas, porém, constitui uma tarefa mais difícil, e sua precisão depende tanto do método usado quanto do objetivo do usuário.

3.2. Alinhando sequências

À primeira vista, o processo de alinhamento entre diferentes sequências parece simples e não sujeito a qualquer tipo de erro. No entanto, esta afirmativa só é verdadeira em casos onde os organismos envolvidos possuem uma baixa taxa evolutiva (Figura 3a-3). Quando consideramos sequências homólogas amostradas de organismos com alta taxa evolutiva, ou até mesmo sequências similares, porém não homólogas, nos deparamos com casos particulares que tornam o processo de alinhamento complexo e, muitas vezes, sujeito a uma interpretação especialmente subjetiva por parte do usuário (Figura 3b-3).

A comparação de sequências homólogas de organismos evolutivamente distantes é um desafio para os programas de alinhamento. As diferentes pressões seletivas moldam os genomas de maneira imprevisível e, muitas vezes, acarretam a perda ou ganho de nucleotídeos ao longo do processo evolutivo. Para estes casos, a adição de lacunas (gaps) em matrizes de alinhamento, representadas por "-", é possível e muitas vezes necessária. As lacunas representam um ou mais eventos de inserção ou deleção de nucleotídeos. Estes eventos, comumente chamados de "indels" (in para inserção, e del para deleção), são fruto de processos mutagênicos (espontâneos ou induzidos) e, dependendo da região atingida, podem ser

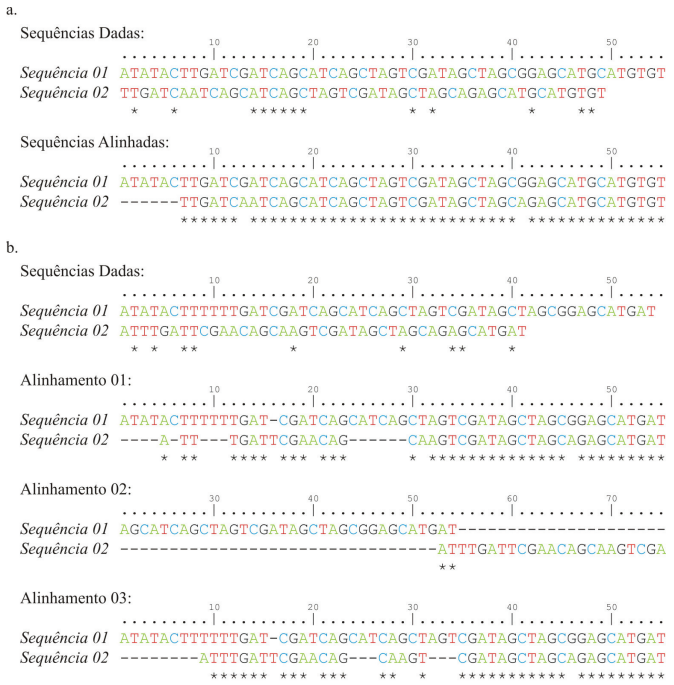


Figura 3-3: Alinhamentos de nucleotídeos. a) Duas sequências homólogas originadas de organismos com baixa taxa de evolução são dadas e seu alinhamento é proposto. b) Duas sequências homólogas amostradas de organismos com alta taxa de evolução são dadas e diferentes alinhamentos são propostos. Os hífens representam eventos de inserção ou deleção únicos na sequência. Os asteriscos identificam colunas com total similaridade dos caracteres.

expressos nas moléculas de RNA e nas proteínas, onde poderão gerar consequências moleculares. Erros de replicação gerados pela DNA-polimerase durante a replicação do DNA, ou mesmo os eventos de recombinação, são os principais fatores atrelados à geração destes indels nos genomas. Em regiões codificadoras, estes eventos podem acarretar mudanças no quadro de leitura da proteína e torná-la não funcional.

Em termos analíticos, a inserção de lacunas dificulta o processo de alinhamento e exige interpretações cautelosas. Para determinados casos, especialmente em análises evolutivas e filogeográficas, é comum que regiões do alinhamento com determinado nível de incerteza, especialmente regiões com grande número de lacunas, sejam eliminadas da análise. Contudo, até o momento não existem programas capazes de lidar com as lacunas de forma coerentemente biológica. Apesar de sabermos que se tratam de eventos evolutivos comuns e bem caracterizados, as incertezas sobre o número de eventos e sua intensidade tornam as lacunas, em grande parte dos casos, um fator de confusão para análises de alinhamento.

Conforme mostrado na Figura 3-3, diferentes alinhamentos são possíveis para um mesmo grupo de sequências. A pergunta que se segue é: como reconhecer o melhor resultado quando nos deparamos com diversos alinhamentos possíveis para um mesmo conjunto de dados? Buscou-se resolver este problema através da criação de um sistema de pontuação para comparar os resultados de diferentes alinhamentos. Caracteres idênticos em sequências diferentes representam igualdades ou correspondências (matches) e, por serem resultados preferenciais durante o processo de alinhamento, são pontuados positivamente. Pelo contrário, caracteres não idênticos que ocupam a mesma coluna são chamados de desigualdades, ou mismatches, e recebem atribuições negativas. Como resultado, o melhor alinhamento possível para duas sequências é aquele que maximiza

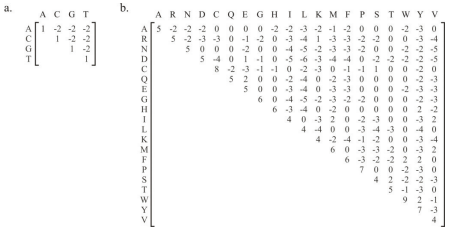


Figura 4-3: Matrizes de custo utilizadas no cálculo de pontuação dos alinhamentos. a) Matriz de custo exemplo utilizada para cálculos de pontuação em alinhamentos de nucleotídeos. b) Matriz de custo BLOSUM62 utilizada para cálculo da pontuação em alinhamentos de aminoácidos.

a pontuação total, somando os valores de matches e debitando os valores de mismatches.

Do ponto de vista biológico, as mudanças entre as bases nitrogenadas nas sequências de nucleotídeos não ocorrem com a mesma probabilidade (Figura 4a-3). Sendo assim, podemos atribuir valores de mismatches diferentes às transições (trocas de purinas por purinas ou pirimidinas por pirimidinas) e às transversões (trocas de purinas por pirimidinas ou pirimidinas por purinas). Para sequências de aminoácidos, é necessário escolher ativamente uma matriz de pontuação específica. Essas matrizes são resultados diretos de estudos de variação proteica e estão diretamente relacionadas à probabilidade de substituição de um aminoácido por outro (matrizes BLOSUM e PAM). Atualmente, as matrizes BLOSUM são as mais disseminadas e aplicadas para os mais diversos casos de comparação entre sequências de aminoácidos (Figura 4b-3).

Ainda, é necessário que as lacunas de alinhamentos recebam determinadas pontuações, pois são frequentemente encontradas em alinhamentos de dados biológicos. Se lacunas podem ser adicionadas em qualquer posição sem qualquer restrição, tanto nas extremidades quanto no interior das sequências, é possível gerar

alinhamentos com mais lacunas do que propriamente caracteres a serem comparados (Figura 3b-3, alinhamento 2). Com o intuito de prevenir inserção excessiva, a adição de lacunas é penalizada durante a atribuição da pontuação de uma sequência, conforme um conjunto de parâmetros, chamado de penalidades por lacuna (gap penalties, PL). A abrangência da lacuna é pontuada pelo respectivo número de indels presentes no alinhamento. A fórmula mais comum para cálculo destas penalizações segue abaixo:

$$PL = g + e\,(L-1)$$

, onde L é o tamanho da lacuna (número de indels presentes na lacuna), g é a penalidade pela abertura da lacuna (necessária para evitar que os alinhamentos contenham lacunas desnecessárias) e e é a penalidade atribuída a cada indel (novamente para evitar grandes lacunas sem necessidade). Os valores de penalidade por lacuna são desenhados para reduzir a pontuação de um alinhamento quando este possui uma quantidade de indels desnecessária. Apesar da disseminação deste conceito, não há qualquer relação matemática ou biológica sustentando este cálculo. É importante destacar que, através da propriedade de "alinhamento livre de colunas em branco" (ou seja, gaps não são alinhados), as penalizações ainda impedem o alinhamento de indels entre as sequências envolvidas na análise. Assim, o melhor alinhamento entre as sequências será dado por um valor que resulta da soma dos valores associados a cada um dos matches, mismatches e lacunas, de acordo com um critério pré-definido (Figura 5-3).

O método de pontuação foi a solução encontrada para avaliar e classificar diferentes alinhamentos em busca da melhor explicação para a relação evolutiva entre as sequências. O próximo problema encontrado foi enumerar todas as possibilidades de alinhamentos para um grupo de dados. Assumindo-se duas sequências com tamanho de 100 caracteres cada, poderíamos enumerar até 10^{77} possíveis alinhamentos, diferentes entre si. A extensão de possibilidades inviabiliza a enumeração de todos os casos devido ao tempo e ao requerimento de enorme processamento destes dados. Apesar da exigência computacional, alguns algoritmos são capazes de realizar tal tarefa e ainda aplicar o método de pontuação para cada um dos casos, em busca do melhor resultado. No entanto, estes algoritmos não são capazes de lidar com sequências que contenham mais que algumas dezenas de caracteres. Em virtude da capacidade de explorar todas as soluções do problema, o processo realizado por estes algoritmos é chamado de "alinhamento ótimo".

Contudo, em virtude da inerente demora do processo, foi necessário desenvolver algoritmos que acelerassem a busca de um alinhamento capaz de explicar de maneira ótima os processos evolutivos para um determinado grupo de sequências sem, no entanto, enumerar todas as possibilidades. Os alinhamentos gerados por estes programas são chamados heurísticos, e compreendem métodos aproximados de busca pelo resultado ótimo. Diferentes métodos foram criados para diferentes tipos de alinhamento (Figura 6-3). Entre estes, devido à eficiência e a rapidez de processamento das informações de um alinhamento, incluindo o cálculo de pontuação, os algoritmos de programação dinâmica são, atualmente, os mais utilizados para este fim, tanto em alinhamentos simples como integrado aos algoritmos de alinhamentos múltiplos.

É fundamental assumirmos, para a maior parte dos problemas em bioinformática, o alinhamento como um modelo de relação evolutiva entre as sequências envolvidas. E como modelo, está sujeito a presença de certos problemas na explicação dos eventos evolutivos reais. Portanto, os alinhamentos devem ser avaliados com extrema cautela. A facilidade e a aparente simplicidade na análise dos programas tornam o processo mecânico e

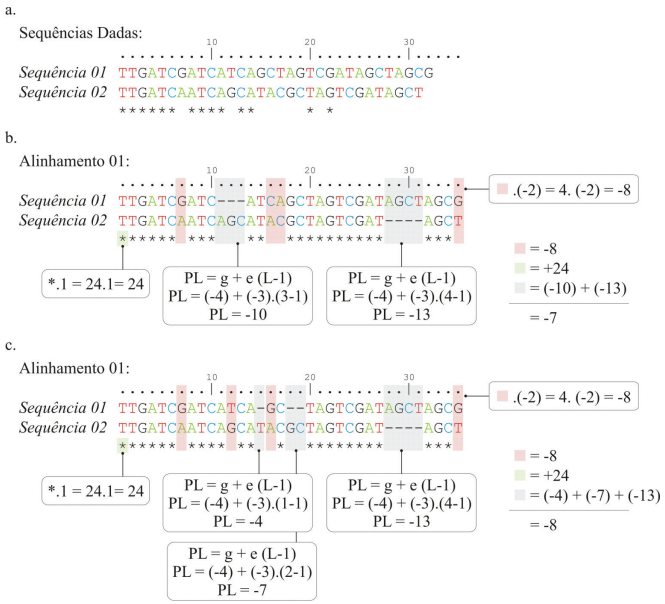


Figura 5-3: Esquema de pontuação para avaliação de alinhamentos. *a*) Duas sequências de desoxirribonucleotídeos não alinhadas. *b*) Proposição de um alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de matches, 4 colunas de mismatches e duas lacunas com 3 e 4 indels. A pontuação total para o alinhamento desta sequência é -7. *c*) Proposição de um segundo alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de matches, 4 colunas de mismatches e três lacunas com 1, 2 e 4 indels. A pontuação total para o alinhamento desta sequência é -8. A partir deste exemplo, o alinhamento com a maior pontuação é o mostrado em *b*. Os valores de pontuação utilizados neste exemplo são especificados na parte inferior da figura.

desvinculado de análises críticas pela maior parte dos usuários. A associação dos métodos de alinhamento a outras análises de bioinformática tende a desvincular a real importância desta técnica e a coloca apenas como um procedimento, e não formalmente como uma técnica sujeita à análise crítica. Isto pode ocasionar na obtenção de modelos incorretos ou mesmo de falsos positivos.

3.3. Tipos de alinhamento

Em estudos de bioinformática, é comum compararmos moléculas de dois ou mais indivíduos, sejam eles da mesma espécie ou de espécies diferentes. Quanto maior o número de sequências comparadas, maior o tempo exigido para conclusão do alinhamento e, dependendo das sequências envolvidas, maior a dificuldade dos algoritmos em encontrar o melhor resultado. Conforme a quantidade de sequências envolvidas, podemos dividir os alinhamentos em dois tipos: alinhamentos simples, ou par-a-par, e alinhamentos múltiplos, ou de múltiplas sequências (Figura 7-3). Os alinhamentos simples descrevem especificamente a relação de similaridade entre duas sequências

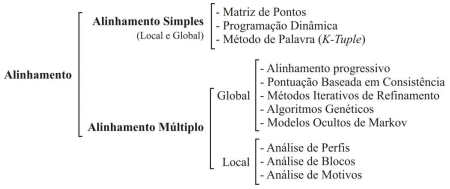


Figura 6-3: Tipos de alinhamento e os algoritmos aplicados à bioinformática.

quaisquer. Já os alinhamentos múltiplos incluem três ou mais sequências na análise de similaridade e, dependendo do objetivo do usuário, podem envolver até centenas de sequências.

Conceitualmente, ainda podemos dividir os alinhamentos, tanto simples, como múltiplos, em dois grandes tipos. Os alinhamentos que levam em consideração toda a extensão das sequências são conhecidos como globais, enquanto aqueles que buscam pequenas regiões de similaridade são chamados de locais (Figura 7-3). Em algoritmos que buscam o alinhamento global de duas sequências, reforça-se a busca do alinhamento completo das sequências envolvidas, procurando incluir o maior número de matches do início ao final das sequências. Quando necessário, estes algoritmos permitem a inserção de lacunas para que as sequências tenham o mesmo tamanho no resultado do alinhamento (Figura 7b-3).

Graficamente, os sítios com caracteres idênticos são representados ligados por barras verticais, enquanto que os sítios que possuem caracteres diferentes nas duas sequências, ou mesmo a presença de uma lacuna em uma delas, permanecem sem qualquer notação (Figura 7-3). O principal algoritmo envolvido no processamento de alinhamentos globais é aquele desenvolvido por Needleman e Wunsch durante a década de 1970. Além de ter uma notável importância metodológica, este algoritmo tem grande importância na história do alinhamento, pois foi o primeiro algoritmo a aplicar o método de programação dinâmica para a comparação de sequências biológicas.

Em seu início, os métodos de alinhamento eram utilizados especialmente para a comparação par-a-par de sequências de proteínas inteiras. No entanto, com a

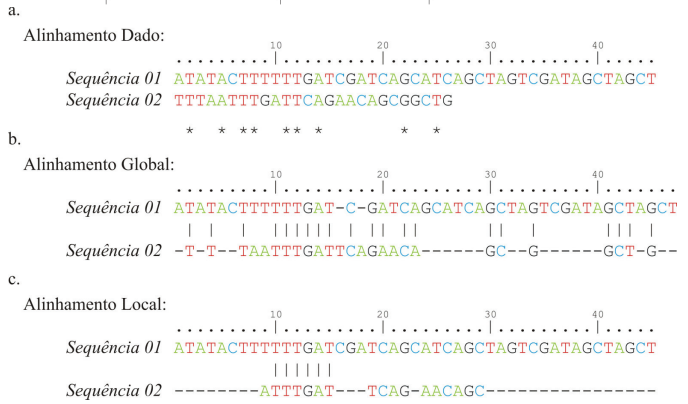


Figura 7-3: Diferenças entre alinhamento local e global. *a*) Duas sequências de nucleotídeos de tamanhos diversos são amostradas e alinhadas por algoritmos diferentes. *b*) No alinhamento local, a prioridade é encontrar as regiões altamente similares, independentemente do tamanho desta região. Neste caso, porções da sequência que não foram alinhadas com alta similaridade foram excluídas do resultado final. *c*) No alinhamento global, as duas sequências são alinhadas por completo, independentemente do número de lacunas que tenham que ser inseridos.

ampliação da disponibilidade de sequências completas de proteínas, foi necessário buscar métodos de alinhamento que privilegiassem a busca de similaridade, não entre sequências completas, mas apenas entre porções isoladas destas sequências. Durante a década de 1980 iniciou-se o desenvolvimento de novos algoritmos de alinhamento, já que os desenvolvidos até aquele momento não eram aplicáveis para esta particularidade. Entre estes novos algoritmos, o desenvolvido por Smith e Waterman, em 1981, ganhou maior destaque e atualmente é o principal algoritmo utilizado por programas para realização de alinhamentos locais. Nestes casos, privilegia-se o alinhamento de partes da sequência, buscando apenas as regiões com a maior similaridade (Figura 7c-3). Em algoritmos para busca local, o alinhamento pára no final das regiões de alta similaridade e substitui as regiões excluídas por hífens (lacunas) no resultado final (Figura 7c-3).

3.4. Alinhamento simples

Para entender como se processa um alinhamento para-a-par e como o grau de similaridade entre elas pode ser computado, apresentamos três dos principais algoritmos desenvolvidos para este fim: algoritmos de programação dinâmica, análise de matriz de pontos (dot matrix) e método de palavra ou ***k***-tuple.

A programação dinâmica é, atualmente, o método mais utilizado por programas para realizar o alinhamento de sequências. Em casos simples (par-a-par), é capaz de encontrar o melhor alinhamento para duas sequências através da aplicação da pontuação de similaridades. É, portanto, um método de execução relativamente rápida nos computadores modernos, requerendo um tempo e memória de processamento proporcional ao produto do tamanho das duas sequências envolvidas.

O método é baseado no princípio de otimização de Bellmann, e propõe a solução de problemas complexos através da resolução dos seus diversos subproblemas. Os subproblemas são resolvidos e seus resultados são armazenados pelo algoritmo. A vantagem funcional da resolução em partes é que, geralmente, problemas complexos combinam uma série de subproblemas. Como o algoritmo acumula os resultados dos diferentes subproblemas, acelera a resolução do problema complexo. Assim, a designação "programação" nada tem a ver com programação de computadores, mas com a organização dos resultados já solucionados para resolução de um problema maior.

Conforme discutimos anteriormente, em determinados casos, duas sequências podem apresentar diferentes alinhamentos. Se não há indels e as sequências são similares, o alinhamento é rápido e não deixa dúvidas. No entanto, quando existe certa diversidade entre as sequências envolvidas e uma quantidade suficiente de indels, a solução para o alinhamento é menos óbvia visualmente. Nestes casos, os algoritmos de programação dinâmica buscarão solucionar os subproblemas envolvidos e fornecerão o melhor resultado.

Para cálculo do melhor alinhamento entre duas sequências, o algoritmo de programação dinâmica necessita da especificação de um esquema de pontuação, seja ele referente a nucleotídeos ou aminoácidos. Da mesma forma, é necessário fornecer um valor de penalidade para a abertura e extensão das lacunas. A partir destas informações, o algoritmo calculará uma relação entre todos os caracteres das sequências e fornecerá o melhor alinhamento como resultado final.

Como exemplo, consideraremos a Figura 8-3. São dadas duas sequências, sequência 1 e sequência 2, um esquema de pontuação e, para facilitar o entendimento do cálculo, um valor único de penalidade por lacuna de -8. O algoritmo toma as sequências e transforma a relação entre elas em uma tabela, onde as linhas são definidas pelos caracteres da sequência 01, e as colunas pelos caracteres da sequência 02. A fim de permitir lacunas no início do alinhamento, o algoritmo impõe a inserção de uma coluna e

de uma linha iniciais contendo o símbolo de indel. A partir deste ponto, para cada um dos elementos da matriz, o algoritmo calculará a melhor pontuação dos subcaminhos associados ao alinhamento: uma substituição, uma inserção na sequência 01 ou uma inserção na sequência 2. Assim, o melhor subcaminho será calculado segundo uma função de pontuação, conforme abaixo:

$$F(i,j) = max \begin{cases} valor\ da\ célula\ na\ diagonal\ superior\ esquerda + pontuação\ da\ similaridade \\ valor\ da\ célula\ acima + valor\ da\ penalidade\ por\ lacuna \\ valor\ da\ célula\ à\ esquerda + valor\ da\ penalidade\ por\ lacuna \end{cases}$$

A partir do elemento (1,1) da matriz e ao longo da primeira linha, apenas a terceira condição é satisfeita (valor da célula à esquerda + valor da penalidade por lacuna). Na primeira coluna, apenas a segunda condição é satisfeita. Para outros elementos, as três condições devem ser calculadas e aquela que resultar no maior valor é escolhida para formar a matriz. Além disso, os procedimentos dos algoritmos de programação dinâmica podem ser representados por pequenas setas para indicar qual o subcaminho obteve o melhor valor (Figura 8-3).

Outro método importante na área de alinhamento de sequências é a análise de matriz de pontos ou matriz dot. É um método simples e bastante eficiente em análises de deleções/inserções e para detectar repetições diretas ou inversas, especialmente em sequências de nucleotídeos. Além disso, vem sendo utilizado para buscar regiões de pareamentos intra-cadeia capazes de formar estruturas secundárias em moléculas de RNA. Este método permite a visualização gráfica das regiões de similaridade entre sequências através da construção de uma matriz de identidade. O número de linhas desta matriz é definido pelo número de caracteres de uma das sequências, e o número de colunas é definido pelo número de caracteres da outra sequência a ser comparada (Figura 9-3). É primariamente um método visual, e não fornece o alinhamento propriamente dito como resultado final, embora sej frequentemente utilizado quando se deseja visualizar as regiões de similaridade entre duas sequências.

Neste método, inicialmente, uma das sequências é disposta na vertical e a outra na horizontal (Figura 9-3). Regiões do gráfico que possuem o mesmo caractere tanto na sequência disposta na horizontal, quanto na sequência disposta na vertical, serão assinalados. Esta marcação representa os possíveis correspondências (matches) entre uma sequência e outra.

Qualquer região de similaridade entre as duas sequências será evidenciada por uma linha diagonal de assinalações. Pontos não dispostos na diagonal representam correspondências aleatórias que não estão relacionadas com a similaridade entre as sequências. A detecção de regiões de alta similaridade pode ser beneficiada, em alguns casos, através da comparação de dois ou mais caracteres ao mesmo tempo. Nestes casos, é necessário escolher um número de caracteres como janela.

Além disso, arbitrariamente, um número de correspondências deve ser escolhido. Por exemplo, para comparar duas sequências com 100.000 caracteres, podemos escolher uma janela de 15 caracteres e 10 correspondências requeridas. O algoritmo varrerá a matriz de 15 em 15 caracteres e, quando, entre estes quinze caracteres, existirem 10 formando correspondências entre as duas sequências, o algoritmo inserirá uma marcação de similaridade. Geralmente, esta variação do método é utilizada para a comparação de longas sequências de DNA.

Por último, outro algoritmo bastante comum no alinhamento par-a-par de dados biológicos é o ***k***-tuple, ou método de palavras. Este método é geralmente mais rápido que o método de programação dinâmica, embora não garanta o melhor alinhamento como resultado. Este tipo de algoritmo é especialmente útil em casos onde se busca similaridade de uma única sequência contra um grande conjunto de dados. Para isso, o algoritmo dividirá uma sequência-alvo em pequenas sequências, geralmente conjuntos entre dois e seis caracteres, chamados de palavras. Da mesma forma, o conjunto total

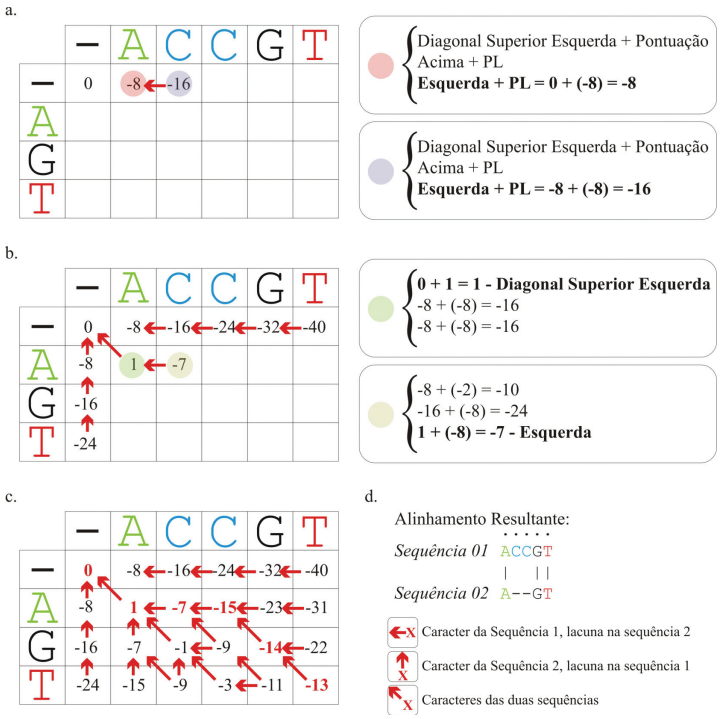


Figura 8-3: Alinhamento de duas sequências de nucleotídeos através do método de programação dinâmica. *a*) As sequências a serem alinhadas são dispostas em uma tabela onde o número de colunas corresponde ao número de caracteres da sequência 1 mais um (devido à adição de uma coluna para uma lacuna) e o número de linhas corresponde ao número de caracteres da sequência 2 mais um. O caracter atribuído à primeira linha e à primeira coluna é, por definição, o símbolo "-", atribuído a uma lacuna. Através da matriz de penalidades calcula-se os valores para as três possibilidades $F(i,j)$, buscando a equação que resulte no maior valor. O valor arbitrário de penalidade por lacuna (PL) é de -8. Em virtude de a primeira linha não possuir valores de comparação na diagonal superior esquerda e acima, considera-se apenas a terceira equação. *b*) O valor demarcado em verde é o primeiro a ser calculado após o preenchimento da primeira linha e primeira coluna, representando o menor valor encontrado no cálculo para $F(i,j)$. Além do cálculo, o algoritmo de programação dinâmica insere informações a respeito da direção da informação. Como o valor "1" foi o maior valor encontrado e representa o cálculo utilizando a informação situada na diagonal superior esquerda, demarcada em verde, insere-se uma seta nesta direção. *c*) O preenchimento completo da tabela e as respectivas setas ilustrando a direção da informação. Algumas casas estão demarcadas com duas setas, pois apresentaram dois valores máximos idênticos na resolução das equações. Ao final dos cálculos, iniciando pelo canto inferior direito, segue-se as setas em busca dos maiores valores. *d*) Relacionando os dados da tabela com a simbologia apresentada, chega-se ao alinhamento final entre as sequências 1 e 2.

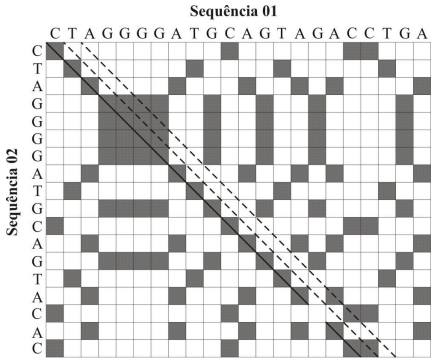


Figura 9-3: Análise de matriz de pontos de duas sequências de DNA. Os pontos assinalados em cinza representam a concordância de caracteres entre a sequência 1 e a sequência 2. A partir da diagonal direita inferior, são traçadas diferentes retas. Aquela que atingir o maior número de pontos assinalados deve ser escolhida como resultado para o alinhamento entre as duas sequências. A linha continua representa a possibilidade mais adequada a esta análise e as linhas tracejadas representam possibilidades de insucesso.

de sequências do banco de dados terá cada uma das sequências subdivida em pequenas palavras. As palavras da sequência-alvo serão comparadas às palavras oriundas do banco de dados. Após a busca de identidade, o algoritmo alinhará as duas sequências completas (sequência oriunda do banco de dados que teve uma palavra similar com umas das palavras da sequência-alvo e a própria sequência-alvo) a partir das palavras similares e estenderá a análise de similaridade para as regiões vizinhas, antes e depois da palavra similar. Através de uma matriz de penalidade, o algoritmo calculará o alinhamento que teve o maior valor de pontuação. É comum, para esta segunda etapa dos cálculos de similaridade, a utilização de algoritmos de programação dinâmica.

3.5. Alinhamento múltiplo global

Da mesma forma que no caso dos alinhamentos simples, o método de programação dinâmica é usualmente utilizado para lidar com múltiplas sequências. Nestes casos, utiliza-se o conceito de soma ponderada dos pares (weighted sum of pairs, WSP). Através deste conceito, para qualquer alinhamento múltiplo de sequências, uma pontuação para cada par possível formado por estas sequências será calculada (Figura 8-3) e, ao final, os valores de similaridade para cada um dos pares serão somados. Apesar de conceitualmente simples, este método exige grande capacidade computacional e, dependendo da quantidade de sequências envolvidas, pode requerer longo tempo para processamento.

Métodos alternativos tiveram que ser criados para acelerar os cálculos para alinhamento de sequências, incluindo-se: alinhamento progressivo, pontuação baseada em consistência (consistency-based scoring), métodos iterativos de refinamento, algoritmos genéticos e modelos ocultos de Markov. Cabe ressaltar que todos estes métodos realizam buscas aproximadas pelo resultado ótimo e, portanto, se tratam de métodos heurísticos.

Alinhamento progressivo

Leva em consideração a relação evolutiva entre as sequências. Os algoritmos utilizam as relações filogenéticas para gerar o resultado de alinhamento. Inicialmente, são realizados alinhamentos par-a-par de todos os possíveis pares. Nesta comparação, verifica-se apenas o número de caracteres diferentes entre as duas sequências (verificar o conceito de distância evolutiva observada no capítulo 6). Estas distâncias serão utilizadas para a construção de uma filogenia (geralmente através do método de neighbor-joining). A partir desta filogenia o alinhamento será construído progressivamente, dependendo da relação entre as sequências sendo, por isso, chamado de alinhamento

progressivo.

Tomemos como exemplo um ramo de uma dada filogenia que inclui duas sequências. O algoritmo construirá um alinhamento através de programação dinâmica para estas duas sequências. A partir deste primeiro alinhamento, estas duas sequências serão agora tratadas como uma, e serão alinhadas à próxima sequência filogeneticamente relacionada. Devemos notar que todo o restante das sequências será alinhado baseando-se neste primeiro par. É um método rápido e amplamente utilizado para alinhar um grande número de sequências. Atualmente, os programas mais populares de alinhamento progressivo são o CLUSTALW e CLUSTALX.

Pontuação baseada em consistência

Baseado no algoritmo de alinhamento progressivo, não leva em consideração apenas o primeiro par de sequências alinhadas. Durante a realização do cálculo, realiza outros alinhamentos par-a-par para aperfeiçoar as comparações entre as sequências. O principal programa a utilizar este algoritmo é o T-COFFEE.

Métodos iterativos de refinamento

Funcionam como os algoritmos de alinhamento progressivo, mas os grupos de sequências são realinhados constantemente ao longo das análises, garantindo que o alinhamento inicial não defina o resultado final. O principal programa a utilizar este algoritmo como base para os cálculos de alinhamento é o MUSCLE.

Algoritmos genéticos

Estes algoritmos buscam simular o processo evolutivo no conjunto de sequências a serem alinhadas, aplicando conceito de seleção e recombinação. É ainda um método lento e, devido à aleatoriedade do processo, não garante o mesmo resultado para diferentes alinhamentos do mesmo conjunto de dados. O programa SAGA é um dos poucos a implementar algoritmos genéticos.

Modelos ocultos de Markov

Modelo baseado em probabilidades estatísticas, destacando os eventos de substituição e inserção ou deleção de caracteres.

3.6. Alinhamento múltiplo local

Na busca por regiões localizadas de similaridade entre diferentes sequências, são aplicados principalmente os seguintes algoritmos: análise de perfis, análise de blocos e análise de motivos.

Análise de perfis

A partir de um alinhamento primário de todas as sequências envolvidas na análise e utilizando uma matriz de custo padrão, o algoritmo seleciona as regiões altamente conservadas e produz uma nova matriz de pontuação (matriz de custo), chamada de perfil. A construção deste perfil pode ser realizada através de dois métodos diferentes (método das médias e método evolutivo) e inclui pontuações para matches, mismatches e lacunas. Assim que produzido, este perfil pode ser utilizado para alinhar sequências entre si utilizando as pontuações calculadas para avaliar a probabilidade em cada posição ou para buscar sequências com o mesmo padrão em um banco de dados.

A desvantagem do método de perfis está na especificidade da nova matriz de custo obtida. Se o alinhamento inicial contiver poucas sequências, pode não representar adequadamente a variabilidade de caracteres em uma determinada posição e prejudicar o algoritmo na busca por similaridade com outras sequências. Este método é principalmente utilizado para alinhamentos de aminoácidos.

Análise de blocos

Assim como a análise de perfis este método requer, inicialmente, a seleção da região de maior similaridade de um alinhamento múltiplo. Estas regiões podem ser chamadas de blocos e diferem dos perfis por não acomodarem indels, que serão automaticamente eliminados das análises. Este método é também capaz de realizar a busca de pequenas regiões de similaridade entre sequências, de maneira semelhante ao método de palavras.

Análise de motivos

Este método é especialmente utilizado na busca por motivos proteicos em sequências de aminoácidos. O método foi desenvolvido através do alinhamento de milhares de sequências de aminoácidos extraídas de grandes bancos de dados de proteínas. A partir deste alinhamento, analisou-se cada uma das colunas para buscar um padrão de substituição entre os aminoácidos. Estes padrões de mudança refletem uma maior probabilidade de substituição. Para proceder ao alinhamento, os algoritmos que aplicam a análise de motivos iniciam o processo por uma análise de blocos. As regiões de alta similaridade são então analisadas para buscar os padrões de substituição descritos inicialmente. O conjunto de padrões resultante da análise das colunas é chamado de motivo. A probabilidade de existência de cada motivo em uma sequência de proteína é estimada através do banco de dados do SwissProt.

3.7. BLAST

O BLAST, ou Ferramenta de Busca por Alinhamento Local Básico (Basic Local Alignment Search Tool) é um algoritmo capaz de realizar buscas baseadas em alinhamento que, apesar de não serem exatas, são confiáveis e muito rápidas, sendo estas suas vantagens em relação a outros métodos. Ele é um dos programas mais usados em Bioinformática devido à velocidade em que consegue responder a um problema fundamental em biologia celular e molecular: comparar uma sequência desconhecida com aquelas depositadas em bancos de dados.

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências ao buscar primeiro por palavras comuns (ou *k*-tuples) na sequência de busca e em cada sequência do banco de dados. Em vez de buscar todas as palavras de mesmo tamanho, o BLAST limita a busca àquelas palavras que são mais significantes. O tamanho de palavra é fixado em 3 caracteres para sequências de aminoácidos e em 11 para sequências de nucleotídeos (3 se as sequências forem traduzidas nos 6 quadros de leitura possíveis). Esses são os tamanhos mínimos para obter uma pontuação por palavras que seja alta o suficiente para ser significativa sem perder fragmentos menores, mas importantes, de sequência.

Funcionamento do algoritmo BLAST

Para funcionar, o BLAST necessita de uma sequência de busca (query) e de sequências-alvo. Comumente, as sequências-alvos são o conjunto de sequências depositadas em um banco de dados, local ou na web. Um dos conceitos principais empregados pelo BLAST é de que alinhamentos estatisticamente significantes contêm pares de segmentos de alta pontuação (HSP, high-scoring segment pairs), e são esses HSPs que o algoritmo busca entre a sequência sendo analisada e aquelas depositadas no banco de dados.

As principais etapas do funcionamento do algoritmo BLAST, para uma sequência proteica genérica incluem:

i. Remoção de repetições ou regiões de baixa complexidade na sequência de busca.

Uma região de baixa complexidade é definida como uma região composta por poucos tipos de elementos. Essas regiões normalmente apresentam pontuações altas que podem confundir o programa em sua busca por sequências com similaridade significativa. Por esse motivo, tais regiões são identificadas antes da próxima

etapa e ignoradas.

ii. Estabelecer uma lista de palavras com *k*-letras.

Sendo este um caso envolvendo sequências proteicas, $k = 3$, ou seja, cada palavra tem tamanho 3. Como mostrado na Figura 10-3, são listadas palavras com comprimento de 3 caracteres, sequencialmente, até que a última letra da sequência de busca seja incluída.

Figura 10-3: Exemplo de lista de palavras geradas pelo BLAST.

iii. Listar as possíveis palavras correspondentes.

Diferente de outros algoritmos (como o FASTA), o BLAST considera apenas as palavras de maior pontuação. As pontuações são estabelecidas por comparação das palavras listadas na etapa *ii* com todas as outras palavras de 3 letras. Uma matriz de substituição (BLOSUM62) é usada para pontuar as comparações entre pares de resíduos. Existem 20^3 possíveis pontuações de correspondência considerando uma palavra de 3 letras. Como exemplo, a comparação das palavras PQG e PEG tem pontuação de 15, enquanto a comparação de PQG com PQA pontua como 12. A seguir, um limiar T para pontuação de palavras vizinhas é usado para reduzir o número de possíveis palavras correspondentes. As palavras cujas pontuações forem maiores que o limiar T serão mantidas na lista de possíveis correspondências, enquanto aquelas cujas pontuações forem menores serão descartadas. Considerando o exemplo anterior, se $T = 13$, PEG será mantida, enquanto PQA será abandonada.

iv. Organizar as palavras de alta pontuação.

As palavras remanescentes, com alta pontuação, são organizadas em uma árvore de busca. Isso permite que o programa compare as palavras com as sequências do banco de dados de maneira rápida.

v. Repetir os passos *iii* e *iv* para cada palavra de *k*-letras originadas da sequência de busca.

vi. Varrer as sequências do banco de dados em busca de correspondências com as palavras remanescentes.

O BLAST realiza uma varredura das sequências depositadas no banco de dados, buscando pelas palavras de alta pontuação (como PEG, no exemplo anterior). Se uma correspondência exata for encontrada, ela será empregada para nuclear um possível alinhamento sem lacunas (gaps) entre a sequência de busca e a depositada no banco de dados.

vii. Estender as correspondências exatas entre pares de segmentos de alta pontuação.

A versão original do BLAST estende o alinhamento para a esquerda e para a direita de onde ocorre uma correspondência exata. A extensão é parada apenas quando a pontuação acumulada pelo HSP começa a diminuir (um exemplo pode ser visto na Figura 11-3).

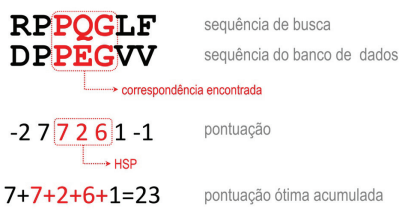


Figura 11-3: Exemplo do esquema de pontuação empregado pelo BLAST.

Para acelerar o processo, a versão atual do BLAST (BLAST2 ou Gapped BLAST) emprega um limiar mais baixo para a vizinhança das palavras, mantendo a sensibilidade na detecção de similaridade de sequências. Assim, a lista de possíveis correspondências obtidas na etapa *iii* é maior. Como observado na Figura 12-3, as regiões de correspondência exata com distância menor que A na mesma diagonal serão unidas como uma nova região, mais extensa. Posteriormente, essas regiões são estendidas da mesma maneira como ocorre no BLAST original, com os HSPs sendo pontuados com base em uma matriz de substituição.

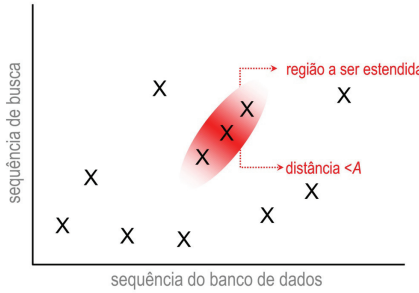


Figura 12-3: Esquema da extensão de zonas de correspondência entre sequências identificadas pelo BLAST.

viii. Listar todos os HSPs do banco de dados cuja pontuação seja alta o suficiente.

Nessa etapa são listados todos os pares de segmentos cuja pontuação seja maior que um determinado ponto de corte S. A distribuição de pontuações obtidas por alinhamento de sequências aleatórias é a base para determinação desse ponto de corte.

ix. Avaliar a significância da pontuação dos HSPs.

A avaliação estatística de cada par de segmentos de alta pontuação explora a Distribuição de Valores Extremos de Gumbel. O valor de confiança estatística e apresentado pelo BLAST, chamado de valor de expectativa, reflete o número de vezes que uma sequência não relacionada presente no banco de dados pode obter, ao acaso, um valor maior que S (ponto de corte). Ou seja, o e reflete o número de falsos positivos entre os resultados de similaridade encontrados. Para $p < 0,1$, o valor e se aproxima da distribuição de Poisson (ver item 4.8).

x. Transformar duas ou mais regiões de HSP em um alinhamento maior.

Em alguns casos, duas ou mais regiões de HSP podem ser combinadas em um trecho maior de alinhamento (uma evidência adicional da relação entre a sequência de busca e a encontrada no banco de dados). Existem dois métodos para comparar a significância das novas regiões ligadas. Se, por exemplo, forem encontradas duas regiões de HSP combinadas com pares de pontuação (67 e 41) e (53 e 45), cada método se comportará de maneira diferente. O método de Poisson conferirá maior significância ao conjunto com valor mínimo maior (45 em vez de 41). O método de soma dos pontos, ao contrário, dará preferência ao primeiro conjunto, pois 108 (67+41) é maior que 98 (53+45). O BLAST original usa o primeiro método, enquanto o BLAST2 emprega o segundo.

xi. Exibir os alinhamentos locais entre a sequência de busca e cada uma das correspondências no banco de dados.

O BLAST original produz apenas alinhamentos sem lacunas (gaps), incluindo cada um dos HSPs encontrados inicialmente, mesmo que mais de uma região de correspondência seja encontrada numa mesma sequência do banco de dados. O BLAST2 produz um único alinhamento com lacunas, podendo incluir todas as regiões de HSP encontradas. É importante destacar que o cálculo da pontuação e do valor e leva em conta as penalidades por abertura de lacunas no alinhamento.

xii. Registrar as correspondências encontradas.

Quando o valor e dos alinhamentos encontrados entre a sequência de busca e as do banco de dados satisfazem o ponto de corte estabelecido pelo usuário, a correspondência é registrada. Os resultados da busca são apresentados de forma gráfica, seguidos por uma lista de correspondências organizada pela pontuação e pelo valor e, e finalizam com os alinhamentos. A Figura 13-3 traz um exemplo de resultado obtido pelo BLAST.

Diferentes tipos de BLAST

O BLAST constitui uma família de programas, que podem ser usados para diferentes fins, dependendo das necessidades do usuário. Esses programas variam quanto ao tipo de sequência de busca, o banco de dados a ser empregado, e o tipo de comparação a ser realizada. As diferentes aplicações disponíveis pelo BLAST incluem:

i. ***blastn***: BLAST nucleotídeo-nucleotídeo. Usando uma sequência de DNA como entrada, dá como resultado as sequências de DNA mais similares presentes no banco de dados especificado pelo usuário.

ii. ***blastp***: BLAST proteína-proteína. Usando uma sequência proteica como entrada,

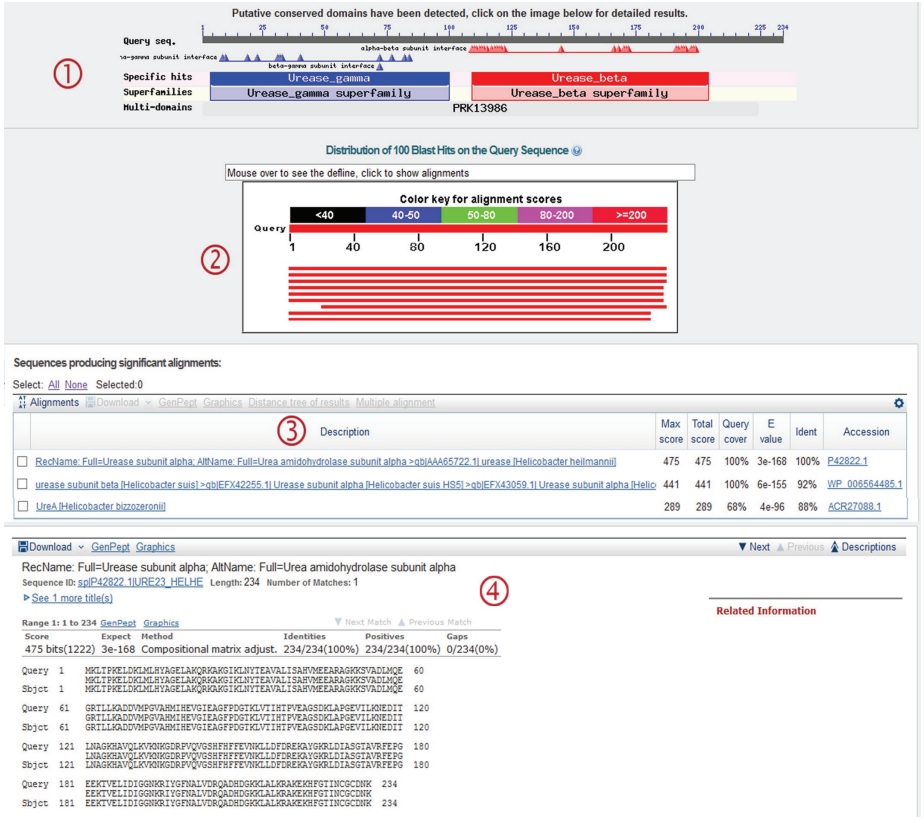


Figura 13-3: Exemplo de um resultado de busca realizada pelo BLAST. Diferentes informações são apresentadas: *1*) representação gráfica de domínios conservados identificados na sequência; *2*) representação gráfica de matches, indicando qualidade do alinhamento e cobertura das sequências identificadas; *3*) informações estatísticas dos resultados encontrados, incluindo identidade e valor ***e***; *4*) alinhamento de cada sequência encontrada com a sequência de busca (*query*).

dá como resultado as sequências proteicas mais similares presentes no banco de dados especificado pelo usuário.

iii. ***blastpgp***: BLAST iterativo com especificidade de posição (PSI-BLAST). Usado para encontrar proteínas distantemente relacionadas. Nesse caso, uma lista de proteínas proximamente relacionadas é criada. Essa lista serve de base para a criação de uma sequência média, que resume as características importantes do conjunto de sequências. A sequência média é usada para buscar sequências similares no banco de dados e um grupo maior de proteínas é encontrado. O grupo maior é usado na construção de uma nova sequência média e o processo é repetido. Ao

incluir proteínas relacionadas na busca, o PSI-BLAST é muito mais sensível na percepção de relações evolutivas distantes que o BLAST proteína-proteína tradicional.

iv. ***blastx***: tradução de nucleotídeos em 6 quadros-proteína. Compara os produtos de tradução conceitual nos 6 quadros de leitura de uma sequência de nucleotídeos contra o banco de dados de sequências proteicas.

v. ***tblastx***: tradução de nucleotídeos em 6 quadros-tradução de nucleotídeos em 6 quadros. O mais lento dos programas BLAST, tem por objetivo encontrar relações distantes entre sequências de nucleotídeos. Ele traduz a sequência de nucleotídeo nos 6 possíveis quadro de leitura e compara os resultados contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco de dados.

vi. ***tblastn***: proteína-tradução de nucleotídeos em 6 quadros. Compara uma sequência de proteína contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco de dados.

vii. ***megablast***: para empregar um grande número de sequências de busca. Quando se compara um grande número de sequências de busca (especialmente no BLAST por linha de comando), o megablast é muito mais rápido que o BLAST executado por várias vezes seguidas. Ele agrupa muitas sequências de busca, formando uma grande sequência, antes de realizar a busca no banco de dados. Os resultados são pós-analisados em busca de alinhamentos individuais.

3.8. Significância estatística

Em determinados casos, especialmente para buscar evidência de homologia entre sequências, o alinhamento é analisado sob o ponto de vista estatístico. Nessa óptica, podemos calcular quão bom pode ser um alinhamento simplesmente levando em consideração as razões de chance de alinhamento entre nucleotídeos quaisquer. Para isso, sequências de nucleotídeos ou aminoácidos são geradas aleatoriamente, alinhadas em conjunto e avaliadas, segundo um determinado esquema de pontuação. Para alinhamentos globais, pouco se sabe a respeito destas distribuições randômicas. No entanto, felizmente, estas técnicas são bem entendidas para casos de alinhamentos locais e, atualmente, são amplamente utilizadas para a avaliação de similaridade, especialmente em bancos de dados que comportam grande quantidade de sequências.

Para analisar a probabilidade associada a determinado alinhamento é necessário, inicialmente, gerar um modelo randômico das sequências em análise. Esses novos alinhamentos serão pontuados seguindo um determinado esquema de pontuação. Neste contexto, será calculada a probabilidade de se obter randomicamente uma pontuação pelo menos igual à pontuação do alinhamento original. O valor associado aos múltiplos testes realizados é chamado de valor ***e*** (*e*-value). Para banco de dados, este valor corresponde ao número de distintos alinhamentos, com uma pontuação igual ou melhor, que são esperados ocorrer na busca por sequências similares simplesmente por razões de chance (aleatórios). Estes cálculos estatísticos levam em consideração a pontuação do alinhamento e o tamanho do banco de dados. Quanto menor o valor ***e***, menor o número de chances de uma determinada sequência ser alinhada aleatoriamente com outras e, portanto, mais significante é o resultado. Por exemplo, um valor ***e*** de 1e-3 ($1x10^{-3}$ ou 0,001) significa que há a chance de 0,001 de que a sequência-alvo seja alinhada com uma sequência aleatória do banco de dados. Por exemplo, em um banco de dados que contém 10.000 sequências, neste caso, esperaríamos encontrar até 10 outras sequências que alinharão significativamente com a sequência-alvo. É importante ressaltar que o fato de

encontrarmos um valor ***e*** próximo de zero na comparação entre duas sequências não necessariamente denota a homologia destas sequências, dado que sequências não relacionadas podem conter similaridades devido à evolução convergente.

3.9. Alinhamento de 2 estruturas

O alinhamento de estruturas é um problema matematicamente complexo que só pode ser resolvido por algoritmos heurísticos. A Figura 14-3 apresenta um exemplo de alinhamento estrutural simples. Diferentes algoritmos oferecem resultados diferentes para o alinhamento, e algumas vezes essas diferenças são grandes. Por esse motivo é importante testar diferentes programas de alinhamento estrutural. Cada um deles tem pontos fortes e fracos, que podem ser explorados a partir da leitura dos artigos que os propuseram originalmente.

Existem três etapas essenciais para as diferentes estratégias de alinhamento estrutural: a representação, a otimização e a pontuação. A representação se refere às maneiras de representar as estruturas de uma forma que não seja dependente de coordenadas espaciais e que seja adequada ao alinhamento. A otimização lida com a amostragem do espaço de possíveis soluções para o alinhamento entre as estruturas. A pontuação lida com a classificação dos resultados obtidos e com sua significância estatística.

DALI: emprega matrizes de distâncias para representar as estruturas, transformando as estruturas 3D em conjuntos 2D de distâncias entre Cα. Se imaginarmos a sobreposição das matrizes, as regiões de sobreposição na diagonal representam similaridades na estrutura secundária (similaridades no esqueleto polipeptídico), e similaridades fora da diagonal representam similaridades na estrutura terciária. As matrizes são então divididas em matrizes menores, de tamanho fixo, com base nas similaridades encontradas. Cada submatriz é unida a outras que sejam adjacentes para obter a matriz de sobreposição com maior abrangência. A significância estatística do alinhamento é calculada com base na distribuição

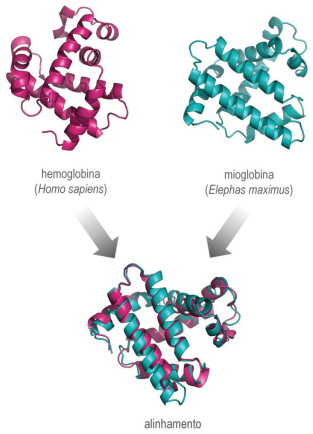


Figura 14-3: Exemplo de alinhamento de duas estruturas proteicas, oriundas de diferentes organismos: hemoglobina humana e mioglobina de elefante-asiático.

encontrada em uma comparação de centenas de estruturas de baixa identidade. A pontuação é apresentada como número de desvios-padrão em relação a tal distribuição.

SSAP: cria vetores ligando resíduos a partir dos Cβ, representando a estrutura em duas dimensões, considerando posição e direção. Um algoritmo de programação dinâmica identifica similaridades entre as matrizes de vetores, gerando uma nova matriz que é posteriormente recalculada considerando as diferenças entre cada posição de similaridade encontrada na primeira etapa em relação às outras posições de similaridade, até que uma matriz ótima seja atingida. A pontuação do SSAP não é estatística, mas foi calibrada em relação ao banco de dados CATH. Assim, uma pontuação maior que 70 indica similaridade entre as estruturas comparadas.

VAST: cria vetores a partir de elementos de estrutura secundária cujo tipo, direção e conexão estão relacionados com a topologia da proteína. Esses elementos (fragmentos) de estrutura secundária são alinhados e comparados com alinhamentos gerados aleatoriamente. Alinhamentos com boa pontuação são agrupados e depois realinhados usando um procedimento de otimização por Monte Carlo. A significância estatística é dada pelo valor ***p*** (assim

como ocorre no BLAST). O valor ***p*** é proporcional à probabilidade de se obter o alinhamento ao acaso.

SARF2: transforma as coordenadas em um conjunto de elementos de estrutura secundária. Posteriormente, avalia pares desses elementos comparando o ângulo entre eles, a menor distância entre seus eixos e as distâncias mínimas e máximas entre cada elemento e a linha média. Um otimizador baseado em grafos é empregado para obter o maior número de conjuntos mutuamente compatíveis, e então o alinhamento final é calculado por adição de mais resíduos até que um valor mínimo de RMSD, definido pelo usuário, seja atingido. A pontuação final do alinhamento é calculada como função do RMSD e do número de Cα pareados entre as estruturas. A significância estatística é obtida por comparação à distribuição de pontuações obtidas pelo alinhamento da proteína leghemoglobina a centenas de estruturas não redundantes.

CE: representa as proteínas como conjuntos de distâncias entre Cα de oito resíduos consecutivos na estrutura. Primeiramente, são identificados todos os pares de octâmeros compatíveis entre as estruturas. Posteriormente, um algoritmo de extensão combinatória identifica e combina os pares mais similares entre as estruturas, adicionando mais pares a cada etapa do cálculo até a obtenção do melhor alinhamento. A significância estatística é dada por comparação às pontuações obtidas em um conjunto de alinhamentos entre estruturas com menos de 25% de identidade de sequência.

MAMMOTH: transforma as coordenadas da proteína em um conjunto de vetores unitários a partir dos Cα de heptâmeros consecutivos. A similaridade entre heptâmeros é calculada pela sobreposição de seus vetores, a matriz de similaridade ótima é identificada e então o melhor alinhamento local entre estruturas é identificado dentro de um valor de RMSD pré-definido. A significância estatística é dada pelo valor ***p***, baseado na comparação com a pontuação de alinhamentos obtidos aleatoriamente.

SALIGN: representa as proteínas por um conjunto de propriedades ou características calculadas a partir da sequência e da estrutura ou definidas arbitrariamente pelo usuário. Tais propriedades incluem tipo de resíduo, distância entre resíduos, acessibilidade da cadeia lateral, estrutura secundária, conformação local da estrutura e característica a ser definida pelo usuário. O programa calcula uma matriz de dissimilaridade entre propriedades equivalentes, e a pontuação da dissimilaridade é calculada pela soma das matrizes de cada característica. A melhor sobreposição de matrizes é obtida por um algoritmo baseado em programação dinâmica. A significância estatística não é calculada pelo SALIGN e o usuário obtém apenas os valores da pontuação de dissimilaridade. O programa fornece, entretanto, um valor adicional de qualidade, apresentado como porcentagem de Cα cuja distância é menor que 3,5 Å entre os pares de estruturas alinhadas.

3.10. Alinhamento de >2 estruturas

A maior parte dos métodos disponíveis para o alinhamento múltiplo de estruturas inicia-se estabelecendo todos os alinhamentos entre pares de estruturas e, então, emprega-os para estabelecer um alinhamento consenso entre todas as estruturas. A Figura 15-3 apresenta um exemplo de alinhamento estrutural múltiplo. Os métodos para obter o alinhamento consenso variam entre os programas de alinhamento. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de estruturas múltiplo.

CE-MC: realiza o refinamento de um conjunto de alinhamentos de pares de estruturas empregando uma técnica de otimização de Monte Carlo. O algoritmo modifica o alinhamento múltiplo aleatoriamente, e as modificações são aceitas se houver melhoria na pontuação do alinhamento. O processo encerra quando o alinhamento múltiplo não puder mais ser melhorado por modificações aleatórias.

MAMMOTH-Mult: essa extensão do MAMMOTH gera inicialmente todos os alinhamentos de estruturas aos pares. Um procedimento de organização por médias é empregado para agrupar as estruturas com base em suas similaridades aos pares, gerando uma árvore. O alinhamento múltiplo é gerado por reorganização dessa árvore, onde ramos similares vão sendo agrupados aos pares, iterativamente.

SALIGN: pode realizar alinhamentos múltiplos de duas maneiras, baseado em uma árvore ou por alinhamento progressivo. O primeiro caso é muito similar ao MAMMOTH-Mult. No alinhamento progressivo, as estruturas são alinhadas na ordem em que são fornecidas para o programa. A vantagem

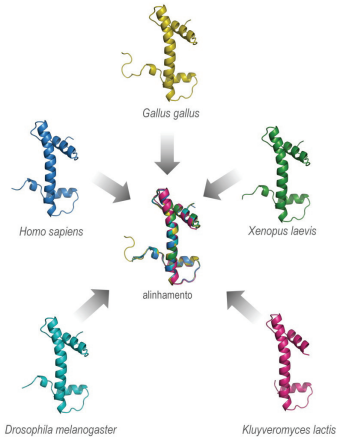


Figura 15-3: Exemplo de alinhamento de múltiplas estruturas proteicas, oriundas de diferentes organismos (histonas H3 de levedura, mosca-da-fruta, homem, frango, sapo-de-garras).

desse método é o de seu custo computacional ser menor que o do método baseado em uma árvore.

3.11. Alinhamento flexível

O alinhamento de estruturas considerando sua flexibilidade está se tornando cada vez mais importante devido à melhor compreensão do enovelamento proteico. Cada vez mais, percebe-se que não existem enovelamentos estanques, mas sim um gradiente densamente populado por variantes conformacionais. Desta forma, torna-se mais difícil definir domínios proteicos, sendo mais adequado descrever as estruturas como conjuntos de estruturas supra-secundárias. Com base nessa proposta, a diferença entre proteínas relacionadas reside na orientação relativa desses subdomínios. A Figura 16-3 demonstra as diferenças que podem ser observadas ao alinhar um par de estruturas de maneira rígida ou flexível.

FATCAT: o algoritmo adiciona "torções" entre pares

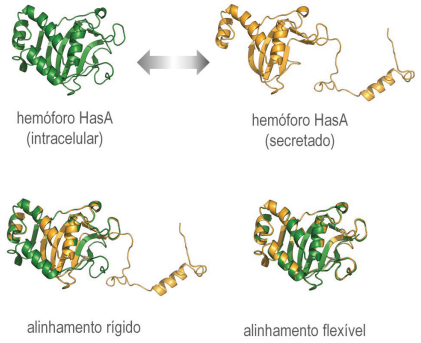


Figura 16-3: Comparação entre alinhamento estrutural rígido e flexível. A estrutura da proteína HasA (um captador bacteriano de grupamentos heme) foi obtida para suas formas intra- e extra-celular. Observe que o alinhamento rígido identifica similaridade parcial entre as estruturas, enquanto o alinhamento flexível detecta o rearranjo espacial de parte da proteína, evidenciando sua identidade.

de fragmentos proteicos alinhados, que são tratados como corpos rígidos. De maneira geral, o programa permite a inclusão dessas torções quando elas diminuem o valor final do RMSD, refletindo em um melhor alinhamento estrutural. O alinhamento final é obtido por programação dinâmica e se baseia na matriz de similaridade entre os fragmentos pareados, obtidos na primeira etapa do cálculo.

FLEXPROT: mantém uma das proteínas rígida, enquanto a outra pode sofrer alterações em busca de maior similaridade estrutural. As regiões potencialmente flexíveis da proteína são detectadas automaticamente e empregadas nas alterações conformacionais.

ALADYN: alinha pares de estruturas com base em sua dinâmica interna e similaridade entre seus movimentos de grande escala. O posicionamento ótimo entre as proteínas é encontrado ao maximizar as similaridades entre os padrões de flutuação estrutural, que são calculados pelo modelo de redes elásticas.

POSA: uma variante do FATCAT para o alinhamento múltiplo flexível de estruturas. Emprega uma metodologia combinada, introduzindo grafos de ordem

parcial para visualizar e agrupar regiões similares entre as estruturas.

3.12. Conceitos-chave

Algoritmo: sequência lógica de instruções necessárias para executar uma tarefa.

Alinhamento: método de organização de sequências ou estruturas biológicas para evidenciar regiões similares e dissimilares. Estes métodos estão geralmente atrelados a inferências funcionais ou evolutivas.

Alinhamento Múltiplo: alinhamento que envolve mais de duas sequências ou estruturas

Alinhamento Simples: alinhamento que envolve apenas duas sequências ou estruturas.

BLAST: *Basic Local Alignment Search Tool* (Ferramenta de Busca por Alinhamento Local Básico), empregado para buscar sequências em bancos de dados com base em sua similaridade.

Homologia: é um termo essencialmente qualitativo que denota uma ancestralidade comum de determinada sequência.

HSP: pares de segmentos de alta pontuação (*high-scoring segment pairs*), zonas de similaridade entre sequências identificadas pelo BLAST.

Identidade: Porcentagem de caracteres similares entre duas sequências (excluindo-se as lacunas).

Indels: identifica inserções e deleções de caracteres ao longo do processo evolutivo.

Lacunas: regiões identificadas por hífens que representam a inserção/deleção de caracteres ao longo do processo evolutivo.

Matches: regiões que apresentam caracteres idênticos entre diferentes sequências.

Mismatches: regiões que apresentam caracteres não idênticos entre diferentes sequências.

Penalidades por lacuna (PL): conjunto de parâmetros necessários para atribuir a pontuação para uma lacuna em um sistema de alinhamento por pontuação.

RMSD: desvio médio quadrático.

Tradução: tradução (*in silico*) de uma sequência de mRNA em sua possível sequência proteica correspondente

3.13. Leitura recomendada

BOGUSKI, Mark S. A molecular biologist visits Jurassic Park. Biotechniques, 12, 668-669, 1992.

CARUGO, Oliviero. Recent progress in measuring structural similarity between proteins. Curr Protein Pept Sci, 8, 219-241, 2007.

MADDEN, Tom. The BLAST sequence analysis tool. In: McENTYRE, Jo; OSTELL, Jim (Eds.), The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information: 2002.

MARTI-RENOM, Marc A.; et al. Structure comparison and alignment. In: GU, Jenny; BOURNE, Philip E (Eds.), Structural Bioinformatics. 2.ed. Hoboken: John Wiley & Sons, 2009.

MAYR, Gabriele; DOMINGUES, Francisco S.; LACKNER, Peter. Comparative analysis of protein structure alignments. BMC Struct Biol, 7, 50, 2007.

MOUNT, David W. Bioinformatics: Sequence and Genome Analysis. 2.ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.

ROSSMANN, Michael G.; ARGOS, Patrick. The taxonomy of binding sites in proteins. Mol Cell Biochem, 21, 161-182, 1978.

10. CURRICULUM VITÆ

I. Formação acadêmica:

Mestrado em Biologia Celular e Molecular pela Universidade Federal do Rio Grande do Sul, de 2009 a 2010. *Comportamento conformacional da urease de 'Canavalia ensiformis'*. Orientadores: Prof. Célia Regina Carlini e Prof. Hugo Verli.

Graduação em Ciências Biológicas - Bacharelado com Ênfase Molecular, Celular e Funcional - pela Universidade Federal do Rio Grande do Sul, de 2005 a 2008. *Ureases: uma análise comparativa*. Orientadora: Prof. Célia Regina Carlini. Co-orientadora: Dra. Evelyn Koeche Schroeder.

II. Formação complementar: (restrito ao período de desenvolvimento da tese)

Computational Molecular Evolution (Curso de curta duração), Technical University of Denmark, Dr. Anders Gorm Pedersen, 2013.

Métodos de Docking Receptor-Ligante, (Curso de curta duração), Laboratório Nacional de Computação Científica, Dr. Laurent Dardenne, Dra. Camila Magalhães, 2010.

Simulação de Proteínas de Biomembranas, (Curso de curta duração), Laboratório Nacional de Computação Científica, Dr. Hubert Stassen, Dr. Rafael Bernardi.

III. Estágios realizados: (restrito ao período de desenvolvimento da tese)

Treinamento em Replica Exchange with Solute Tempering (REST2), sob supervisão do Dr. Francesco Musiani e coordenação do Dr. Stefano Ciurli, outubro de 2013. Laboratório de Química Bioinorgânica, Università di Bologna, Itália.

IV. Prêmios recebidos: (restrito ao período de desenvolvimento da tese)

3º colocado "Top-25 Hottest Articles - January to December 2012 full year, Toxicon", Ligabue-Braun, R.; Verli, H.; Carlini, C.R., Toxicon, 59 (2012) 680-695, SciVerseScience Direct - Elsevier.

2º colocado "Top-25 Hottest Articles - April to June 2012, Toxicon", Ligabue-Braun, R.; Verli, H.; Carlini, C.R., Toxicon, 59 (2012) 680-695, SciVerseScience Direct - Elsevier.

6º colocado "Top-25 Hottest Articles - July to September 2012, Toxicon", Ligabue-Braun, R.; Verli, H.; Carlini, C.R., Toxicon, 59 (2012) 680-695, SciVerseScience Direct - Elsevier.

10º colocado "Top-25 Hottest Articles - October to December 2012, Toxicon", Ligabue-Braun, R.; Verli, H.; Carlini, C.R., Toxicon, 59 (2012) 680-695, SciVerseScience Direct - Elsevier.

V. Trabalhos científicos apresentados em congressos:

a. Nacionais

Ligabue-Braun, R; Sachett, LG; Carlini, CR; Verli, H. Interaction of the major cat allergen (Fel d 1) with calcium. XLII Annual Meeting of SBBq, Foz do Iguaçu, 2013.

Ligabue-Braun, R; Carlini, CR; Verli, H. Structural characterization of the urease activation complex. XLI Annual Meeting of SBBq, Foz do Iguaçu, 2012.

Andreis, FC; Ligabue-Braun, R; Carlini, CR; Verli, H. Phylogenetic Characterization of Ureases. XLI Annual Meeting of SBBq, Foz do Iguaçu, 2012.

Ligabue-Braun, R; Andreis, FC; Carlini, CR; Verli. Phylogenetical, Structural and Conformational Characterization of Ureases. XL Annual meeting of SBBq, Foz do Iguaçu, 2011.

Ligabue-Braun, R; Schroeder, EK; Carlini, CR; Verli. Molecular dynamics of Jack bean (*Canavalia ensiformis*) urease. XXXIX Annual Meeting of SBBq, Foz do Iguaçu, 2010.

Lima, TMF; Feder, V; Olivera-Severo, D; Pinto, PM; Becker-Ritt, AB; Ligabue-Braun, R; Vainstein, MH; Carlini, CR. Characterization and structural

aspects from purified urease of *Cryptococcus gattii* R265. XXXIX Annual Meeting of SBBq, Foz do Iguaçu, 2010.

Ligabue-Braun, R; Schroeder, EK; Carlini, CR. Modeled structures and molecular dynamics of plant ureases. XXXVIII Annual Meeting of SBBq, Águas de Lindóia, 2009.

b. Internacionais

Schroeder, EK; Ligabue-Braun, R; Carlini, CR. 'In silico' comparative studies of *Canavalia ensiformis* and *Glycine max* ureases. XVI World Congress of the International Society on Toxinology and X Congresso da Sociedade Brasileira de Toxinologia, Recife, 2009.

VI. Publicações em periódicos especializados:

a. Internacionais

Ligabue-Braun, R; Andreis, FC; Verli, H; Carlini, CR. 3-to-1: unraveling structural transitions in ureases. *Naturwissenschaften*, v. 100, p. 459-467, **2013**.

Ligabue-Braun, R; Real-Guerra, R; Carlini, CR; Verli, H. Evidence-based docking of the urease activation complex. *Journal of Biomolecular Structure & Dynamics*, v. 31, p. 854-861, **2013**.

Martinelli, AHS; Kappaun, K; Ligabue-Braun, R; Defferrari, MS; Piovesan, AR; Stanisçuaski, F; Demartini, DR; Dal Belo, CA; Almeida, CGM; Follmer, C; Verli, H; Carlini, CR; Pasquali, G. Structure-function studies on Jaburetox, a recombinant insecticidal and antifungal peptide derived from jack bean (*Canavalia ensiformis*) urease. *Biochimica et Biophysica Acta. G, General Subjects*, no prelo (doi:pii:S0304-4165(13)00493-5. 10.1016/j.bbagen.2013.11.010) , **2013**.

Ligabue-Braun, R; Verli, H; Carlini, CR. Venomous mammals: A review. *Toxicon*, v. 59, p. 680-695, **2012**.

Postal, M; Martinelli, AH; Becker-Ritt, AB; Ligabue-Braun, R; Demartini, DR; Ribeiro, SF; Pasquali, G; Gomes, VM; Carlini, CR. Antifungal properties of *Canavalia ensiformis* urease and derived peptides. *Peptides*, v. 38, p. 22-32, **2012**.

Salvadori, JDM; Defferrari, MS; Ligabue-Braun, R; Lau, EY; Salvadori, JR; Carlini, CR. Characterization of entomopathogenic nematodes and symbiotic bacteria active against *Spodoptera frugiperda* (Lepidoptera: Noctuidae) and contribution of bacterial urease to the insecticidal effect. *Biological Control*, v. 63, p. 253-263, **2012**.

Mulinari, F; Becker-Ritt, AB; Demartini, DR; Ligabue-Braun, R; Stanisçuaski, F; Verli, H; Fragoso, RR; Schroeder, EK; Carlini, CR; Grossi-de-Sá, MF. Characterization of JBURE-IIb isoform of *Canavalia ensiformis* (L.) DC urease. *Biochimica et Biophysica Acta. Proteins and Proteomics*, v. 1814, p. 1758-1768, **2011**.

VII. Orientações de Iniciação Científica:

Fábio Carrer Andreis. Caracterização filogenética de ureases. Centro de Biotecnologia, UFRGS, de agosto de 2010 a julho de 2013.

VIII. Representação de classe:

Representação discente junto ao Conselho Científico do Centro de Biotecnologia da UFRGS, maio de 2010 a abril de 2012.

Representação discente junto à Comissão de Pós-Graduação do Programa de Pós-Graduação em Biologia Celular e Molecular, UFRGS, maio de 2012 a março de 2014.

IX. Educação e popularização de ciência e tecnologia:

Co-organização, Curso de Férias PPGBCM, "Você conhece a célula?", público-alvo: professores de ciências e alunos de ensino médio, 2013.

Co-organização, Curso de Férias PPGBCM, "Plantas... Para que mesmo?", público-alvo: professores de ciências e alunos de ensino médio, 2011.

Co-organização, Curso de Férias PPGBCM, "Plantas... Como funcionam?", público-alvo: professores de ciências e alunos de ensino médio, 2011.

Co-organização, Curso de Férias PPGBCM, "Você conhece a célula?", público-alvo: professores de ciências e alunos de ensino médio, 2010.

IX. Bolsas recebidas:

Bolsa CNPq de doutorado (Edital MCT/CNPq nº 70/2009), Programa de Expansão da Pós-Graduação em Áreas Estratégicas), Programa de Pós-Graduação em Biologia Celular e Molecular pelo Centro de Biotecnologia da UFRGS, julho de 2010 a março de 2014.

Bolsa PROPG-UFRGS de missão científica de curta duração no exterior para estudantes dos Programas de Pós-Graduação da UFRGS, outubro de 2013.

Bolsa CAPES de mestrado, Programa de Pós-Graduação em Biologia Celular e Molecular pelo Centro de Biotecnologia da UFRGS, março de 2009 a junho de 2010.

Bolsa PIBITI (CNPq-UFRGS) de iniciação tecnológica, Centro de Biotecnologia da UFRGS, agosto de 2008 a dezembro 2008.