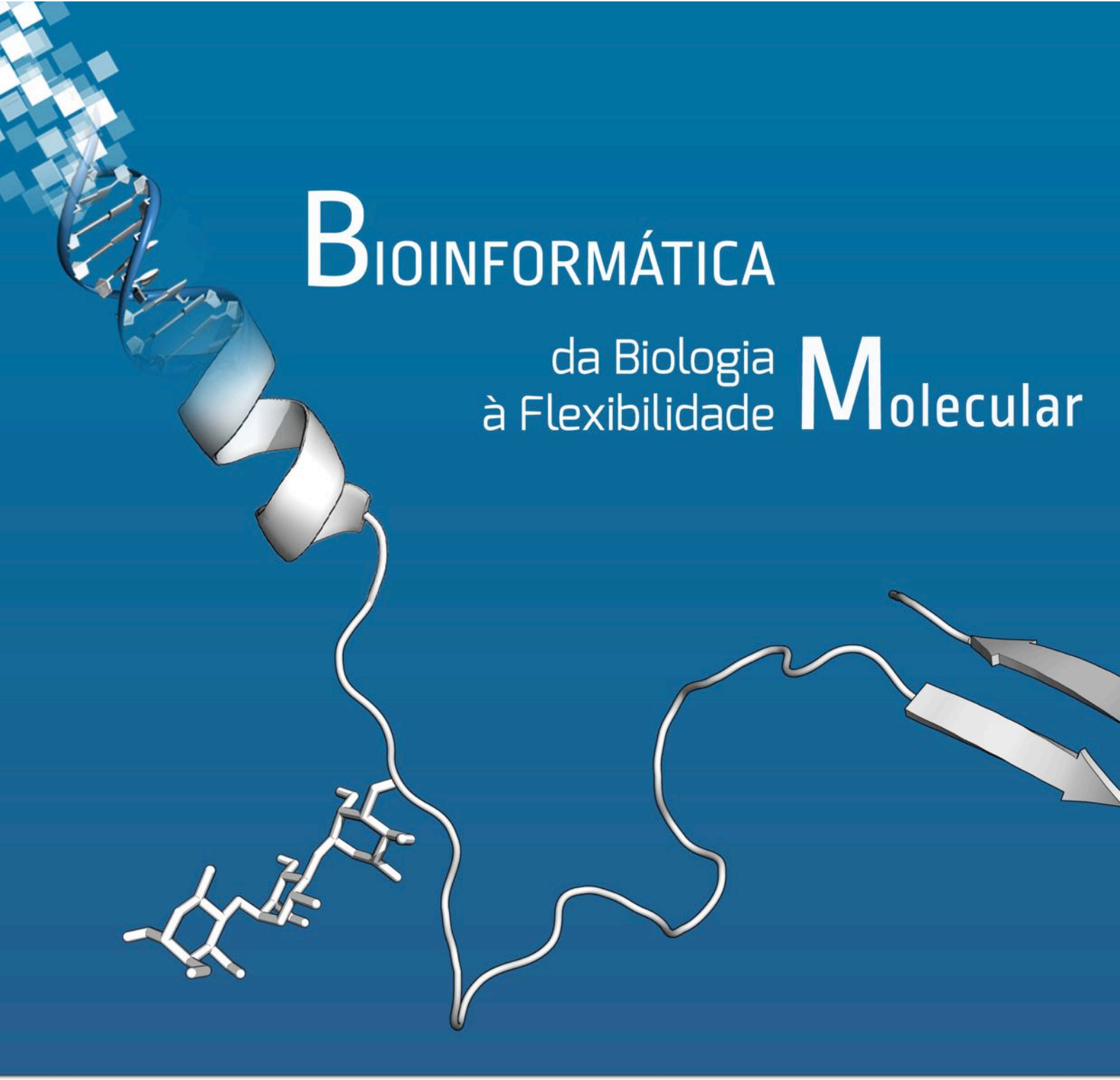

Bioinformática
da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

ISBN 978-85-69288-00-8
9 788569 288008

Sociedade Brasileira de Bioquímica e Biologia Molecular — SBBq

Apoio:
CNPq
CAPES
FAPERGS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

CDU 575.112
ISBN 978-85-69288-00-8

1. O que é Bioinformática?

"O todo sem a parte não é todo,
A parte sem o todo não é parte,
Mas se a parte o faz todo, sendo parte,
Não se diga, que é parte, sendo todo."

Hugo Verli

Gregório de Matos Guerra (1636-1696)

1.1. Introdução

Gregório de Matos, poeta brasileiro que viveu no século XVII, há quase 400 anos apresentou, na frase de epígrafe deste capítulo, seu entendimento sobre a indissociabilidade das partes para compreensão do todo. No nosso caso, o todo é a bioinformática. As partes, contudo, não são tão óbvias quanto se possa imaginar em um primeiro momento. Tampouco há consenso sobre estas. Assim, nossa discussão sobre o que é bioinformática não pretende estabelecer definições rígidas, mas guias para que o leitor entenda o quão complexa e dinâmica é esta jovem ciência.

Esta complexidade usualmente nos passa despercebida. Por exemplo, quando pensamos no impacto do projeto genoma humano, uma das principais implicações é a melhoria dos processos terapêuticos acessíveis à população. Mas a identificação de um novo gene ou mutação em um gene conhecido, por mais que seja associado a um processo patológico, está a uma grande distância de um novo fármaco. A partir da sequência, o paradigma mais moderno para desenvolvimento de novos fármacos passa pela caracterização da estrutura tridimensional da proteína codificada. Esta estrutura é então empregada para guiar o planejamento racional de novos compostos, como se um chaveiro construísse uma chave (o fármaco) a partir da fechadura. Por mais que a analogia seja simples, ainda serve como base para algumas das mais frequentes estratégias de planejamento de fármacos. E, embora a ideia de que este processo é flexível, e não rígido (mais como uma mão encaixando em uma luva, sendo a mão o fármaco e a luva o receptor) date da década de 1960, são processos tão complexos que demoramos em torno de 15 anos para lançar um novo fármaco no mercado (e este tempo não está diminuindo).

Assim, ao invés de procurar definições restritivas, este livro se propõe a empregar definições amplas, que sirvam de suporte para um entendimendo da grande gama de potencialidades e aplicações da bioinformática, buscando suportar inclusive futuras aplicações da metodologia, ainda em desenvolvimento ou por serem desenvolvidas.

Ao mesmo tempo que sequências codificantes geram seus efeitos biológicos como estruturas tridimensionais, o estudo destas pode e muito se beneficiar do estudo de sequências de proteínas relacionadas (por exemplo, alças flexíveis tendem a apresentar uma elevada variabilidade filogenética). Mesmo o estudo de sequências não codificantes pode se beneficiar do conhecimento de estruturas tridimensionais, visto que a regulação de sua expressão é realizada por fatores de transcrição proteicos. Assim, há uma retroalimentação entre as informações originadas em sequências biológicas e em suas respectivas estruturas 3D.

Em linhas gerais, este livro parte do entendimento de que a bioinformática se refere

ao emprego de ferramentas computacionais no estudo de problemas e questões biológicas, abrangendo também as aplicações relacionadas à saúde humana como o planejamento de novos fármacos.

Neste caminho, da sequência de nucleotídeos até estruturas proteicas, alcançando por fim fármacos, diversas áreas do conhecimento estão envolvidas. Biologia molecular, biologia celular, bioquímica, química, física e computação são talvez as principais grandes áreas do saber envolvidas nesse processo, cada uma contribuindo com diversas especialidades.

1.2. Origens

O que apresentaremos neste livro como bioinformática pode ser separado em duas grandes vertentes:

i) a bioinformática tradicional, ou clássica (pela primazia do nome bioinformática), que aborda principalmente problemas relacionados a sequências de nucleotídeos e aminoácidos, e

ii) a bioinformática estrutural, que aborda questões biológicas de um ponto de vista tridimensional, abrangendo a maior parte das técnicas compreendidas pela química computacional ou modelagem molecular.

Podemos traçar como momento chave para ambas as vertentes da bioinformática o início da década de 1950, quando a revista *Nature* publicou o trabalho clássico sobre a estrutura em hélice da molécula de DNA por James Watson e Francis Crick (Figura 1-1). Neste momento, as bases moleculares para o entendimento estrutural da replicação e tradução do material genético foram apresentadas, permitindo-nos entender como aquela "sequência de letras" (as bases do DNA) se organizam tridimensionalmente.

Este trabalho, contudo, deve ser visto como parte de um momento histórico, composto por diversas contribuições fundamentais para o nosso entendimento de moléculas biológicas e suas funções. Dentre estas destacam-se os trabalhos de Linus Pauling e Robert Corey, no início da década de 1950, e de Gopalasamudram N. Ramachandran, no início da década de 1960, que ofereceram as bases para a compreensão da estrutura tridimensional de proteínas.

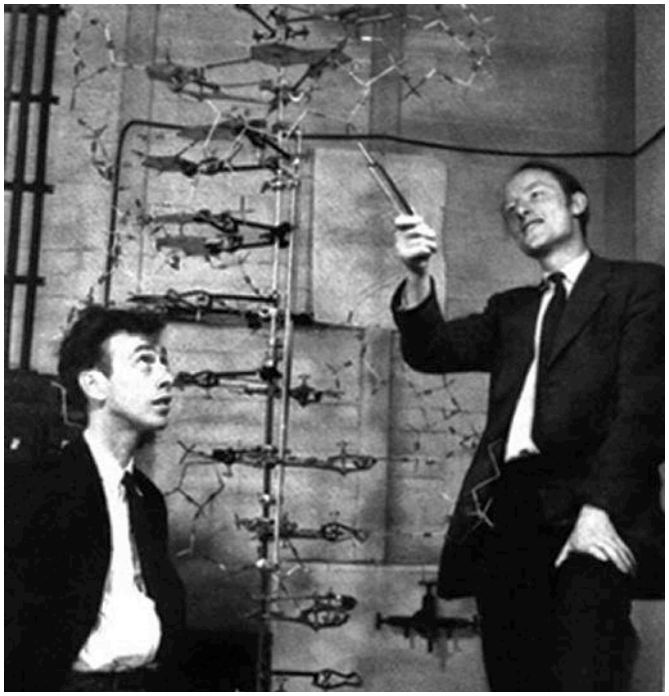

Figura 1-1: Watson e Crick em frente a um modelo da hélice de DNA. Cavendish Laboratory, Universidade de Cambridge, 1953, reproduzida sob licença.

Desde estes trabalhos até a primeira vez em que se relatou o uso de programas de computadores para visualizar estruturas tridimensionais de moléculas passaram-se mais de 10 anos quando, em 1966, Cyrus Levinthal publica na revista *Scientific American* o trabalho desenvolvido no *Massachusetts Institute of Technology* por John Ward e Robert Stotz.

Ainda nesta década se dá o primeiro esforço de sistematização do conhecimento acerca da estrutura tridimensional dos efetores da informação genética, as proteínas, em 1965, com o *Atlas of Protein Sequence and Structure*, organizado por diversos autores, dentre os quais destacaremos Margaret Dayhoff.

Este destaque se deve ao fato do papel-chave exercido pela Dra. Dayhoff na formação das raízes do que entendemos hoje por

bioinformática, tanto em sua faceta voltada para sequências quanto para estruturas. Foi uma das pioneiras no uso de computadores para o estudo de biomoléculas, incluindo tanto ácidos nucleicos quanto proteínas. Por exemplo, é ela que inicia o uso da representação de uma única letra para descrever cada aminoácido (Tabela 1-1), ao invés das usuais três letras, em uma época em que os dados eram armazenados em cartões perfurados (Figura 2-1). Desenvolveu as primeiras matrizes de substituição e fez importantes contribuições no desenvolvimento dos estudos filogenéticos. Também teve participação importante no desenvolvimento de métodos para o estudo de moléculas por cristalografia de raios-X (como veremos no capítulo 13).

Com o desenvolvimento de computadores mais poderosos e com o avanço no entendimento dos determinantes da estrutura e da dinâmica proteica, tornam-se possíveis os primeiros estudos acerca da dinâmica e do enovelamento de proteínas por simulações de dinâmica molecular por Michael Levitt e Arieh Warshel, nos anos de 1970, estudos estes agraciados com o prêmio Nobel de Química em 2013 (Figura 3-1).

A partir dos trabalhos destes e de outros pesquisadores, diversos avanços foram feitos progressivamente nos anos que se seguiram, tanto no entendimento de biomoléculas quanto no emprego de técnicas computacionais para retroalimentar este entendimento. Por exemplo, o aumento na obtenção de informações de alta qualidade sobre a estrutura 3D de biomoléculas vem servindo de suporte para o desenvolvimento de campos de força cada vez mais precisos, enquanto novas abordagens vêm possibilitando o alinhamento de sequências cada vez mais distantes evolutivamente.

Contudo talvez possamos afirmar que, a partir destas bases, os maiores impactos da área na ciência estejam se delineando neste exato período da história, em que dois importantes fatores se manifestam: o avanço (e barateamento) no poder computacional e os projetos genoma.

Computadores cada vez mais rápidos e mais baratos nos permitem abordar problemas, literalmente, inimagináveis há poucos anos. Os métodos e a dimensão dos problemas abordados por um aluno de iniciação científica serão, em sua maioria, totalmente obsoletos ao final de seu doutoramento (considerado o mesmo nível de impacto dos veículos de divulgação). A cada ano que passa podemos abordar problemas mais complexos, de forma mais completa, e mais pesquisadores com menos recursos podem trabalhar nestas áreas de pesquisa, o que torna a bioinformática uma das áreas do conhecimento mais acessíveis para pesquisadores em início de carreira.

Em contrapartida, esta situação acarreta na necessidade de atualização e renovação dos procedimentos computacionais constantemente para nos mantermos competitivos na comunidade científica da área. O trabalho

Tabela 1-1: Nomes dos 20 aminoácidos codificadores de proteínas junto a suas representações em 1 e 3 letras.

Aminoácido	Representação de 3 letras	Representação de 1 letra
Alanina	Ala	A
Cisteína	Cys	C
Ác. aspártico	Asp	D
Ác. glutâmico	Glu	E
Fenilalanina	Phe	F
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Lisina	Lys	K
Leucina	Leu	L
Metionina	Met	M
Asparagina	Asn	N
Prolina	Pro	P
Glutamina	Gln	Q
Arginina	Arg	R
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Triptofano	Trp	W
Tirosina	Tyr	Y

Figura 2-1: IBM 7090, computador que Margaret Dayhoff utilizou no início de seus trabalhos (NASA Ames Resarch Center, 1961).

que alguém tenha publicado com simulações por dinâmica molecular (capítulo 8) alguns anos atrás, com uma simulação de, digamos, 10 ns, hoje estaria totalmente desatualizado, exigindo no mínimo uma ordem de grandeza a mais (idealmente, com replicatas e/ou condições adicionais como controle). Como consequência, as conclusões obtidas em um trabalho não necessariamente se manteriam em um novo trabalho. Similarmente, uma árvore filogenética obtida a partir de um determinado alinhamento e matriz de pontuação há 20 anos poderia ser diferente hoje, com ferramentas mais robustas de alinhamento (como será visto no capítulo 3). Esta é uma situação bastante desafiadora, assim como uma grande oportunidade, para os futuros bioinformatas.

Mas esta situação por si não é suficiente para o aumento explosivo do emprego de estratégias computacionais no estudo de sistemas biológicos, o que é principalmente devido ao projeto Genoma Humano. A partir deste, e da popularização de outros projetos genoma (capítulo 4), criou-se um gigantesco e crescente volume de sequências de genes cujas relações evolutivas e funcionais precisam ser elucidadas, como ponto de partida para novos desenvolvimentos terapêuticos. Hoje, é possível identificar um novo candidato a receptor alvo de novos fármacos a partir de organismos muito distantes evolutivamente de nós, como leveduras, bactérias ou mesmo plantas.

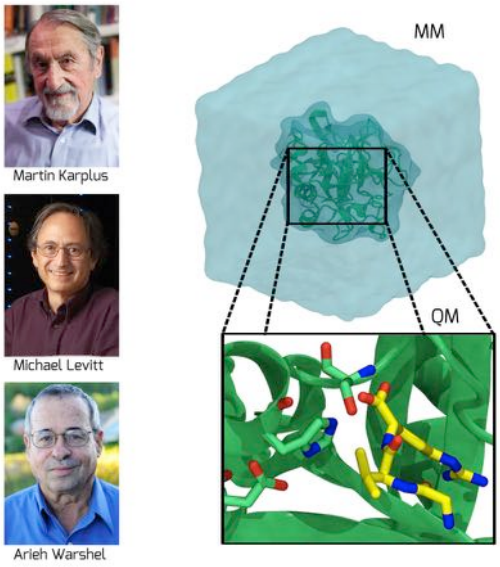

Figura 3-1: Agraciados pelo prêmio Nobel de química de 2013, os Professores Martin Karplus, Michael Levitt e Arieh Warshel.

O crescimento deste volume de informações ainda está longe de cessar. Estudos de transcriptoma, metaboloma ou glicoma ainda têm muito a agregar no nosso conhecimento do funcionamento de sistemas biológicos, potencializando tanto aplicações terapêuticas quanto biotecnológicas. Contudo, isto exigirá cada vez mais avanços da bioinformática, seja em *hardware*, *software* ou em estratégias de análise de dados e construção de modelos.

Um exemplo neste sentido envolve a gigantesca defasagem entre nossa capacidade de lidar com sequências e com estruturas 3D. Enquanto em um computador pessoal simples podemos realizar alinhamentos com algumas centenas de sequências sem maiores dificuldades, localmente ou na *web*, dependendo do método, e recebendo a resposta quase que imediatamente, para realizar uma simulação por dinâmica molecular de uma única proteína precisaríamos, neste mesmo computador, de alguns meses.

Um último aspecto importante nesta contextualização inicial da bioinformática, dentro da proposta apresentada por este livro, diz respeito à importância relativa das diferentes biomoléculas na manifestação da informação genética, mantendo a homeostasia e servindo como alvo de modulação far-

macológica ou emprego biotecnológico. Tradicionalmente, os ácidos nucleicos e as proteínas receberam a maior atenção enquanto alvos da bioinformática, os primeiros como repositórios da informação biológica e as últimas como efetores desta informação. Esta percepção, contudo, vem sendo progressivamente relativizada. Membranas e carboidratos, a despeito de não estarem codificados diretamente no genoma (não há um códon para um fosfolipídeo ou para um monossacarídeo), são fundamentais à homeostasia da grande maioria dos organismos em todos os domínios da vida. E entender estes papéis vem se tornando um importante alvo da bioinformática.

1.3. Problemas alvo

Considerando o tipo de informação manipulada, os problemas e questões abordados pela bioinformática podem ser agrupados entre aqueles relacionados a sequências de biomoléculas e aqueles relacionados à estrutura de biomoléculas (Figura 4-1). À primeira vista, considerando que de forma geral estruturas de proteínas são determinadas por seus genes, poderíamos imaginar que lidar com estruturas 3D seria redundante a manipular sequências, conjuntos de informações 1D. Esta percepção é limitada e não se configura como verdade para diversas questões. Na verdade, existem aspectos únicos em cada conjunto de informação, não diretamente transferíveis para o outro.

Inicialmente, como veremos adiante (item 1.4 e capítulo 2), o enovelamento de proteínas é um fenômeno extremamente complexo e ainda não totalmente compreendido, de forma que não somos capazes de transformar uma sequência linear de aminoácidos (codificada por seu gene) em uma estrutura 3D (salvo para algumas situações específicas, que serão vistas ao longo do livro).

Outro aspecto importante é que o enovelamento de proteínas, em muitas situações, depende de mais do que sua sequência de aminoácidos, envolvendo aspectos como o ambiente e o local onde a proteína estará na célula ou organismo, a ocorrência de modificação co- ou pós-traducionais e a sua interação com chaperonas. Para ilustrar o quanto este fenômeno é complexo, embora diversas sequências com identidade mínima possam ter estruturas 3D extremamente parecidas, em alguns casos a troca de um ou poucos resíduos de aminoácidos pode modificar totalmente a função, chegando até a interferir na forma tridimensional que uma proteína adota.

Em contrapartida, algumas informações presentes em sequências gênicas ou mesmo peptídicas não são necessariamente observáveis em estruturas tridimensionais. Por exemplo, regiões promotoras ou reguladoras da expressão gênica são facilmente descritas como informações 1D, e peptídeos sinal ou íntrons estão normalmente ausentes nas formas nativas de proteínas, sendo mais facilmente observáveis por sequências das biomoléculas em questão.

Adicionalmente, estruturas 3D de moléculas são formas muito mais complexas de serem manipuladas que sequências 1D, o que agrega uma série de dificuldades nos estudos de bioinformática. Assim, diversas tarefas tendem a ser muito simplificadas (ou mesmo de outra forma não seriam possíveis atualmente) quando trabalhamos com sequências em vez de estruturas. Por exemplo, a identificação de uma assinatura para modificação pós-traducional é muito mais ágil em uma sequência do que em um conjunto de milhares de átomos distribuídos em um espaço tridimensional.

Por fim, talvez o motivo mais prático para separarmos as duas abordagens se refere à facilidade de obtenção das informações. Os métodos experimentais para sequenciamento de ácidos nucleicos estão muito mais avançados do que os métodos para determinação da estrutura 3D de biomoléculas. A diferença de capacidade de determinação dos dois conjuntos de dados é de ordens de grandeza.

Questões relacionadas a sequências

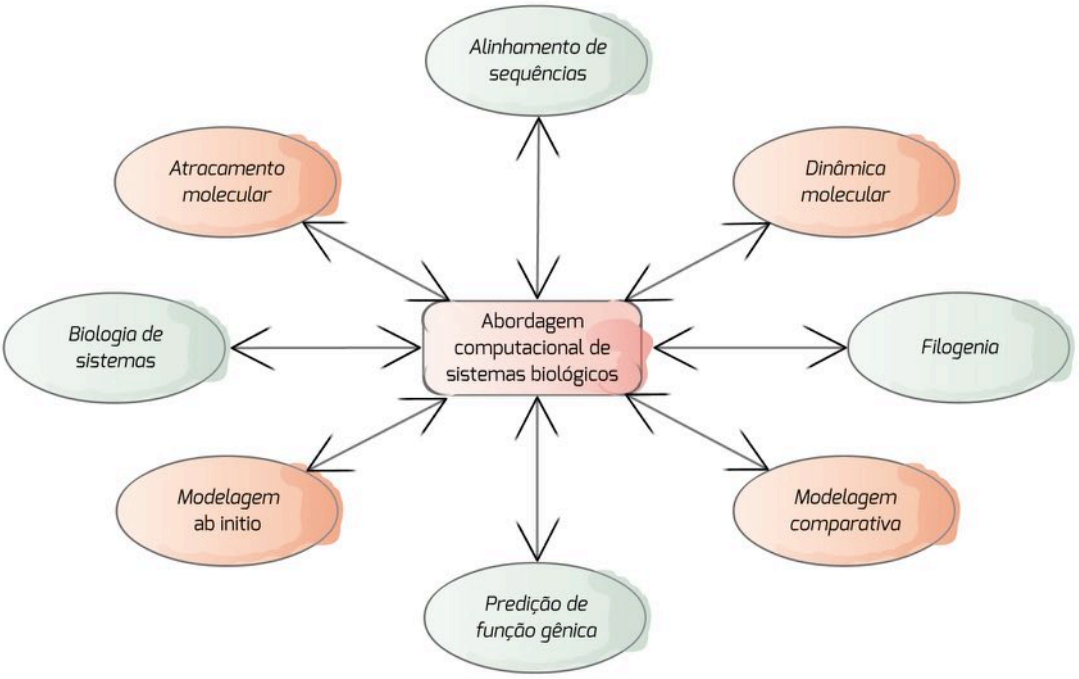

Figura 4-1: Representação de algumas das principais áreas da bioinformática. As metodologias que lidam majoritariamente com estruturas 3D estão representadas em laranja, enquanto as metodologias envolvidas principalmente com sequências estão representadas em verde. Devemos lembrar, contudo, que esta separação é imperfeita. Por exemplo, a modelagem comparativa parte de sequências, a função de um gene pode ser determinada pela estrutura da proteína associada.

A manipulação de sequências é menos custosa computacionalmente, nos possibilitando lidar com genomas inteiros. Isto permite realizar análises em indivíduos ou mesmo populações de indivíduos, nos aproximando do entendimendo dos organismos em sua complexidade biológica. Podemos traçar a história evolutiva de um conjunto de organismos ou construir redes de interação entre centenas ou milhares de moléculas de um determinado organismo, tecido ou tipo celular. Em linhas gerais, os objetos de estudo relacionados a sequências de biomoléculas incluem:

i) comparações entre sequências (alinhamento);
ii) identificação de padrões em sequências (assinaturas);
iii) caracterização de relações evolutivas (filogenia);
iv) construção e anotação de genomas;
v) construção de redes (biologia de sistemas).

Vale destacar que estas análises podem receber a contribuição de estudos envolvendo a estrutura das biomoléculas de interesse ou mesmo ser validadas por estas. Por exemplo, resíduos conservados evolutivamente possuem grande chance de possuírem papel funcional (como atuando na catálise) ou estrutural (estabilizando a estutura proteica). Assim, comparar um alinhamento à estrutura 3D pode tanto explicar quanto oferecer novas abordagens e considerações ao significado de conservações de resíduos maiores ou menores em conjuntos de sequências.

Questões relacionadas a estruturas

Ao contrário da manipulação de sequências, estruturas exigem um maior poder de processamento para serem manipuladas. Na prática, podemos manipular uma ou um pequeno punhado de estruturas simultaneamente (embora este número venha crescendo progressivamente). Neste caso, o foco costuma ser o entendimento de moléculas e dos eventos mediados por estas, individualmente, incluindo:

i) obtenção de modelos 3D para proteínas e outras biomoléculas (por exemplo, modelagem comparativa);
ii) identificação do modo de interação de moléculas (atracamento);
iii) seleção de compostos com maior potencial de inibição (atracamento);
iv) caracterização da flexibilidade molecular (dinâmica molecular);
v) avaliação do efeito de mudanças na estrutura e ambiente molecular na dinâmica e função de biomoléculas (dinâmica molecular).

O uso de sequências para alimentar estudos estruturais é mais comum na construção de modelos tridimensionais de proteínas a partir de suas sequências codificadoras, no método denominado modelagem comparativa (capítulo 7). Contudo, outras relações extremamente úteis podem ser estabelecidas. Por exemplo, por serem estruturas usualmente flexíveis, alças tendem a possuir uma maior capacidade de acomodar mutações ao longo da evolução. Isto permite uma comparação entre resultados de alinhamentos e, por exemplo, perfis de flexibilidade observáveis através de simulações por dinâmica molecular.

1.4. Tendências e desfios

Como uma área em rápido desenvolvimento, a bioinformática exige de seu praticante uma constante atenção a novas abordagens, métodos, requerimentos e tendências. Programas podem se tornar rapidamente ineficientes comparados a novas ferramentas ou mesmo obsoletos. Avanços de *hardware* podem (e na verdade vem fazendo isso) catapultar o nível de exigência metodológica pelas revistas de ponta. E há algumas áreas em específico nas quais a comunidade científica vem concentrando esforços. São por conseguinte áreas de grande impacto potencial e grande competição na literatura científica, dentre as quais destacaremos algumas abaixo.

Processamento em CPU e GPU

CPUs (*Central Processing Unit*s ou unidades de processamento central) ou simplesmente processadores (ou ainda microprocessadores) são partes dos computadores responsáveis pela execução das instruções estabelecidas pelos programas. Desde seu surgimento em torno da metade do século XX, as CPUs tornaram-se progressivamente mais complexas, confiáveis, rápidas e baratas. Esse processo foi previsto pioneiramente por Gordon E. Moore, no que ficou sendo conhecido desde então como a lei de Moore. Segundo esta lei, o número de transistores em um processador (na verdade em qualquer circuito integrado) dobra aproximadamente a cada 2 anos (Figura 5-1). O impacto do fenômeno descrito nesta observação na vida moderna é enorme, envolvendo desde nossos computadores, celulares e câmeras digitais até a precisão de estudos climáticos (com impacto na prevenção de catástrofes e na agricultura), medicina, engenharia, indústria bélica e aeroespacial. Com o aumento da velocidade e barateamento das CPUs, podemos a cada ano construir modelos mais precisos de fenômenos biológicos progressivamente mais complexos. Na prática, o avanço da bioinformática está ligado intrinsecamente à lei de Moore.

Em uma CPU podemos encontrar não somente um microprocessador, mas mais de um, o que é chamado multi-processamento e estas CPUs de processadores de múltiplos núcleos (*multi-core processing*). Hoje, a grande maioria dos processadores empregados em computadores, *notebooks* e celulares já possui múltiplos núcleos. Se o programa que estamos utilizando for adaptado para este tipo de processamento, o cálculo poderá ser distribuído pelos núcleos de processamento, tornando o cálculo significativamente mais rápido. A grande maioria dos aplicativos em bioinformática já possui versões compatíveis com processamento em múltiplos núcleos, e devemos estar atentos à escolha destas versões e à instalação de forma que essa característica esteja funcional, sob pena de subutilização da CPU.

Já GPUs (*Graphical Processing Unit*s ou unidades de processamento gráfico) são microprocessadores desenvolvidos inicialmente

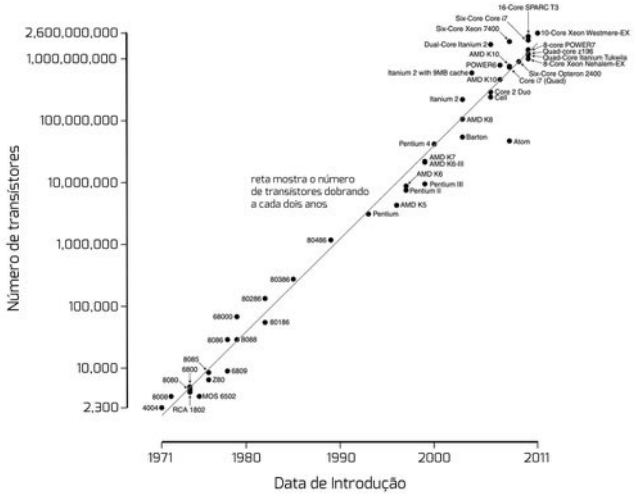

Figura 5-1: Representação da lei de Moore, indicando o aumento no número de transistores em microprocessadores no período de 1971 a 2011. Adaptada de William Wegman, 2011 (*Creative Commons*).

como unidades especializadas na manipulação de representações gráficas em computadores. Estão, assim, normalmente localizadas nas placas de vídeo de nossos computadores. O termo GPU foi popularizado a partir de 1999 com o lançamento da placa de vídeo GeForce256, comercializada pela Nvidia.

O desenvolvimento das GPUs remonta ao início dos anos de 1990, com o aumento do emprego de gráficos em 3D nos computadores e videogames. De fato, alguns dos primeiros exemplos de *hardware* dedicado ao processamento em 3D estão associados a consoles como PlayStation e Nintendo 64. Atualmente, enquanto CPUs possuem até em torno de uma dezena de núcleos de processamento, GPUs podem facilmente alcançar centenas ou mesmo milhares de núcleos de processamento, permitindo uma grande aceleração na manipulação de polígonos e formas geométricas, encontradas em aplicações 3D (como os jogos) e sua renderização (Figura 6-1). Tal aumento de performance ao dividir a carga de trabalho em um grande número de núcleos de processamento abriu um grande horizonte de possibilidades em computação científica, implicando em grande aumento na velocidade de manipulação de dados.

Diversos aplicativos em bioinformática vêm sendo portados para trabalhar com GPUs. Desde o alinhamento de sequências à filogenia, do atracamento molecular à dinâmica molecular, múltiplos pacotes estão disponíveis, tanto pagos quanto gratuitos, capazes de explorar a computação em GPU, e este número vem crescendo a cada ano, apontando para uma nova tendência na área. O usuário deve, contudo, observar seu problema alvo, pois a aceleração fornecida pela GPU dependerá das características do problema em questão e da eficiência e portabilidade do código empregado.

A combinação de CPUs e GPUs com múltiplos núcleos fez com que a capacidade de processamento de alguns supercomputadores de há alguns anos já esteja disponível para computadores pessoais, nos chamados supercomputadores pessoais.

Predições a partir de sequências

Quando estudamos uma sequência de nucleotídeos de DNA desconhecida é importante determinar seu papel funcional, por exemplo, se codificante de proteínas ou não. E, sendo codificante, qual proteína é produzida ao final da tradução e qual sua função. Tais predições são realizadas a partir de algoritmos construídos a partir de bancos de dados

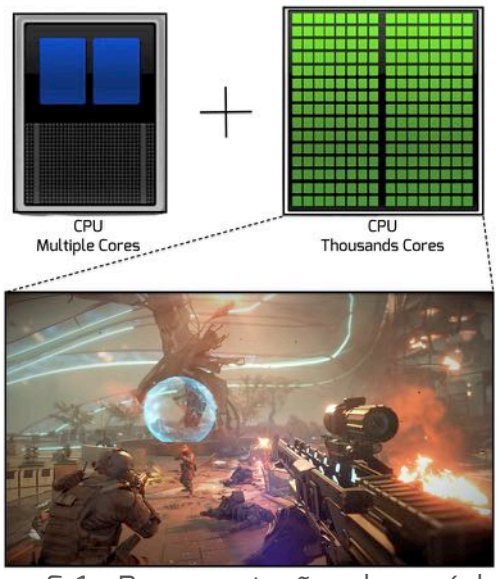

Figura 6-1: Representação dos núcleos de processamento em CPUs e GPUs. O grande número de núcleos em GPUs permite a realização de cálculos complexos rapidamente.

existentes, relacionando determinada sequência a características e propriedades específicas. Contudo, somente uma pequena quantidade de organismos teve seu genoma sequenciado até o momento e, destes, somente uma pequena parte de genes teve sua função determinada experimentalmente. Devemos, portanto, lembrar que as predições destes modelos estão relacionadas a quão completos foram os bancos de dados que os basearam. E que estes estão em contínuo avanço (ou seja, uma predição feita há 5 anos não necessariamente será igual a uma predição hoje que, por sua vez, pode ser diferente de uma predição de função gênica daqui a 5 anos - discutiremos no capítulo 3 alguns indicadores da qualidade dessas associações).

Predição de energia livre

Os fenômenos moleculares são regidos pela termodinâmica, tanto para reações químicas na síntese de um novo fármaco quanto à ação da DNA polimerase ou ao enovelamento de proteínas. Entender termos como entropia, entalpia e energia livre torna-se, assim, fundamental na adequada descrição destes fenômenos e, a partir desta, sua previsão computacional. Quando a medida destas variáveis se tornar precisa o bastante, poderemos esperar a substituição de diversos experimentos em bancada por cálculos em computadores mas, infelizmente, ainda não chegamos neste momento.

Predições de energia livre tem impacto direto na identificação da estrutura 2ária de moléculas de RNA, na localização de regiões do DNA para ligação de reguladores da transcrição, para a especificidade de enzimas por substratos e receptores por ligantes ou moduladores (fisiológicos ou terapêuticos, isto é, fármacos). Assim, diversos métodos foram desenvolvidos para a obtenção destas medidas, tais como a perturbação da energia livre, a integração termodinâmica, a energia de interação linear, a metadinâmica e diversas estratégias empíricas voltadas ao pareamento de nucleotídeos ou atracamento molecular.

A despeito desta diversidade de estratégias, a predição da energia livre em processos moleculares continua sendo um grande desafio. Em decorrência do elevado custo computacional associado a estes cálculos, diferentes tipos de simplificações e generalizações precisam ser realizadas, comprometendo nossa capacidade de empregá-los de forma ampla e fidedigna.

Enovelamento de proteínas

Como veremos adiante no livro, o enovelamento de proteínas é um dos processos mais complexos conhecidos pelo ser humano. O número de estados conformacionais possíveis para uma proteína pequena é gigantesco, dos quais um ou alguns poucos serão observáveis em solução em condições nativas. Os métodos experimentais usualmente empregados para tal, a cristalografia de raios-X e a ressonância magnética nuclear, são métodos caros e ainda possuem algumas limitações importantes em determinadas situações, apontando para a Bioinformática um potencial e importante papel na determinação da estrutura de biomoléculas.

Mas para que precisamos saber como é a estrutura tridimensional de uma determinada biomolécula? Esta pergunta possui muitas respostas, incluindo a compreensão de como a natureza evoluiu, como os organismos funcionam, como os processos patológicos se desenvolvem (e podem ser tratados) e como as enzimas exercem suas funções catalíticas. Tomemos este último caso como exemplo.

Com o entendimento de como proteínas se enovelam, será possível construir novas proteínas, capazes de adotar formas que a natureza não previu até o momento, enzimas aptas a catalizar reações de importância econômica, com menor toxicidade, o que terá por si impacto ambiental. Ainda, abre-se a possibilidade de planejamento racional de enzimas e proteínas envolvidas na detoxificação de áreas. Esta linha de pesquisa está em seu início, e o número de grupos de pesquisa dedicados ao redor do mundo para trabalhar na

engenharia de proteínas vem aumentando gradativamente. Mas, infelizmente, ainda não possuimos uma base teórica que nos permita entender e prever, com precisão e de forma ampla, a estrutura 3D de proteínas.

Contudo, esta problemática vem sendo abordada a cada ano com maior sucesso. Para proteínas com no mínimo em torno de 30% de identidade com outras proteínas de estrutura 3D já determinada, podem ser obtidos modelos de qualidade próxima àquela de métodos experimentais. Em outros casos, estruturas cristalográficas podem ser refinadas por métodos computacionais, agregando explicitamente informações ausentes nos experimentos (como a flexibilidade molecular). Outro exemplo é a construção de alças flexíveis, de difícil observação experimental mas que podem ser abordadas por diferentes métodos computacionais.

Para ácidos nucleicos, a construção computacional de estruturas 3D de moléculas de DNA é tarefa relativamente simples, que usualmente não requer os custos associados a experimentos de cristalografia e ressonância magnética. Para moléculas de RNA, contudo, a elevada flexibilidade traz consigo desafios adicionais. Mesmo assim, em diversos casos as estratégias computacionais possuem vantagens em lidar com moléculas muito flexíveis. Talvez o caso mais emblemático neste sentido sejam as membranas biológicas. Estas macromoléculas biológicas não são observáveis nos experimentos usuais capazes de determinar estruturas com resolução atômica, embora através de simulações por dinâmica molecular tenham suas estruturas descritas com elevada fidelidade.

Outro caso em que os métodos computacionais parecem possuir vantagens em relação aos experimentais envolve os carboidratos. Embora sejam moléculas em vários aspectos mais complexos que proteínas, carboidratos biológicos não parecem sofrer enovelamento nem adotar tipos de estrutura 2^{ária} em solução (embora o façam em ambiente cristalino), o que os torna na prática um problema estrutural mais simples que proteínas. De fato, vem sendo possível prever a estrutura de glicanas com graus variados de complexidade com grande precisão, um campo no qual os métodos experimentais possuem grandes dificuldades em abordar.

Validação experimental

Em linhas gerais, métodos computacionais devem ser comparados a dados experimentais para validação. Esta afirmação, embora tomada geralmente como um axioma, é bastante simplista, e não expressa claramente a complexidade e desafio nesta tarefa. Alguns pontos específicos incluem:

i) nem sempre há dados experimentais disponíveis para validar os cálculos e simulações realizados. Por exemplo, este é o caso com frequência para alinhamentos de sequências, para relações filogenéticas, para predições *ab initio* da estrutura de proteínas e para a descrição da flexibilidade de biomoléculas obtidas por dinâmica molecular. Nem sempre há fósseis ou outras evidências arqueológicas para validar antepassados evidenciados por estudos filogenéticos. Por outro lado, não há métodos experimentais com resolução atômica e temporal, de forma que a validação de simulações por dinâmica molecular é em grande medida indireta (uma estrutura obtida por cristalografia é única, sem variação temporal, enquanto os modelos oriundos de ressonância magnética nuclear correspondem a médias durante o período de coleta do dado);

ii) os dados experimentais devem ser adequados ao estudo computacional empregado. Assim, se estamos estudando a formação de um complexo fármaco-receptor, resultados *in vivo* devem ser evitados, enquanto os experimentos *in vitro* preferidos. Se administramos um determinado fármaco por via oral a um camundongo, este fármaco passará por diversos processos farmacocinéticos (absorção, distribuição, metabolização e excreção) que muito provavelmente irão interferir na ação

frente ao receptor alvo. Portanto, para estudos de atracamento, dados *in vivo* devem ser evitados;

iii) a margem de erro do dado experimental deve ser considerada quando comparada aos dados computacionais. Frequentemente a margem de erro para experimentos na bancada é maior que para aqueles realizados em computadores, limitando a extensão da validação. Usando novamente o exemplo de estudos de atracamento, se a afinidade experimental de um fármaco por seu receptor é de 0,11 ± 0,04 μM, valores teóricos de 97 nM a 105 nM estarão corretos. Por outro lado, frequentemente os resultados experimentais são expressos como a menor dose testada, por exemplo, > 5 μM. Assim, qualquer valor maior que 5 μM será validado pelo dado experimental, o que cria uma grande dificuldade de validação (como comparar 5 a, digamos, 1.000?);

iv) as condições nas quais os experimentos foram realizadas devem ser observadas com estrito cuidado. Temperatura, contaminantes, sais e concentrações diferentes daquelas no ambiente nativo são frequentemente requeridas por alguns métodos experimentais, e podem interferir nos resultados. Por exemplo, a melitina (principal componente do veneno da abelha *Apis mellifera*) aparece como uma hélice em estudos cristalográficos mas é desenovelada no plasma humano, como pode ser confirmado por experimentos de dicroismo circular com força iônica compatível com o plasma.

Assim, a despeito do axioma da exigência de validação experimental para estudos computacionais, não é infrequente que um dado computacional apresente maior precisão que um dado obtido na bancada. Na realidade, um modelo computacional, frequentemente chamado de teórico em oposição aos métodos ditos experimentais, não é nada além de um experimento computacional que, infelizmente, nem sempre tem contraparte em experimentos de "bancada". E esses adjetivos não carregam consigo qualificações quanto à confiabilidade dos resultados gerados.

1.5. Leitura recomendada

KHATRI, Purvesh; DRAGHICI, Sorin. Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. ***Bioinformatics***, 21, 3587-3593, 2005.

MORGON, Nelson H.; COUTINHO, K. ***Métodos de Química Teórica e Modelagem Molecular***. São Paulo: Editora Livraria da Física, 2007.

MIR, Luis. ***Genômica***. São Paulo: Atheneu, 2004.