Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Proteases Pr1 de *Metarhizium anisopliae*: Evolução, Estrutura e Função

Tese de Doutorado

Fábio Carrer Andreis

Porto Alegre, novembro de 2021

Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Proteases Pr1 de *Metarhizium anisopliae*: Evolução, Estrutura e Função

Tese submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para obtenção do grau de Doutor em Ciências.

Fábio Carrer Andreis

Augusto Schrank - Orientador

Claudia Elizabeth Thompson - Co-orientadora

Porto Alegre, novembro de 2021

Este trabalho foi desenvolvido na Unidade de Biologia Teórica e Computacional e no Laboratório de Biologia Celular e Molecular de Fungos Filamentosos, no Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (CBiot/UFRGS), sob orientação conjunta dos Doutores Augusto Schrank e Claudia Elizabeth Thompson. Apoio financeiro: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

Banca Examinadora

Charley Christian Staats
Universidade Federal do Rio Grande do Sul

Glória Regina Franco
Universidade Federal de Minas Gerais

Rodrigo Ligabue-Braun
Universidade Federal de Ciências da Saúde de Porto Alegre

Fernanda Cortez Lopes
Universidade Federal do Rio Grande do Sul
(Relatora e Suplente)

Sumário

Resumo

A família Pr1 de proteases tem papel importante na patogenicidade e virulência de artropatógenos como *Metarhizium anisopliae*. Esses fatores de virulência geralmente atuam na penetração da cutícula do hospedeiro, etapa essencial do processo infectivo. Há 11 proteoformas de Pr1 (Pr1A a Pr1K) descritas. Essa família é dividida em duas classes, sendo que a Classe II (tipo-proteinase K) inclui 10 parálogos subdivididos em três subfamílias. Essas proteoformas agem em sinergia e com outros fatores de virulência, conferindo patogenicidade a diferentes hospedeiros. À medida em que a virulência coevolui por seleção recíproca com os hospedeiros, a seleção positiva pode levar à evolução de novas famílias de proteases ou parálogos das existentes que possam superar as defesas do hospedeiro. Essa hipótese é suportada por proteínas Pr1 da Classe II, pois evidenciamos: (i) seleção positiva em seis dos dez parálogos de Pr1 (em maioria no domínio proteolítico); (ii) divergência funcional Tipo I em comparações intra-subfamília, com suporte a uma potencial nova proteoforma de Pr1J; (iii) localizações projetadas em estrutura terciária próximas ao sítio catalítico, com potencial impacto na catálise. Uma abordagem mista computacional-experimental foi desenvolvida para caracterizar essa nova proteoforma de Pr1J e identificar os efeitos de sua evolução, utilizando simulações comparativas por dinâmica molecular e expressão heteróloga em *Escherichia coli*. Até o momento, diferentes elementos estruturais e comportamentos conformacionais podem ser observados entre ambas as proteoformas de Pr1J. Ademais, as construções de vetores plasmidiais foram bem-sucedidas, embora a sua expressão não tenha sido possível. Em conjunto, essa abordagem mista pode apontar os efeitos da duplicação e diversificação proteica nas capacidades proteolíticas de *M. anisopliae*. Até o momento, os resultados implicam a existência de pressão seletiva diferencial em genes *pr1* e uma potencial nova proteoforma, provavelmente afetando especificidade de

hospedeiros, virulência ou ainda adaptando o organismo a diferentes estilos de vida independentes do hospedeiro.

Abstract

The Pr1 family of proteases plays an important role in pathogenicity and virulence of arthropathogens such as *Metarhizium anisopliae*. These virulence factors are active during the penetration of the host cuticle, an essential step in the infective process of this fungus, which possesses 11 Pr1 proteoforms (Pr1A through Pr1K). This family is divided in two classes, with Class II (proteinase K-like) comprising 10 paralogs further split into three subfamilies. These proteoforms act synergistically and with other virulence factors, conferring pathogenicity to multiple hosts. As virulence coevolves by reciprocal selection with hosts, positive selection may lead to the evolution of new protease families or paralogs of extant ones that can withstand host defenses. This hypothesis is supported in Class II Pr1 proteins, as we have evidenced: (i) positive selection in six out of ten Pr1 paralogs (mostly located on the proteolytic domain); (ii) Type I functional divergence in intra-subfamily pairwise comparisons, also supporting a potential novel Pr1J proteoform; (iii) tertiary structure projected locations being closely located to the enzyme's catalytic cleft, potentially impacting catalysis. A mixed computational-experimental approach was developed in order to characterize this novel Pr1J protein proteoform and identify the effects of positively selected sites and residues related to functional divergence, by means of comparative molecular dynamics simulations and heterologous expression in bacterial vectors. While these analyses are currently underway, different structural elements and conformational behaviors can already be observed among both Pr1J proteoform. Furthermore, plasmidial vector constructions have been successful, although their expression is yet to be accomplished. In conjunction, this mixed approach should provide a comprehensive view of the effects of protein duplication and diversification regarding proteolytic capabilities in *M. anisopliae*. So far, our results imply the existence of differential selective pressure acting on *pr1* genes and a potential novel proteoform, likely affecting host specificities, virulence or even adapting the organism to different host-independent lifestyles.

Lista de Abreviaturas e Símbolos

ºC	Grau(s) Celsius
aa	Aminoácido(s)
Ala	Alanina
cDNA	DNA complementar
CDS	do inglês, *Coding DNA Sequence*
DEPC	DiEtilPiroCarbonato
DM	Dinâmica Molecular
dN	taxa de substituições não-sinônimas
DNA	do inglês, *DeoxyriboNucleic Acid*
dNTP	do inglês, *DeoxyNucleosideTriPhosphate*
DOI	do inglês, *Digital Object Identifier*
DOPE	do inglês, *Discrete Optimized Potential Energy*
dS	taxa de substituições sinônimas
DTT	DiTioTreitol
e.g.	do latim, *exempli gratia*; por exemplo
EDTA	do inglês, *EthyleneDiamine Tetraacetic Acid*
EST(s)	do inglês, *Expressed Sequence Tags*
FAP(s)	Fungo(s) Artropatogênico(s)
g	grama(s)
h	Hora(s)
i.e.	do latim, *id est;* isto é
IPTG	Isopropil β-D-1-tiogalactopiranosídeo
K	Kelvin
L	Litro(s)
LB	Luria-Bertani; meio de cultura
M	Molar; mol/L
MiA	Milhões de Anos
min	Minuto(s)
MM	Mecânica Molecular, usada intercambiavelmente com a

	expressão em inglês, *Molecular Mechanics*
NCBI	do inglês, *National Center for Biotechnology Information*
NPT	*Ensemble com* número de partículas, pressão e temperatura constantes
NVT	*Ensemble* número de partículas, volume e temperatura constantes
pb	pares de bases
PCR	do inglês, *Polymerase Chain Reaction*
PDB	Protein DataBank
PEG	PoliEtilenoGlicol
pH	potencial de hidrogênio; *-log [H+]*
Phe	Fenilalanina
pK	constante de dissociação; *-log K*
Pro	Prolina
QM	do inglês, *Quantum Mechanics*
RAF	Região Adicionalmente Favorável
rbs	do inglês, *ribosome binding site*
Rg	Raio de Giração, ou Raio de Giro
RMF	Região Mais Favorável
RMSD	do inglês, *Root Mean Square Deviation*
RMSF	do inglês, *Root Mean Square Fluctuation*
RNA	do inglês, *RiboNucleic Acid*
RNA-Seq	do inglês, *RNA Sequencing*
rpm	Rotações Por Minuto
s	Segundo(s)
SASA	do inglês, *Solvent-Acessible Surface Area*
SDS-PAGE	do inglês, *Sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis*
sf1 / Sf1	Subfamília 1
sf2 / Sf2	Subfamília 2
sf3 / Sf3	Subfamília 3
t	tempo

TAE	Tris-Acetato-EDTA
U	unidades

Índice de Figuras

Índice de Tabelas

Introdução

Evolução de Fungos

No contexto da vida na Terra (aproximadamente esférica[1]), é inegável a importância dos fungos nos mais diversos ecossistemas que habitam. Reconhecidos como um reino taxonômico, Fungi[2], desempenham papel vital na bioquímica global, reciclando carbono e mobilizando nitrogênio, provendo suporte à vida de plantas na rizosfera ou de forma endofítica, por exemplo (NARANJO-ORTIZ; GABALDÓN, 2019). Em outra frente, características particulares de seus metabolismos permitiram que a sociedade humana produzisse antibióticos para tratamento de infecções, além de alimentos e bebidas fermentadas para saciedade ou prazer, ou ainda que explorasse a própria biomassa fúngica para alimentação. Essas singularidades metabólicas trazem consigo possibilidades danosas, contudo. Patógenos fúngicos podem dizimar populações animais e vegetais, extinguindo espécies ou ameaçando cadeias de produção alimentícia (NARANJO-ORTIZ; GABALDÓN, 2019). Entretanto, mesmo essas características inicialmente negativas podem ser exploradas de maneira benéfica, a exemplo da utilização de espécies fúngicas para biocontrole de pragas agrícolas (IWANICKI et al., 2019), a ser melhor explorada em seções subsequentes.

A classificação taxonômica de fungos passou por diversas revisões. Parte do problema envolve a própria definição do que é um fungo “verdadeiro”. Embora existam exceções, características comuns a fungos incluem: parede celular composta de quitina e β-glicanas[3], apresentação unicelular ou crescimento como micélio, presença da rota do aminoadipato para biossíntese de lisina, e cristas mitocondriais achatadas (ADL et al., 2012, 2019). Enquanto o percurso de classificação dos fungos verdadeiros foi baseado em traços morfológicos e reprodutivos em seus primórdios, dados de sequenciamento genômico e

[1] Infelizmente, se faz necessária essa distinção, dada a abismal disseminação de pseudociência.
[2] Juntamente a Animalia, Plantae, Protista, Bacteria e Archaea.
[3] Nos esporos, pelo menos.

metagenômico permitiram sua divisão taxonômica em nove filos principais (Figura 1): Basidiomycota, Blastocladiomycota, Chytridiomycota, Glomeromycota, Mucoromycota, Neocallimastigomycota, Opisthosporidia, Zoopagomycota e Ascomycota, ao qual direcionam-se os estudos deste trabalho.

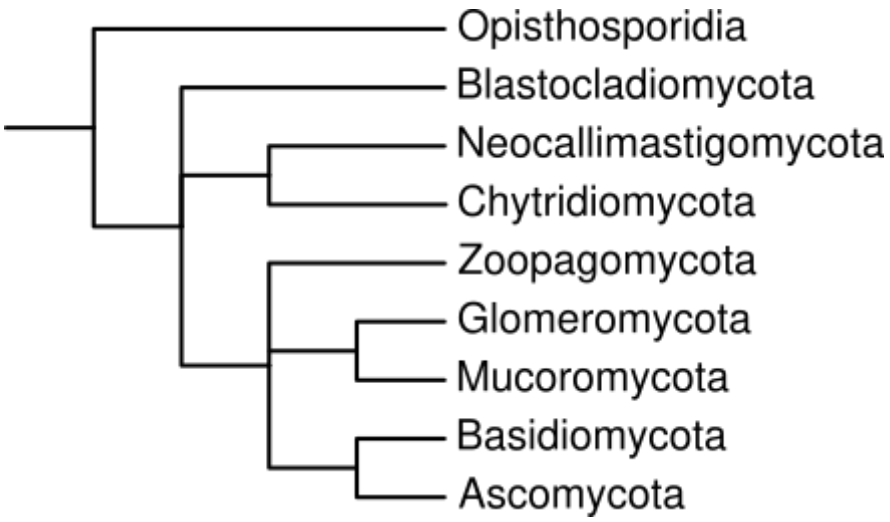


Figura 1: Reino *Fungi*. Árvore esquemática representando os nove filos majoritários de fungos e suas relações evolutivas. Unidades taxonômicas fora do nível de filo foram omitidas por simplicidade. Adaptado de (NARANJO-ORTIZ; GABALDÓN, 2019).

Contendo aproximadamente dois terços de todas as espécies fúngicas descritas (LUTZONI et al., 2004; SCHOCH et al., 2009), o filo Ascomycota contém organismos-modelo de notável importância científica e econômica (i.e. *Aspergillus nidulans, Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* e *Neurospora crassa*, dentre outros). Cruzamentos entre espécies de ascomicetos induzem a formação de hifas dicarióticas[4], que levam à formação de ascos contendo esporos (ascosporos) de origem meiótica. Porém, não é incomum a ocorrência de reprodução assexuada em vários membros deste filo. De particular interesse, o subfilo Pezizomycotina[5] possui a maior diversidade dentre os ascomicetos e sua anatomia básica é filamentosa e anastomizada (ADL et al., 2012; HEALY et al.,

[4] Característica do sub-reino Dikarya, composto por Ascomycota e Basidiomycota, onde a fusão de hifas é disjunta da meiose (NARANJO-ORTIZ; GABALDÓN, 2019).
[5] Os outros sendo Taphrinomycotina e Saccharomycotina, este sendo clado-irmão de Pezizomycotina (NARANJO-ORTIZ; GABALDÓN, 2019).

2013; LIU et al., 2008), com ascos normalmente protegidos por ascocarpos. Único a este grupo taxonômico fúngico é a evolução mesosintênica[6], evidenciada por genômica comparativa (HANE et al., 2011), e metabolismo secundário altamente desenvolvido (SBARAINI et al., 2016; WISECAVER; ROKAS, 2015; WISECAVER; SLOT; ROKAS, 2014). Os Pezizomycotina possuem 13 classes conhecidas, cada uma abrangendo diversas ordens, porém há em torno de 5000 membros desse subfilo sem classificação exata (NARANJO-ORTIZ; GABALDÓN, 2019). Por objetividade, focaremos na classe Sordariomycetes, que abrange em torno de 2,2 x 10^4 organismos conhecidos, o segundo mais abundante desse subfilo (NARANJO-ORTIZ; GABALDÓN, 2019). Morfologicamente diversos, sordariomicetos possuem estilos de vida saprotrófico, patogênicos a plantas, parasitas de animais ou fungos, endofítico ou até formando líquens (NARANJO-ORTIZ; GABALDÓN, 2019).

A ordem Hypocreales, da classe Sordariomycetes, abrange uma cornucópia de fungos fitopatogênicos, endofíticos, micoparasitas, simbiontes de insetos e patógenos de artrópodes (KEPLER et al., 2012; SUH; NODA; BLACKWELL, 2001). Em particular, a entomopatogenicidade evoluiu de maneira independente nas famílias Cordycipitaceae, Ophiocordycipitaceae e Clavicipitaceae, de tal forma consistente com transições repetidas entre hospedeiros plantas, fungos ou insetos (SUH; NODA; BLACKWELL, 2001). Direcionando o foco especialmente à família Clavicipitaceae, destaca-se que é composta majoritariamente por espécies patogênicas, incluindo patógenos de plantas, fungos, répteis e artrópodes (KEPLER et al., 2014). O gênero *Metarhizium*, por exemplo, inclui representantes patogênicos a répteis (*Metarhizium viride* e *Metarhizium granulomatis*) (SCHMIDT et al., 2017), a cogumelos (*Metarhizium marquandii*, também saprófito) e a insetos (*Metarhizium acridum*). *Metarhizium* spp. são tradicionalmente considerados patógenos de insetos que apresentam esporos assexuais esverdeados, divergindo de representantes endofíticos há aproximadamente 307 milhões de anos (MiA) (ST

[6] Manutenção de conteúdo gênico sem conservação de ordem de ocorrência.

LEGER; WANG, 2020), e vem sendo de particular interesse no estudo da patogenicidade e em aplicações comerciais voltadas ao biocontrole de pragas.

Artropatógenos e *Metarhizium* spp.

Fungos artropatogênicos (FAP)[7] vem sendo utilizados para biocontrole há mais de um século, oferecendo alternativas ambientalmente corretas se comparados a pesticidas sintéticos convencionais para controle de pragas artrópodes. Um foco especial vem sendo dado à ordem Hypocreales, devido a mecanismos eficientes de infecção, abrangência de hospedeiros e facilidade de produção em massa (BUTT; JACKSON; MAGAN, 2001). Aproximadamente 80% das aplicações comerciais de FAPs são baseadas nos gêneros *Metarhizium* e *Beauveria* (FARIA; WRAIGHT, 2007). Especialmente, *Metarhizium anisopliae* foi a primeira espécie de FAP a ser produzida em massa e aplicada a esse propósito. Sendo uma das espécies de melhor caracterização quanto à virulência e especificidade de hospedeiros em nível molecular, *M. anisopliae* tem grande importância no estudo da artropatogenicidade (ST LEGER; WANG, 2020)

O banco de dados Mycobank[8] apresenta 85 espécies reconhecidas de *Metarhizium* spp. (*Name status: Legitimate*; incluindo formas e variantes[9]), cujas sequências genômicas já foram determinadas para 8 espécies (Tabela 1). Estudos filogenômicos (HU et al., 2014) apontam para a existência de grupos considerados hospedeiro-específicos ou especialistas[10] como primitivas às demais espécies amostradas. Fungos cuja abrangência de hospedeiros é mais diversa são considerados (hospedeiro-)generalistas, abrangendo 7 ordens de artrópodes[11]. Integrantes com número intermediário de hospedeiros[12] são

[7] Usado como alternativa para “entomopatogênicos”, visto que os hospedeiros de alguns FAPs incluem não-insetos (e.g. *Metarhizium*). É, porém, um termo não-usual.
[8] http://www.mycobank.org/; consulta em 5 de agosto de 2020.
[9] Do latim, respectivamente, *forma* e *varietas*, abreviados como f. e var.
[10] *Metarhizium album* e *Metarhizium acridum*, patogênicos às ordens Hemiptera e Orthoptera, respectivamente.
[11] *M. anisopliae, Metarhizium brunneum* e *Metarhizium robertsii*, abrangendo Diptera, Lepidoptera, Coleoptera, Hymenoptera, Orthoptera, Hemiptera e Ixodida.
[12] *Metarhizium majus* e *Metarhizium guizhouense,* ambos abrangendo Lepidoptera e Coleoptera.

definidos como espécies transicionais no processo de diversificação de hospedeiros no gênero *Metarhizium*. Presume-se que a especiação direcionada à generalização do leque de hospedeiros, partindo de especialistas, ocorreu paralelamente à diversificação dos artrópodes em um processo coevolutivo. Em espectro mais abrangente, a ampliação do rol de hospedeiros dos *Metarhizium* gerou aumento no tamanho genômico, número total de genes e expansão de famílias gênicas – especialmente as associadas à interação com os hospedeiros.

Tabela 1: Genomas disponíveis para *Metarhizium* spp. conforme o banco de dados do NCBI. Consulta em 18/06/2021.

Espécie	Montagens	ID	Comprimento total (Mb)	Proteínas	GC%
M. acridum	1	2443	39,42	9849	49,80
M. album	1	11737	30,45	8472	52,70
M. anisopliae	6	2190	38,59	11415	50,90
M. brunneum	2	15954	37,43	11054	51,08
M. guizhouense	1	15953	43,47	11787	49,60
M. majus	1	37185	42,06	11535	51,00
M. rileyi	2	44815	31,91	8854	49,65
M. robertsii	2	13329	40,99	12036	51,10

Em particular, *M. anisopliae* foi a primeira espécie de FAP a ser produzida em massa e a ser utilizada de forma bem-sucedida para fins de biocontrole (ZIMMERMANN; PAPIEROK; GLARE, 1995). Esse organismo é um dos FAPs melhores caracterizados e é aplicado globalmente em programas de controle biológico (BEYS-DA-SILVA et al., 2020), podendo ser encontrado no solo[13], endofiticamente ou de maneira infectiva ou saprotrófica (ST LEGER, 2008). Em particular, a forma infectiva (Figura 2) pode ser dividida em 3 grandes etapas (BUTT et al., 2016; SÁNCHEZ-PÉREZ et al., 2014): penetração, crescimento e reprodução. O estágio de penetração inclui o reconhecimento da superfície do hospedeiro suscetível, adesão e germinação do conídio, desenvolvimento de apressório e tubo germinativo e a penetração propriamente dita. Em seguida, na

[13] Habitando a rizosfera de plantas.

etapa de crescimento, ocorre a diferenciação das hifas em blastosporos[14], crescimento intenso do FAP e consumo de recursos do hospedeiro até sua morte. Por fim ocorre nova diferenciação de hifas, que emergem através da cutícula e desenvolvem conidióforos, cuja liberação de conídios no ambiente permite a reprodução através de novo hospedeiro (BEYS-DA-SILVA et al., 2020; BOOMSMA et al., 2014). Entretanto, esse ciclo só é possível se o conídio resistir às diversas defesas, induzidas ou pré-existentes, do hospedeiro (BUTT et al., 2016). A primeira e principal dessas é a cutícula, composta principalmente por lipídeos[15], proteínas e quitina[16]. Para que ocorra a penetração física dessa complexa camada, conídios de *M. anisopliae* formam apressórios, estruturas especializadas que exercem simultaneamente pressão mecânica e secreção de enzimas lipolíticas, quitinolíticas e proteolíticas (BUTT et al., 2016; SCHRANK; VAINSTEIN, 2010). Esse balanço químico entre parasita e hospedeiro, no que se refere respectivamente a fatores de virulência e moléculas de defesa, em prol do agente invasor é essencial para a manutenção e evolução desse estilo de vida. Os genes codificantes de proteases fúngicas, em particular, apresentam-se como modelos promissores no estudo da evolução adaptativa de famílias multigênicas e seu impacto nos processos de especiação (VILCINSKAS, 2010).

As proteases, em particular, desempenham funções essenciais na patogenicidade de *M. anisopliae* (ROSAS-GARCÍA et al., 2014). No processo infectivo desse fungo filamentoso, são conhecidos pelo menos três tipos distintos de proteases: serino-proteases dos tipos subtilisina (Pr1[17]) e tripsina (Pr2[18]), além de um tipo de metaloprotease (ST LEGER; BIDOCHKA; ROBERTS, 1994). As proteínas tipo-tripsina e tipo-subtilisina pertencem a distintas superfamílias de serino-proteases, cuja evolução convergiu independentemente a mecanismos catalíticos similares. As funções dessas enzimas vão além da penetração cuticular, também disponibilizando as proteínas do hospedeiro para a nutrição e

[14] Formas leveduriformes que se difundem passivamente pela hemolinfa do artrópode.
[15] Na epicutícula
[16] Na procutícula
[17] EC 3.4.21.62
[18] EC 3.4.21.4

agindo em resposta às defesas do inseto através da hidrólise de peptídeos antimicrobianos e inibidores de proteases, a citar alguns (BUTT et al., 2016; VILCINSKAS, 2010).

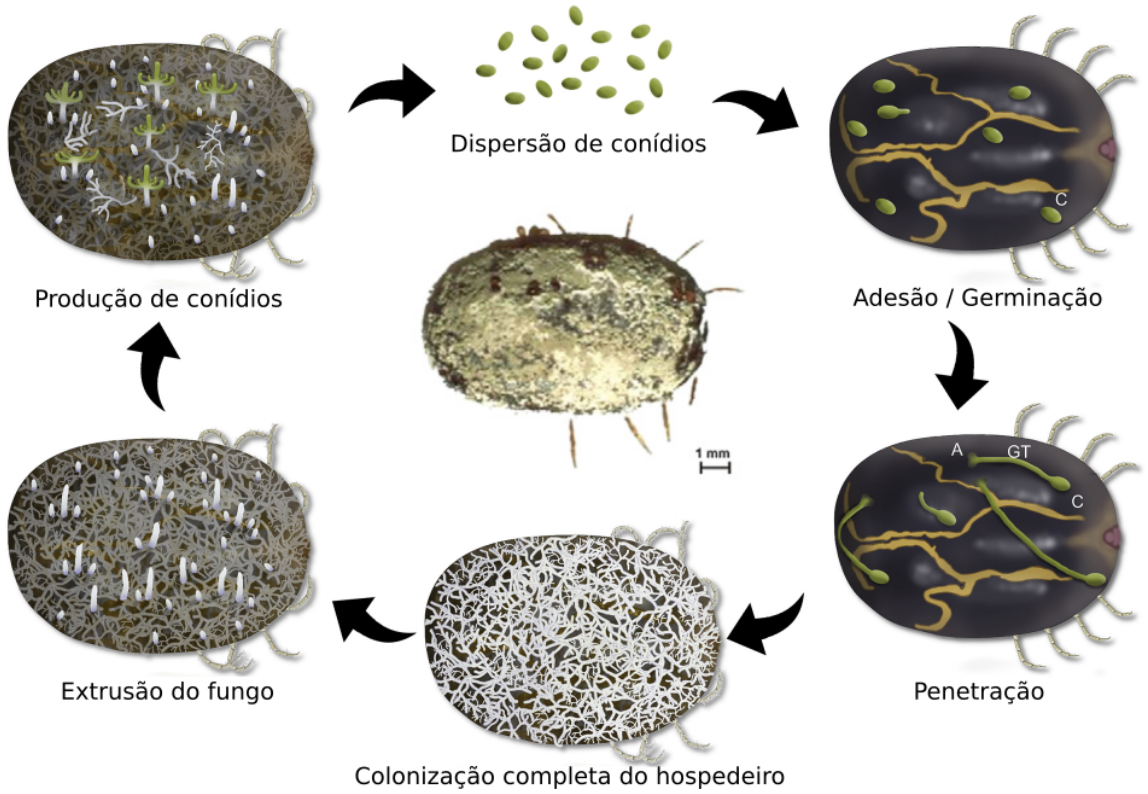


Figura 2: Visão esquemática da forma infectiva de *M. anisopliae*. Retrata-se o ciclo em forma cartunizada do carrapato bovino *Rhipicephalus microplus*. Ao centro, visão ampliada do estágio final de colonização, onde a superfície do cadáver apresenta-se tomada pelos conidióforos esverdeados de *M. anisopliae*. C - conídio; A - apressório; GT - tubo germinativo (sigla do do inglês, *germ tube*). Adaptado de (BEYS-DA-SILVA et al., 2020; SCHRANK; VAINSTEIN, 2010).

Proteases Pr1

A família Pr1 de proteases está intimamente ligada à patogenicidade de *M. anisopliae.* Análises de ESTs indicam a presença de 11 parálogos de Pr1, sendo mais prevalentes os transcritos de Pr1A, seguidos de Pr1J - quase oito vezes menos abundantes (FREIMOSER et al., 2003). Estudos filogenéticos posteriores empregando três linhagens de *M. anisopliae* (ARSEF 2575, ARSEF 324 e ARSEF

820) (BAGGA et al., 2004) sugerem a divisão dessa família em duas classes (Tabela 2). A Classe I (tipo-bacteriana) contém a proteoforma Pr1C, enquanto a Classe II, tipo-proteinase K, subdivide-se em três subfamílias que abrangem as isoenzimas remanescentes.

A subfamília extracelular 1 (Sf1) compreende Pr1 A, B, G, I e K, sendo caracterizada por genes contendo de dois a três íntrons e quatro resíduos de cisteína conservados nas proteínas codificadas (BAGGA et al., 2004). Em especial, Pr1A é reconhecida como um importante fator de virulência de *M. anisopliae*, sendo que linhagens superexpressando essa proteoforma reduziram o tempo de morte do hospedeiro em até 25%, ocorrendo também redução da alimentação em 40% enquanto vivos, parâmetro crítico na avaliação de pesticidas comerciais (ST LEGER et al., 1996). Adicionalmente, mutantes espontâneos com deleção dos genes *pr1A* e *pr1B* apresentaram virulência reduzida - sem efeito na patogenicidade - para *Tenebrio molitor*, porém não para *Galleria mellonella* (WANG; TYPAS; BUTT, 2002), implicando que as diferentes proteoformas de Pr1 podem ter efeito na especificidade a determinados hospedeiros. O mesmo trabalho também aponta que linhagens apresentando deleções de alguns genes *pr1* ainda são capazes de infectar seus respectivos hospedeiros, embora com virulência reduzida. O processo é, portanto, intrincado e multifatorial.

A subfamília extracelular 2 (Sf2) compreende Pr1 D, E, F e J. Genes neste grupo podem possuir até dois íntrons, não havendo conservação de códons de cisteína. Em particular, os genes *pr1E* e *Pr1F*[19] estão organizados *in tandem*, distantes cerca de 800 pb, implicando que sua origem decorreu de duplicação de um gene ancestral similar a *pr1F*. É proposto que um ancestral de *Metarhizium* spp. herdou *pr1D* e um gene similar a *pr1J*, este sofrendo duplicação, resultando em gene ancestral similar a *pr1F* (BAGGA et al., 2004). Também existem indícios de que a duplicação que deu origem às subfamílias 1 e 2 ocorreu há aproximadamente 400 MiA no filo Ascomycota, anteriormente à divergência entre leveduras e ascomicetos filamentosos (HU; ST. LEGER, 2004).

[19] Ambos sem íntrons

Tabela 2: Classificação da família Pr1 de proteases, conforme proposto por (BAGGA et al., 2004)

Classe	Subfamília	Proteoforma	Localização
Classe I (Tipo-Subtilisina)	N/A	C	Extracelular
Classe II (Tipo-Proteinase K)	1 (sf1)	A	Extracelular
		B	Extracelular
		G	Extracelular
		I	Extracelular
		K	Extracelular
	2 (sf2)	D	Extracelular
		E	Extracelular
		F	Extracelular
		J	Extracelular
	3 (sf3)	H	Intracelular

N/A - Não-Aplicável

A última subdivisão da Classe II é composta apenas pela isoenzima Pr1H, ocupando a subfamília intracelular 3 (Sf3), independente das demais. A natureza intracelular dessa proteína é inferida pela ausência de peptídeo-sinal identificável (BAGGA et al., 2004). Há indícios de que a evolução dos genes associados às proteínas tipo-subtilisina encontradas em ascomicetos ocorreu a partir de proteínas intracelulares até as extracelulares, o que pode ter facilitado o uso dessas proteases como fatores de virulência por espécies patogênicas (LI et al., 2010, 2017).

O único representante da Classe I em *M. anisopliae* (Pr1C) não foi incluído nas análises de 2004 (BAGGA et al., 2004) por ser altamente divergente das demais e introduzir um viés nos alinhamentos. Sendo assim, não foi possível

relacioná-lo com os demais membros da família sob um mesmo olhar. Essa proteoforma em particular apresenta maior similaridade com sequências de *Bacillus* spp., não apresentando íntrons ou sítios conservados de cisteína (BAGGA et al., 2004). Análises *in silico* apontam potenciais funções regulatórias e interação com outras proteases tipo-subtilisina[20], dentre outras proteínas em etapas iniciais de infecção (BEYS-DA-SILVA et al., 2014).

As proteínas Pr1, em sua maioria, são sintetizadas como pré-pró-peptídeos (Figura 3) (SIEZEN; LEUNISSEN, 1997). A região "pré" é localizada nos primeiros 15 a 22 aminoácidos (aa) dos polipeptídeos e corresponde a um peptídeo-sinal, este ausente em Pr1H (BAGGA et al., 2004; LI et al., 2010). Sequencialmente, temos a região "pró", localizada entre o peptídeo-sinal e a região da protease madura e com 60 a 80 aa de extensão (BAGGA et al., 2004). Esse segmento, sem assinatura evidente em Pr1E e F, possui função no enovelamento proteico como chaperona intramolecular (LI et al., 1995) e atua como inibidor temporário do domínio proteolítico (KOJIMA; MINAGAWA; MIURA, 1997) até que assuma sua conformação final, de aproximadamente 280 aa (ST. LEGER, 1995).

Embora não se conheça completamente a função individual de cada proteoforma de Pr1, espera-se que as atividades das enzimas que forem parcial ou totalmente não inibidas complementem funcionalmente umas às outras no processo infectivo (VILCINSKAS, 2010). Potencialmente, esse processo pode aumentar a adaptabilidade e alcance de hospedeiros, ou até mesmo favorecendo a sobrevivência em diferentes *habitats* externos. Diferenças de estabilidade, adsorção e afinidade por substratos, dentre outras características, sugerem que as proteínas tipo-subtilisina interagem sinergisticamente com outras enzimas associadas à cutícula, hidrolisando mais eficientemente as componentes cuticulares (BAGGA et al., 2004; BUTT et al., 2016; LI et al., 2010).

[20] Pr1A, B, I e J.

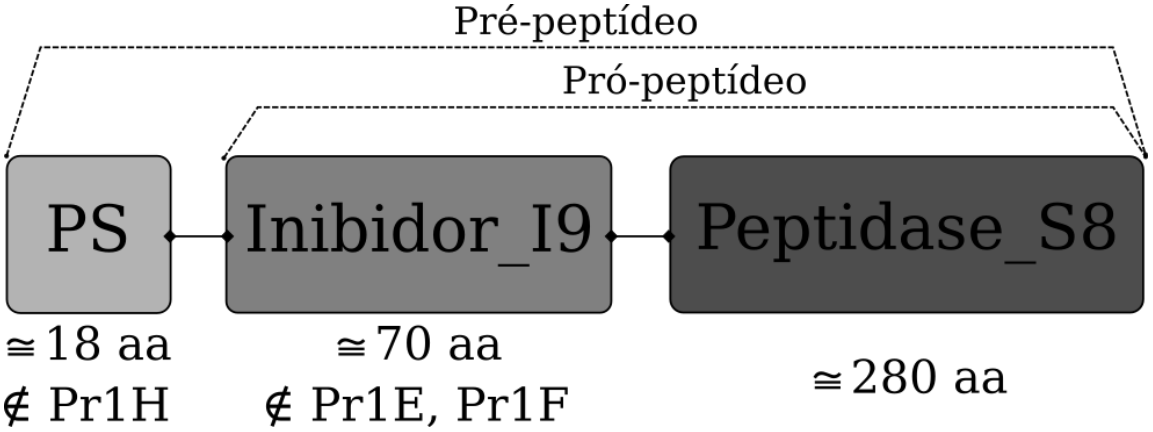


Figura 3: Arquitetura típica da família Pr1 de proteases. Em geral, o polipeptídeo transcrito possui peptídeo-sinal (PS) para secreção, este ausente na proteoforma H, um domínio dito inibitório (Inibidor_I9), ausente nas proteoformas E e F, e o domínio proteolítico (Peptidase_S8) da superfamília das subtilases. A nomenclatura dos domínios segue a empregada no PFAM, das referências PF05922 (*Inhibitor_I9*) e PF00082 (*Peptidase_S8*).

A expressão diferencial de proteases por *Metarhizium* spp. em diferentes substratos pode ser interpretada como uma adaptação fisiológica aos inibidores de proteases dos hospedeiros (VILCINSKAS, 2010). Análises proteômicas de *M. anisopliae* durante a infecção de *Dysdercus peruvianus* (BEYS-DA-SILVA et al., 2014) sugerem aumento da expressão de Pr1A, B, C e I, além de redução de expressão Pr1J nos estágios iniciais de infecção (48 h). Adicionalmente, Pr1B continua aumentada após 96 h do início da infecção. Na infecção de *Spodoptera exigua* (JAVAR et al., 2015), verifica-se aumento gradual na expressão de Pr1A nos estágios iniciais e atinge pico 1.000 vezes maior na etapa de conidiação, em relação ao início do processo, sugerindo que essa proteoforma também tem ação no processo de extrusão de hifas nos estágios finais, possivelmente disponibilizando nutrientes. Da mesma forma, em *Galleria mellonella* (SMALL; BIDOCHKA, 2005), verifica-se expressão aumentada de Pr1A na formação de apressórios e na conidiação. Experimentos de RNA-Seq a partir de material extraído de culturas de *M. anisopliae* em cutículas de *Rhipicephalus microplus* (STAATS et al., 2014) evidenciam superexpressão das proteoformas C, I, J e K a 48 h pós-infecção, voltando a níveis basais em 144 h pós-infecção. Já em

Manduca sexta (FREIMOSER; HU; ST. LEGER, 2005), observa-se alteração generalizada da expressão da maioria dos parálogos da família Pr1 de proteases em relação ao controle (Figura 4).

Esses dados mostram conjuntamente que a expressão de serino-proteases da família Pr1 está relacionada a diferentes composições de cutícula dos hospedeiros artrópodes. Potencialmente, isso envolve um processo de "amostragem" do meio, afetando a especificidade de *M. anisopliae* (BEYS-DA-SILVA et al., 2014; FREIMOSER; HU; ST. LEGER, 2005; SANTI et al., 2010). Contudo, se o hospedeiro não possuir capacidade inibitória relativa a alguma protease em particular que esteja associada a um determinado patógeno, de tal maneira que não ocorra a sua inativação completa, esta se tornará o fator mais relevante. Isso se dá independentemente de sua concentração absoluta ou atividade, podendo ser notavelmente menores que as dos demais fatores, visto que a secreção de enzimas de cujos inibidores o hospedeiro dispõe em grandes quantidades acarreta no desperdício de recursos (VILCINSKAS, 2010).

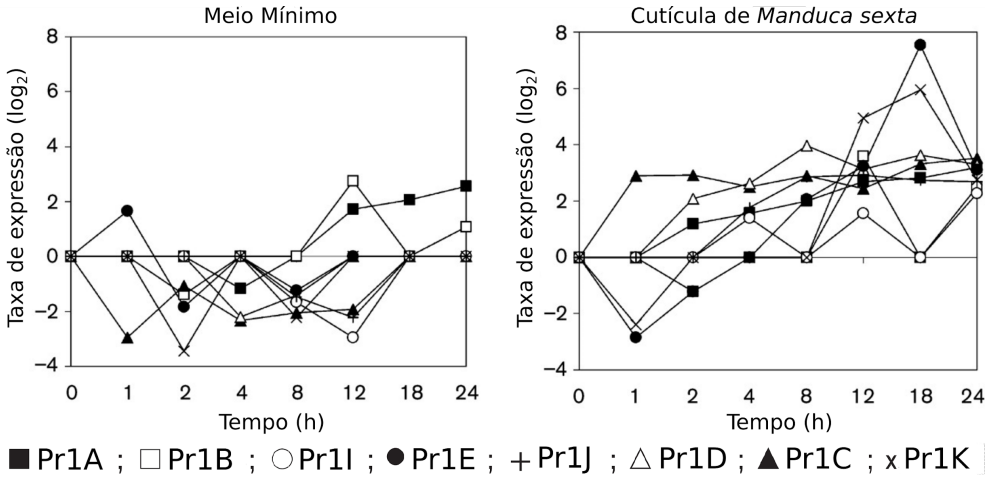


Figura 4: Diferenças na expressão de genes *pr1*. As taxas de expressão médias para cada gene, comparativamente entre meio mínimo (à esquerda) e em cutículas de *M. sexta* (à direita), demonstram a plasticidade da transcrição de acordo com o ambiente. Adaptado de FREIMOSER; HU; ST. LEGER, 2005.

Provavelmente a virulência de FAPs coevolui por resultado da seleção recíproca com o organismo-alvo. Sob esse paradigma, é possível postular que

exista seleção positiva na direção da evolução de novas proteases[21] que não são inativadas pelos fatores do hospedeiro (VILCINSKAS, 2010). Avaliações prévias a este trabalho da pressão seletiva na família Pr1 de proteases apontam que, calculando $\omega = d_N/d_S$ médio entre todos os códons, a divergência dessa família seguiu o que se espera de evolução neutra, com níveis variáveis de seleção purificadora (BAGGA et al., 2004; HU; ST. LEGER, 2004). Contudo, como os próprios autores apontam, ponderar ω entre todos os códons mascara as diferentes regiões da cadeia polipeptídica sob pressões seletivas diferenciadas. Mais recentemente, avaliações filogenômicas incluindo 7 espécies de *Metarhizium* [22] apontam Pr1K, apenas, sob seleção positiva em todas as linhagens (HU et al., 2014). Ademais, há evidências de seleção positiva atuando em serino-proteases de fungos aprisionadores de nematódeos (LI et al., 2010). Esse mesmo estudo aponta propriedades compartilhadas com FAPs no âmbito do parasitismo, inclusive demonstrando a capacidade de diversos entomopatógenos infectarem ovos de nematódeos. Da mesma forma, mostrou-se que fungos patogênicos a nematódeos são capazes de infectar artrópodes. Com isso em mente, um extensivo estudo comparativo entre as diversas proteoformas de serino-proteases tipo-subtilisina da família Pr1, incorporando espécies fora do complexo *Metarhizium*, pode evidenciar novos padrões de diversificação, contribuindo no entendimento do papel individual dessas enzimas no processo infectivo.

[21] Sejam novas famílias ou mesmo novos parálogos de famílias existentes.
[22] *M. acridum, M. album, M. majus, M. guizhouense, M. brunneum, M. robertsii e M. anisopliae*

Tese

Se a virulência fúngica coevolui com seus hospedeiros, havendo pressão seletiva positiva para o surgimento de novas variantes,
então essa pressão seletiva pode ser quantificada e alterações decorrentes desse processo causam efeitos conformacionais e funcionais observáveis e detectáveis na estrutura proteica.

Objetivos

Gerais: evidenciar padrões evolutivos de seleção positiva e divergência funcional na família Pr1 de proteases, bem como seus efeitos na estrutura proteica e função enzimática.

Específicos:

- Descrever a evolução molecular da família Pr1 de serino-proteases em *Metarhizium* spp.;
- Evidenciar regiões de variação e alteração estrutural e funcional da proteína;
- Verificar o papel de resíduos positivamente selecionados e associados à divergência funcional na estrutura e função de proteoformas de Pr1J;
- Caracterizar e diferenciar a função enzimática de proteoformas de Pr1J em *Metarhizium* spp.

Organização Geral

Doravante, o trabalho se divide em três capítulos, correspondendo à divisão de temas inerente ao estudo proposto. Neles serão apresentados resultados e metodologias pertinentes, seguidos de discussão conjunta das implicações das observações. A Figura 5 traz, de forma gráfica, o panorama do fluxo metodológico seguido. De forma resumida, são como segue:

1. *Evolução*

   Inferência filogenética baseada em sequências de proteases Pr1, com posterior aferição de pressão seletiva positiva e avaliação de divergência funcional entre as 10 proteoformas da Classe II. Por fim, discorre-se sobre as implicações desses resíduos em estrutura proteica homóloga. Será apresentado na forma de artigo, já publicado.

2. *Estrutura*

   Baseado nos resultados decorrentes do Capítulo 1, foi realizada a construção de modelos comparativos de estrutura proteica, seguidos de simulação por Dinâmica Molecular de proteoformas (potenciais) de Pr1J.

3. *Função*

   Complementar ao Capítulo 2, tentou-se realizar expressão heteróloga das Pr1J fúngicas, com posterior isolamento e caracterização da atividade enzimática em substratos conhecidos, visando observar diferenças funcionais entre as potenciais proteoformas de Pr1J. Este segmento encontrava-se em andamento até a pandemia do SARS-CoV2 e está interrompido por tempo indeterminado.

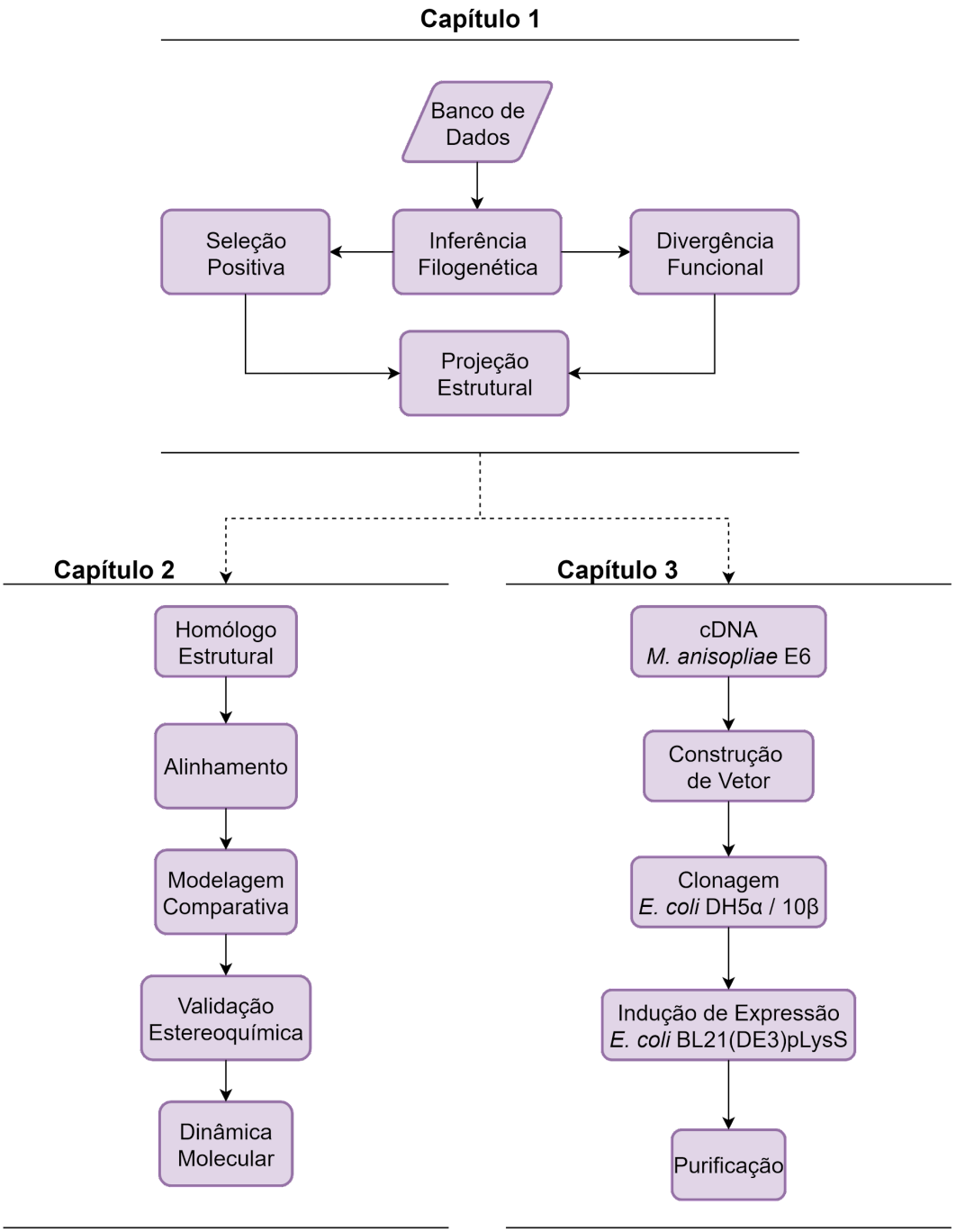


Figura 5: Organização Geral da Tese.

Capítulo 1: Evolução

O artigo que constitui essa sessão foi publicado no periódico Molecular Genetics and Genomics sob o DOI 10/c3zn. Neste Capítulo são descritas as primeiras análises evolutivas envolvendo a Classe II da família Pr1 de proteases em *M. anisopliae* e outras espécies relacionadas, incorporando informações existentes em bancos de dados públicos. Esse trabalho aborda aspectos evolutivos gerais dos 10 parálogos (A-K, exceto C), caracterizando as relações das entidades moleculares entre si, bem como entre as subfamílias, analisando-se as ramificações associadas ao gênero *Metarhizium*. A seguir, analisa-se a presença e localização de seleção positiva e divergência funcional tipo II. Verifica-se, então, a relação entre pressão seletiva e localização na estrutura proteica através de modelagem estrutural comparativa. Por fim, discorre-se sobre as implicações dos processos evolutivos sobre a função proteica e sobre a potencial existência de uma proteoforma não descrita de Pr1J.

Molecular Genetics and Genomics (2019) 294:901–917
https://doi.org/10.1007/s00438-019-01546-y

ORIGINAL ARTICLE

Molecular evolution of Pr1 proteases depicts ongoing diversification in *Metarhizium* spp

Fabio Carrer Andreis[1,2] · **Augusto Schrank**[1,2] · **Claudia Elizabeth Thompson**[1,2,3,4]

Received: 14 January 2019 / Accepted: 8 March 2019 / Published online: 28 March 2019

Abstract
The Pr1 family of serine endopeptidases plays an important role in pathogenicity and virulence of entomopathogens such as *Metarhizium anisopliae* (Ascomycota: Hypocreales). These virulence factors allow for the penetration of the host cuticle, a vital step in the infective process of this fungus, which possesses 11 Pr1 isoforms (Pr1A through Pr1K). The family is divided into two classes with Class II (proteinase K-like) comprising 10 isoforms further split into three subfamilies. It is believed that these isoforms act synergistically and with other virulence factors, allowing pathogenicity to multiple hosts. As virulence coevolves through reciprocal selection with hosts, positive selection may lead to the evolution of new protease families or isoforms of extant ones that can withstand host defenses. This work tests this hypothesis in Class II Pr1 proteins, focusing on *M. anisopliae*, employing different methods for phylogenetic inference in amino acid and nucleotide datasets in multiple arrangements for *Metarhizium* spp. and related species. Phylogenies depict groups that match the taxonomy of their respective organisms with high statistical support, with minor discrepancies. Positively selected sites were identified in six out of ten Pr1 isoforms, most of them located in the proteolytic domain and spatially close to the catalytic residues. Moreover, there was evidence of functional divergence in the majority of pairwise comparisons. These results imply the existence of differential selective pressure acting on Pr1 proteins and a potential new isoform, likely affecting host specificities, virulence, or even adapting the organism to different host-independent lifestyles.

Keywords *Metarhizium anisopliae* · Serine endopeptidases · Molecular evolution · Positive selection · Functional divergence

Communicated by Stefan Hohmann.

Augusto Schrank and Claudia Elizabeth Thompson share senior authorship.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-019-01546-y) contains supplementary material, which is available to authorized users.

✉ Claudia Elizabeth Thompson
cthompson@ufcspa.edu.br; thompson.ufcspa@gmail.com

[1] Rede Avançada em Biologia Computacional (RABICÓ), Petrópolis, RJ, Brazil

[2] Centro de Biotecnologia, Programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM), Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[3] Departamento de Farmacociências, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, RS, Brazil

[4] Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, RS, Brazil

Introduction

Entomopathogenic fungi have been used for biocontrol purposes for over a century. Organisms such as *Metarhizium anisopliae* and *Beauveria bassiana* (Ascomycota:Hypocreales) are the most frequently used worldwide mainly due to their efficient infection mechanisms (Faria and Wraight 2007; Schrank and Vainstein 2010; Sánchez-Pérez et al. 2014). Among these, *M. anisopliae* was the first to be mass produced and applied for such purposes (Zimmermann et al. 1995). Globally distributed, *M. anisopliae* can be isolated from soil, inhabiting plant rhizospheres (St Leger 2008) or as a saprobe on arthropod cadavers (Schrank and Vainstein 2010), while also being pathogenic to arthropods and capable of adopting an endophytic lifestyle (Barelli et al. 2016). Some of its hosts are of economic importance, such as the cattle tick *Rhipicephalus microplus* or the cotton-stainer bug *Dysdercus peruvianus* (Schrank and Vainstein 2010), while others are of major sanitary

importance, such as *Aedes aegypti*, the primary vector of Dengue, Chikungunya and Zika viruses (Greenfield et al. 2015). However, different species of *Metarhizium* display different host ranges, correlated with the genus evolution: specialists such as *M. acridum* and *M. album* are pathogenic, respectively, to the orders Orthoptera and Hemiptera; while transitional species (*M. guizhouense, M. majus*) are able to infect Lepidoptera and Coleoptera; and finally generalist representatives (*M.anisopliae, M. brunneum, M. robertsii*) can parasitize over seven arthropod orders. Specialist traits are considered plesiomorphic with respect to transitional and generalist characters, the latter being the most recent in the evolutionary chronology. This directional widening of host ranges is thought to have occurred alongside the diversification of the hosts, in a coevolutionary fashion, encompassing increases in genome size, number of genes, and number of protein families (Hu et al. 2014).

A successful infection requires that the conidium transposes the cuticle, the first and foremost layer of defense, while resisting a series of defense mechanisms employed by the hosts, including cuticle melanization and secretion involving fungistatic compounds, antimicrobial peptides, and protease inhibitors (Butt et al. 2016). After adhesion to the epicuticle, conidia develop structures named appressoria, which are responsible for exerting mechanical pressure and secreting an array of proteinases, chitinases, and lipases for the degradation of the protein-chitin exoskeleton (Schrank and Vainstein 2010; Butt et al. 2016).

M. anisopliae synthesizes at least three different kinds of proteases in the infective process: subtilisin-like serine proteases of the Pr1 family, trypsin-like serine proteases of the Pr2 family, and a metalloprotease (St Leger et al. 1994; Schrank and Vainstein 2010). *M anisopliae* possesses 11 known Pr1 isoforms (Pr1A through K), the highest number ever observed in fungi (Freimoser et al. 2003), which are divided in Class I (bacterial-like), containing exclusively Pr1C, and Class II (proteinase K-like), comprising the remaining isoforms further divided into three subfamilies. The extracellular subfamily 1 (sf1) contains Pr1s A, B, G, I, and K; the extracellular subfamily 2 (sf2) comprises Pr1s D, E, F, and J; and finally Pr1H composes subfamily 3 (sf3), which is the only endocellular Pr1 (Bagga et al. 2004). This family of proteases has been widely associated with fungal virulence (Rosas-García et al. 2014) and their overall expression levels increase during the initial and later stages of infection in cuticle models, namely through cuticle penetration and conidiation, although expression patterns for each isoform change accordingly to hosts (Freimoser et al. 2005; Small and Bidochka 2005; Beys-da-Silva et al. 2014; Javar et al. 2015). While the individual functions of Pr1 isoforms are not fully understood, it is presumed that they act synergistically with other hydrolytic enzymes for an efficient degradation of the cuticle (Butt et al. 2016). Their different expression patterns suggest an environment sampling mechanism with direct effects on host specificity (Santi et al. 2010).

In this work, we sought to unveil the patterns of selection in the molecular evolution of the Pr1 family of proteases regarding *M. anisopliae* and related species. Presumably, virulence coevolves through reciprocal selection with the host and positive selection is responsible for the evolution of new proteases or isoforms capable of withstanding the host defense mechanisms, namely protease inhibitors (Vilcinskas 2010). In this sense, we performed a thorough phylogenetic evaluation of these proteases, focusing on positively selected sites and functional divergence, and evidenced that six out of ten Class II Pr1s are under positive selection, some on functionally-related regions, with many residues involved in the functional divergence of these proteins. Our findings might contribute to the understanding of the intricate interaction of fungal pathogens and their arthropod hosts by means of differential selective pressures on virulence factors.

Materials and methods

Database construction

A local database of subtilisin amino acid sequences was constructed using the *hmmsearch* software of the online HMMER platform (Eddy 1998; Finn et al. 2011) based on the Peptidase_S8 HMM profile (PF00082) downloaded from the Pfam database (Finn et al. 2014). We realized searches in the Reference Proteomes database applying significance *E* value thresholds of 0.01 and 0.03 for sequences and hits, respectively. In total, our database comprised 35,560 sequences containing the subtilase family proteolytic domain. We then searched this database for homologous sequences to each Pr1 contained on the *Metarhizium anisopliae* E6 genome (Staats et al. 2014) (Table 1) using the *blastp* software from the BLAST + package (Camacho et al. 2009) with the following parameter thresholds: $E \leq 10^{-06}$, identity $\geq 60\%$, and cover $\geq 60\%$. The resulting sequences were filtered by size, removing sequences 40% larger or smaller than the average size of the references (Table 1). We then retrieved the corresponding nucleotide sequences from the NCBI database and clustered identical entries into a single representative sequence. *Metarhizium* spp. sequences from Hu and coworkers (2014) were maintained for comparison.

Phylogenetic analysis

Amino acid sequence alignments were performed for Pr1 A, B, D, E, F, G, H, I, J, and K, as well as for joint datasets for sf1, 2, or 3 with PRANK (Löytynoja 2014) and were

Table 1 *blastp* and filtering results, evolutionary model analyses and selected phylogenies for different Pr1 datasets

Subfamily	Isoform	Accession	Size (aa)	#BLAST+	#Filtered	Model (aa)	Model (nt)	Selected tree
Sf1	**A**	KFG86683.1	390	100	33[a]	LG + I + G + F[b]	GTR + I + G[c]	nt-BI
	B	KFG82971.1	386	33	13	JTT + G[d]	HKY + I[e]	nt-BI
	G	KFG79277.1	399	9	7	JTT + I[d]	HKY + G[e]	nt-BI
	I	KFG81480.1	388	97	12	WAG + G[f]	GTR + I[c]	nt-BI
	K	KFG84128.1	391	32	26	WAG + I + G[f]	GTR + I + G[c]	nt-BI
	Σ	N/A	N/A	N/A	91	LG + I + G[b]	GTR + I + G[c]	nt-BI
Sf2	**D**	KFG83000.1	406	14	9	JTT + G[d]	HKY + G[e]	nt-BI
	E	KFG79137.1 KFG81922.1 KFG84469.1	386[g]	34	16	JTT + G[d]	GTR + G[c]	nt-ML
	F	KFG79136.1 KFG83510.1	329[g]	24	13	WAG + G[f]	GTR + G[c]	aa-BI
	J	KFG80219.1 KFG85392.1	398[g]	23	16	WAG + G[f]	GTR + I + G[c]	nt-ML
	Σ	N/A	N/A	N/A	54	WAG + I + G[f]	GTR + I + G[c]	nt-BI
Sf3	**H**	KFG81188.1	533	179	118[a]	LG + I + G[b]	GTR + I + G[c]	nt-BI
Global		N/A	N/A	N/A	263	LG + I + G[b]	GTR + I + G[c]	nt-BI

Σ union of all sequences belonging0020to the subfamily, *N/A* Non-Applicable

[a]After additional filtering at 95% identity

[b]Le and Gascuel 2008

[c]Tavaré 1986

[d]Jones et al. 1992

[e]Hasegawa et al. 1985

[f]Whelan and Goldman 2001

[g]average of the sizes of the references

manually inspected and edited (when deemed necessary) with AliView (Larsson 2014). For the global sf123 dataset we employed GUIDANCE v2 (Sela et al. 2015) to assess alignment reliability, using the PRANK (Löytynoja 2014) algorithm for alignment with 100 bootstrap replicates and variable gap penalties, the GUIDANCE2 (Landan and Graur 2008) method for confidence measurements with a score cutoff of 0.70. The corresponding nucleotide alignment was constructed with TranslatorX (Abascal et al. 2010). The best-fit amino acid or nucleotide evolutionary model was estimated using ProtTest v3.2 (Guindon and Gascuel 2003; Abascal et al. 2005; Darriba et al. 2011) and JModelTest v2 (Guindon and Gascuel 2003; Darriba et al. 2012), respectively. Phylogenetic reconstruction through Maximum Likelihood (ML) was performed using PhyML v3 (Guindon and Gascuel 2003) with 1000 bootstrap replicates. Additional parameters include the estimation of gamma distribution (+G), proportion of invariable sites (+I), observed residue frequencies (+F) when applicable (Abascal et al. 2005), 4 substitution rate categories, optimization of tree topologies, branch lengths, and substitution model parameters. Starting tree construction uses the BioNJ algorithm (Gascuel 1997), and tree topology search using nearest-neighbor interexchange. Phylogenetic reconstruction by Bayesian Inference was performed using MrBayes v3.2.5 (Altekar et al. 2004; Ayres et al. 2012; Ronquist et al. 2012), sampling trees for at most 10^7 generations, except for Pr1H-aa-BI that required further sampling for convergence. The sampling occurs every hundredth generation, with a stopping value at a 0.01 average standard deviation of split frequencies threshold, discarding the 25% initial samples as burn-in, and summarizing topologies and parameters afterwards. All obtained phylogenies had their branches collapsed at an 800 bootstrap or with the use of a 0.8 posterior probability threshold using the TreeGraph v2 software (Stöver and Müller 2010). A meta-tree was constructed using the four collapsed trees (amino acid and nucleotide datasets for both ML and BI) with the MetaTree software (Nye 2008). The generated consensus topology (centermost node of the meta-tree) and the phylogeny displaying fewer polytomies were used for positive selection analyses. Tree manipulation softwares include FigTree v1.4.2 (Rambaut 2016) and SeaView v4 (Gouy et al. 2010).

Positive selection

In order to estimate positively selected sites on the individual Pr1 coding nucleotide sequences we used the *codeml*

software of the PAML package (Yang 2007), following the procedures depicted on the manual, for the MetaTree consensus topology and selected tree alike. First, we calculated branch lengths according to the M0 (ω) model, which were used as input for d_N/d_S calculations according to the M1a (nearly neutral), M2a (positive selection), M3 (discrete), M7 (β) and M8 (β&ω) likelihood (*l*) models. Likelihood Ratio Tests (LRTs; 2Δln *l*) were conducted comparing nested models M0 vs M3, M1a vs M2a, and M7 vs M8 following χ^2 tests with 4, 2, and 2 degrees of freedom, respectively, as recommended by Anisimova et al. (2002). For rejection of the null hypothesis of neutrality, we considered $p \leq 0.05$, which was calculated using the *chi2* software of PAML. The probability of positively selected sites was calculated according to the Naïve Empirical Bayes (NEB) and Bayes Empirical Bayes (BEB) methods implemented in PAML, where we only considered sites with posterior probabilities of 0.95 or higher as positively selected.

Functional divergence

Type I functional divergence evaluates altered functional constraints in one or more residues when comparing two genes, where any given site is conserved in one gene and highly variable in other, regardless of the underlying evolutionary mechanisms (Gu et al. 2006). We employed DIVERGE v3.0 software (Gu et al. 2013) to analyze functional divergence of the subfamilies 1 and 2. The Gu99 method (Gu 1999) was applied to each subfamily alignment and extracted subtrees from the global phylogeny due to better resolution (in most scenarios) when compared against their joint phylogenies. Polytomies in these topologies were represented as zero-length branches. For each dataset, we selected monophyletic clades corresponding to each isoform (labeled A, B, G, I, and K for sf1; D, E, F, and J for sf2) for ML pairwise estimation of type I functional divergence coefficient (θ_I), gamma parameter for among-site rate variation (α), standard error of the estimate (SE_θ), and LRT for θ_I. The likelihood ratio test yields a log-score approximately following a χ^2 distribution with 1 *df* and is constructed under the null hypothesis of no functional divergence ($\theta_I = 0$) against the existence of functional divergence ($\theta_I > 0$). We considered $p \leq 0.05$, also calculated using the *chi2* software (see previous section), for rejection of the null hypothesis. In this case, a successful test means that functional constraints have shifted between two homologous genes (Gu 2001).

Structural projection

In order to pinpoint the location of positively selected sites, we projected the location of residues onto an experimentally determined homologous structure. Thus for each Pr1 isoform a generalist, transitional, and specialist *Metarhizium* representative was selected (*M. anisopliae* E6, *M. guizhouense* ARSEF 977, and *M. acridum* CQMa 102, respectively). An online *phmmer* (Eddy 1998; Finn et al. 2011) search within the Protein Data Bank (PDB; Berman et al. 2003, 2007) was then conducted for each sequence at default settings to find a reliable homologous structure. Criteria for structure selection followed best-practices for template selection in template-based modeling (Khan et al. 2016): over 30% alignment identity and the best possible resolution. Afterwards, a structure-based multiple sequence alignment was performed using the online PROMALS3D (Pei et al. 2008) software with default parameters on our representative sequences plus a structural homolog. Amino acid sites under positive selection were located in this alignment and their locations highlighted in the homologous position of the crystallographic structure using PyMOL v2.2.0 (Schrödinger 2015).

Imaging software

Final figures were constructed using Inkscape v0.92 (http://www.inkscape.org) and converted using GIMP v2.8.14 (http://www.gimp.org).

Results

Filtering, alignment, and evolutionary model selection

Sequences corresponding to each Class II Pr1 from *M. anisopliae* E6 were used as queries against our local protein database using *blastp* (Camacho et al. 2009). Initially 545 sequences were recovered. This number was reduced to approximately 65% (359) through identity and size filtering. Our resulting sequences for isoforms A, B, and I (sf1) displayed some intersections, solved by constructing preliminary phylogenies and analyzing the resulting clusters. Additionally, due to over-representation of some taxonomic groups (especially *Metarhizium* spp. and *Fusarium* spp.), a representative sequence was chosen for groups displaying ≥ 95% sequence identity for Pr1A and Pr1H (*Metarhizium* spp. sequences from Hu and coworkers (2014) were maintained for comparison, regardless of identity). This results in a final number of 263 sequences (roughly 48% of the initial) (Table 1). Moreover, assessing the reliability of our global alignment of Class II isoforms rendered a reduction of 76.55% in alignment size (1966 to 461 sites) (Online Resource 1–13 and 1–14).

Isoforms E, F, and J exist in multiplicity in the genome of *M. anisopliae* E6 (3, 2, and 2, respectively), as well as in some related species. For Pr1E there are three identified sequences for *M. brunneum* ARSEF 3297, three for

M. robertsii ARSEF 23, and two sequences for *M. robertsii* ARSEF 2575. In a similar fashion for Pr1F we obtained two sequences belonging to *M. anisopliae* ARSEF 549, two for *M. brunneum* ARSEF 3297 and two for both lineages ARSEF 23 and ARSEF 2575 of *M. robertsii*. Additionally, many species displayed two sequences for isoform J, namely *M. brunneum* ARSEF 3297, *M. guizhouense* ARSEF 977, *M. acridum* CQMa 102, and to both *M. robertsii* ARSEF 23 and ARSEF 2575. Finally, there were two annotated Pr1H sequences for *M. anisopliae* E6, KFG81188.1 and KFG83511.1, although the latter displayed no homologs in our database and was thus discarded.

All sequence accessions for data used in this work, along with their taxonomic classifications, can be found in Online Resource 2. Results for all evolutionary model analyses based either on our amino acid alignments (Online Resource 1) or on our retroalignments are summarized in Table 1.

Phylogenetic reconstruction

Selected phylogenies for each dataset are listed in Table 1.

Subfamily 1

In order to better assess the evolutionary history of Pr1 proteases we constructed a joint alignment, comprising all sequences contained in sf1 (Online Resource 1–10). However, we could not unequivocally define relationships among them. In Sf1-nt-Bi (Fig. 1a) and Sf1-MT (Fig. 1b), we observed a separation of representatives for each isoenzyme in a [K, (A, B, G, I)] pattern, although some of these groups appear fragmented. Pr1A members split into four groups in Sf1-nt-BI: A1 (including *Metarhizium, Pochonia, Metacordyceps, Epichloë, Claviceps, Beauveria, Lecanicillium, Fusarium, Engyodontium, Ophiocordyceps, Hirsutella*, and *Purpureocillium* genera; corresponding to sets A1.1, A1.2, and A1.3 in Sf1-MT), A2 (*Tolypocladium;* identical in Sf1-MT), A3 (*Eutypa, Cordyceps*, and *Sarocladium*; A3.1 and A3.2 in Sf1-Mt), and A4 (*Trichoderma;* identical in Sf1-MT). Even though group K is depicted as monophyletic in Sf1-nt-BI, it is branched into four groups in the meta-tree: K.1 (including *Metarhizium, Epichloë, Claviceps, Villosiclava*, and *Trichoderma* genera), K.2 (*Fusarium* and *Nectria*), K.3 (*Acremonium*), and K.4 (*Colletotrichum*). The remaining groups (B, G and I) composed similar monophyletic clades in both topologies. There were also no major differences regarding clustering for OTUs of each isoform in comparison to individual datasets and the observed subgroups in Fig. 1b correspond in most cases to existing branches in the individual trees (Online Resource 3–1 and 3–2). It is noteworthy that host-specialist species (*M. acridum* and *M. album*) in the sf1 phylogeny appear as sisters to host-generalist (*M. anisopliae, M. brunneum*, and *M. robertsii*) and transitional species (*M. guizhouense* and *M. majus*) in B, G, I, and K clades. In clade A, however, even though Pr1s belonging to specialists are depicted as sisters to the others, there is a mix of generalist and transitional OTUs in a high posterior probability clade.

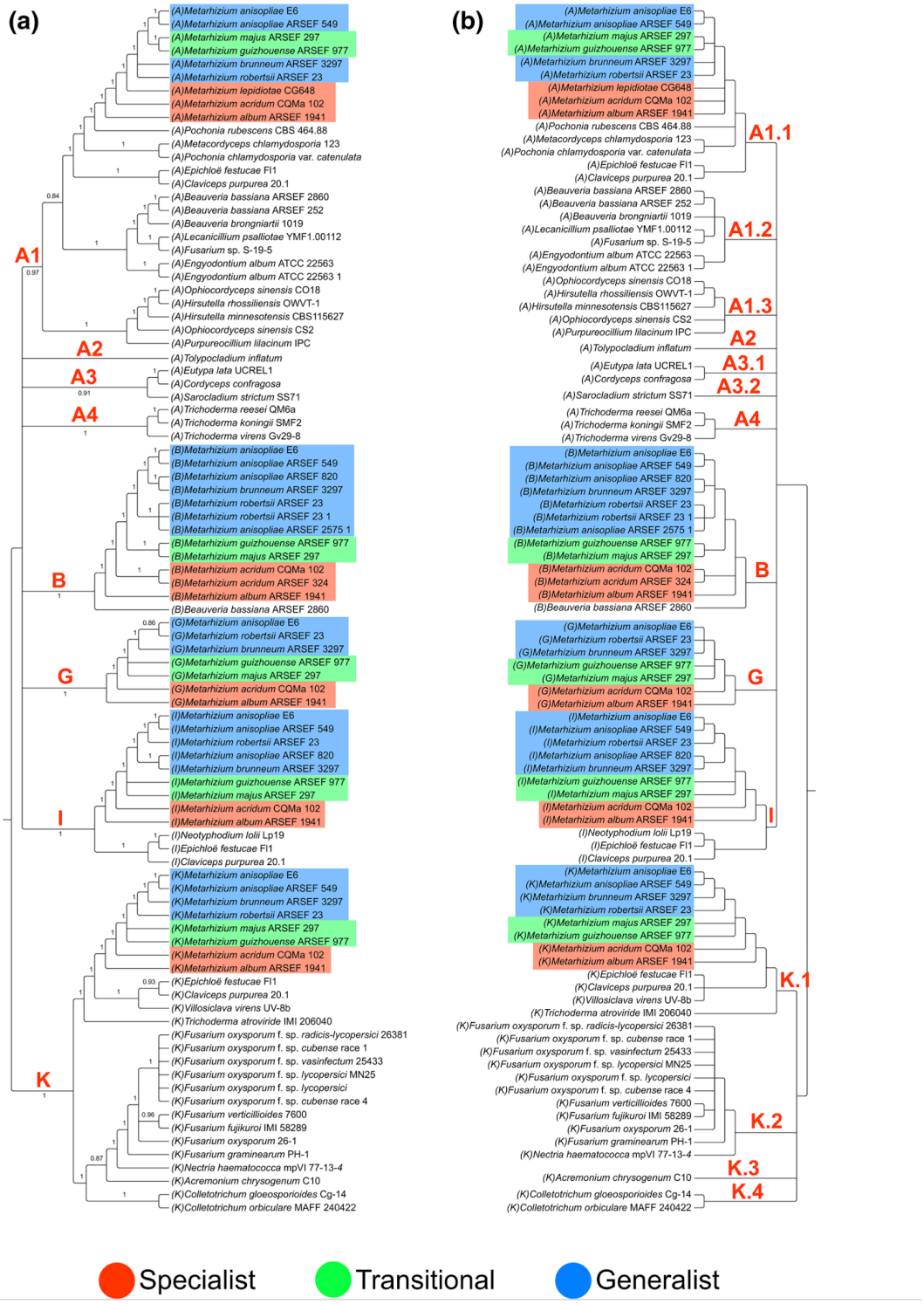
A similar trend was also observed in the proposed individual reconstructions (Online Resource 3). In general, phylogenies depicted similar groupings when comparing the MetaTree for each dataset to its selected topology, although with some different inter-relationships. For Pr1B-MT specifically, no reasonable clusters were observed, meaning that there was no substantial agreement among all our individual phylogenetic reconstructions for the Pr1B dataset (Online Resource 3–1). Overall, the majority of operational taxonomical units (OTUs) belonged to the Hypocreales order and were clustered according to their families in our individual analyses.

Subfamily 2

Jointly, phylogenetic analyses for sf2 isoforms (Fig. 2) displayed concordant topologies regarding large groups. We could verify in both Sf2-nt-BI (Fig. 2a) and Sf2-MT (Fig. 2b) the (J, (D, (E, F))) clustering pattern, there being an identifiable specialist-transitional-generalist trend for *Metarhizium* spp. inside all groups. The OTU composition for this subfamily contained only *Metarhizium* spp. and *Epichloë festucae* (Pr1J only), which are representatives from the Hypocreales order and Clavicipitaceae family. In general, all branchings for monophyletic clades correspond to their individual counterparts. Due to Pr1D inferences (Online Resource-3) depicting two groups of two nearly identical topologies, differing by a single branch, no consensus topology could be inferred for this dataset. Trees belonging to Pr1E (Online Resource 3–3) branched into three main groups: groups 1 and 2 were composed exclusively by generalist species, while group 3 contained specialist, transitional, and generalist representatives, there being a closer relationship among non-specialists. Much like Pr1E, resulting phylogenies for our Pr1F analyses (Online Resource 3–4) displayed multiple groups, one following the specialist-transitional-generalist pattern and the other being a heterogeneous cluster composed of one specialist and five generalist sequences. Finally, regarding Pr1J (Online Resource 3–4) we observed two monophyletic groups following the overall trend for *Metarhizium* spp., with *E. festucae* placed in between.

Subfamily 3/Pr1H

The sf3 dataset was the largest and most taxonomically diverse of all our data, encompassing a different phylum and several classes (Fig. 3a). The Eurotiomycetes and


(a)
(A)Metarhizium anisopliae E6
(A)Metarhizium anisopliae ARSEF 549
(A)Metarhizium majus ARSEF 297
(A)Metarhizium guizhouense ARSEF 977
(A)Metarhizium brunneum ARSEF 3297
(A)Metarhizium robertsii ARSEF 23
(A)Metarhizium lepidiotae CG648
(A)Metarhizium acridum CQMa 102
(A)Metarhizium album ARSEF 1941
(A)Pochonia rubescens CBS 464.88
(A)Metacordyceps chlamydosporia 123
(A)Pochonia chlamydosporia var. catenulata
(A)Epichloë festucae FI1
(A)Claviceps purpurea 20.1
(A)Beauveria bassiana ARSEF 2860
(A)Beauveria bassiana ARSEF 252
(A)Beauveria brongniartii 1019
(A)Lecanicillium psalliotae YMF1.00112
(A)Fusarium sp. S-19-5
(A)Engyodontium album ATCC 22563
(A)Engyodontium album ATCC 22563 1
(A)Ophiocordyceps sinensis CO18
(A)Hirsutella rhossiliensis OWVT-1
(A)Hirsutella minnesotensis CBS115627
(A)Ophiocordyceps sinensis CS2
(A)Purpureocillium lilacinum IPC
(A)Tolypocladium inflatum
(A)Eutypa lata UCREL1
(A)Cordyceps confragosa
(A)Saroclaadium strictum SS71
(A)Trichoderma reesei QM6a
(A)Trichoderma koningii SMF2
(A)Trichoderma virens Gv29-8
(B)Metarhizium anisopliae E6
(B)Metarhizium anisopliae ARSEF 549
(B)Metarhizium anisopliae ARSEF 820
(B)Metarhizium brunneum ARSEF 3297
(B)Metarhizium robertsii ARSEF 23
(B)Metarhizium robertsii ARSEF 23 1
(B)Metarhizium anisopliae ARSEF 2575 1
(B)Metarhizium guizhouense ARSEF 977
(B)Metarhizium majus ARSEF 297
(B)Metarhizium acridum CQMa 102
(B)Metarhizium acridum ARSEF 324
(B)Metarhizium album ARSEF 1941
(B)Beauveria bassiana ARSEF 2860
(G)Metarhizium anisopliae E6
(G)Metarhizium robertsii ARSEF 23
(G)Metarhizium brunneum ARSEF 3297
(G)Metarhizium guizhouense ARSEF 977
(G)Metarhizium majus ARSEF 297
(G)Metarhizium acridum CQMa 102
(G)Metarhizium album ARSEF 1941
(I)Metarhizium anisopliae E6
(I)Metarhizium anisopliae ARSEF 549
(I)Metarhizium robertsii ARSEF 23
(I)Metarhizium anisopliae ARSEF 820
(I)Metarhizium brunneum ARSEF 3297
(I)Metarhizium guizhouense ARSEF 977
(I)Metarhizium majus ARSEF 297
(I)Metarhizium acridum CQMa 102
(I)Metarhizium album ARSEF 1941
(I)Neotyphodium lolii Lp19
(I)Epichloë festucae FI1
(I)Claviceps purpurea 20.1
(K)Metarhizium anisopliae E6
(K)Metarhizium anisopliae ARSEF 549
(K)Metarhizium brunneum ARSEF 3297
(K)Metarhizium robertsii ARSEF 23
(K)Metarhizium majus ARSEF 297
(K)Metarhizium guizhouense ARSEF 977
(K)Metarhizium acridum CQMa 102
(K)Metarhizium album ARSEF 1941
(K)Epichloë festucae FI1
(K)Claviceps purpurea 20.1
(K)Villosiclava virens UV-8b
(K)Trichoderma atroviride IMI 206040
(K)Fusarium oxysporum f. sp. radicis-lycopersici 26381
(K)Fusarium oxysporum f. sp. cubense race 1
(K)Fusarium oxysporum f. sp. vasinfectum 25433
(K)Fusarium oxysporum f. sp. lycopersici MN25
(K)Fusarium oxysporum f. sp. lycopersici
(K)Fusarium oxysporum f. sp. cubense race 4
(K)Fusarium verticillioides 7600
(K)Fusarium fujikuroi IMI 58289
(K)Fusarium oxysporum 26-1
(K)Fusarium graminearum PH-1
(K)Nectria haematococca mpVI 77-13-4
(K)Acremonium chrysogenum C10
(K)Colletotrichum gloeosporioides Cg-14
(K)Colletotrichum orbiculare MAFF 240422
A1
A2
A3
A4
B
G
I
K
(b)
A1.1
A1.2
A1.3
A2
A3.1
A3.2
A4
B
G
I
K.1
K.2
K.3
K.4
Specialist
Transitional
Generalist

◂**Fig. 1 Subfamily 1 joint phylogenetic reconstructions. a** Phylogenetic reconstruction using Bayesian Inference with the nucleotide dataset, following the GTR + I + G evolutionary model, and sampling 364,000 generations. Estimated posterior probabilities are adjacent to their corresponding branches; **b** Meta-tree combining all constructed topologies (Maximum Likelihood and Bayesian Inference methods using both amino acid and nucleotide datasets). Component groups are identified in red labels above their corresponding branches. Host range information for *Metarhizium* spp. is depicted in red, blue, and green for specialist, generalist, or transitional OTUs, respectively

Dothideomycetes classes formed sister groups, both of which composed a sister clade to Leotiomycetes. Sordariomycetes, which included *Metarhizium* spp., was inferred to be a sister group to the [Leotiomycetes, (Eurotiomycetes, Dothideomycetes)] clade. Moreover, the Saccharomycetes class shared a basal node with [Sordariomycetes, (Leotiomycetes, (Eurotiomycetes, Dothideomycetes))]. Due to our rooting, a member of the Basidiomycota phylum (*Rhodotorula mucilaginosa*) was positioned within Saccharomycetes, which was labeled as an outlier (Online Resource 3–5). Sf3-MT (Fig. 3b), comparatively, depicted the same relationships regarding Saccharomycetes, Sordariomycetes, and Leotiomycetes, although the latter was polytomically split into two groups. Also, Eurotiomycetes and Dothideomycetes were positioned at the same level and both are divided into two groups. Overall, higher taxonomical orders were also grouped in Sf3-nt-BI (Online Resource 3–5).

Groupings within the Hypocreales clade were organized according to taxonomical families. In Sf3-nt-BI, Cordycipitaceae OTUs (*Beauveria, Cordyceps*, and *Isaria* genera) shared a basal node with a polytomic group composed of Nectriaceae (*Fusarium* and *Nectria*), Hypocreaceae (*Trichoderma*), Ophiocordycipitaceae (*Ophiocordyceps*), and Clavicipitaceae [*Metarhizium, Epichloë, Claviceps*, and *Ustilaginoidea* (outlier)], the last two being sister clades. Differences observed in our Sf3 meta-tree included the loss of a tree level, polytomically positioning Cordycipitaceae with the remaining families, while maintaining the (Ophiocordycipitaceae, Clavicipitaceae) clade. Just as observed for the other Pr1 isoforms, *Metarhizium* spp. were grouped as expected regarding host ranges in Sf3-nt-BI and Sf3-MT alike.

Global analysis

A joint phylogenetic analysis of all Class II Pr1 serine proteases, employing only the most reliable alignment regions (estimated with GUIDANCE), allowed us to observe the relationships at the subfamily and isoform levels in a highly-supported tree. According to Li and coworkers (2010), sf3 (Clade E in their work) is the earliest diverging group of subtilisin-like serine proteases so that our trees were rooted in the Sf3 clade. Figure 4 displays three monophyletic clades corresponding to Sf1, Sf2, and Sf3, the first two depicted as sister groups and sharing a basal node with Sf3. This pattern was observed for the BI topology constructed with the nucleotide sequence alignment (Sf123-nt-BI; Fig. 4a) and meta-tree (Sf123-MT; Fig. 4b) alike.

Outer branches of Sf123-nt-BI supported the monophyly of each isoform (posterior probabilities ≥ 0.98), rendering the evolution of Class II subtilisin-like serine proteases as (H, [(J, (D, (E, F))), (K, (A, (B, (G, I))))))]. Isoform branchings were identical in Sf123-MT, with the exception of Sf1, which polytomically related all members of this subfamily and split the equivalent Pr1A clade of Sf123-nt-BI into eight distinct branches, corresponding to highly supported ramifications in the selected tree (Online Resource 3–6). Additionally, observed branchings in Sf123-nt-BI were roughly the same to the individual or joint subfamily analyses for sf1 and sf2, and similar at the class level for sf3.

Positive selection (PS)

The phylogenetic trees obtained for meta-tree (MT) and selected topology (Sel) alike were employed in positive selection analyses using the PAML package. These, along with retroalignments for each respective dataset, allowed for a statistical measurement of selective pressures under which each coding site is subjected by means of LRTs (Likelihood Ratio Test; using ω as descriptor). We point out that M0 vs M3 tests evaluate ω heterogeneity among sites, not positive selection, and this test displayed $p < 0.05$ for all datasets. Log-likelihood values, LRTs, average ω, estimated parameters, and positively selected sites (when applicable) are displayed in individual tables for each analysis in Online Resource 4. The following nomenclatures will be employed: Pr1X-PS-MT (when the meta-tree was used) and Pr1X-PS-Sel (selected topology), where X designates the corresponding isoform. Table 2 depicts a qualitative framework of statistic test results, while Fig. 5 summarizes identified positively selected sites and their respective locations according to typical protein architecture.

Positive selection was identified in four out of five members of sf1 (Pr1A, B, G, and I), employing MT and selected topologies alike. There was an overall convergence among identified sites for Pr1A-PS-MT and Pr1A-PS-Sel. Sites 133, 233, and 391 corresponded to positions N126, G216, and K353, respectively, in *M. anisopliae* E6, located in the proteolytic domain (S8). For Pr1B we detected positive selection in site 76 in Pr1B-PS-MT, but not in Pr1B-PS-Sel, corresponding to position 16A in our reference sequence, located in the signal peptide region. Results for Pr1G-PS-MT and Pr1G-PS-Sel concurred, displaying site 211 (position 211R, S8 domain) as positively selected. Analyses for Pr1I point towards sites 231, 328, and 379 for both Pr1I-PS-MT and Pr1I-PS-Sel, and 233 and 328 for Pr1I-PS-MT exclusively. These sites corresponded, in ascending order, to positions

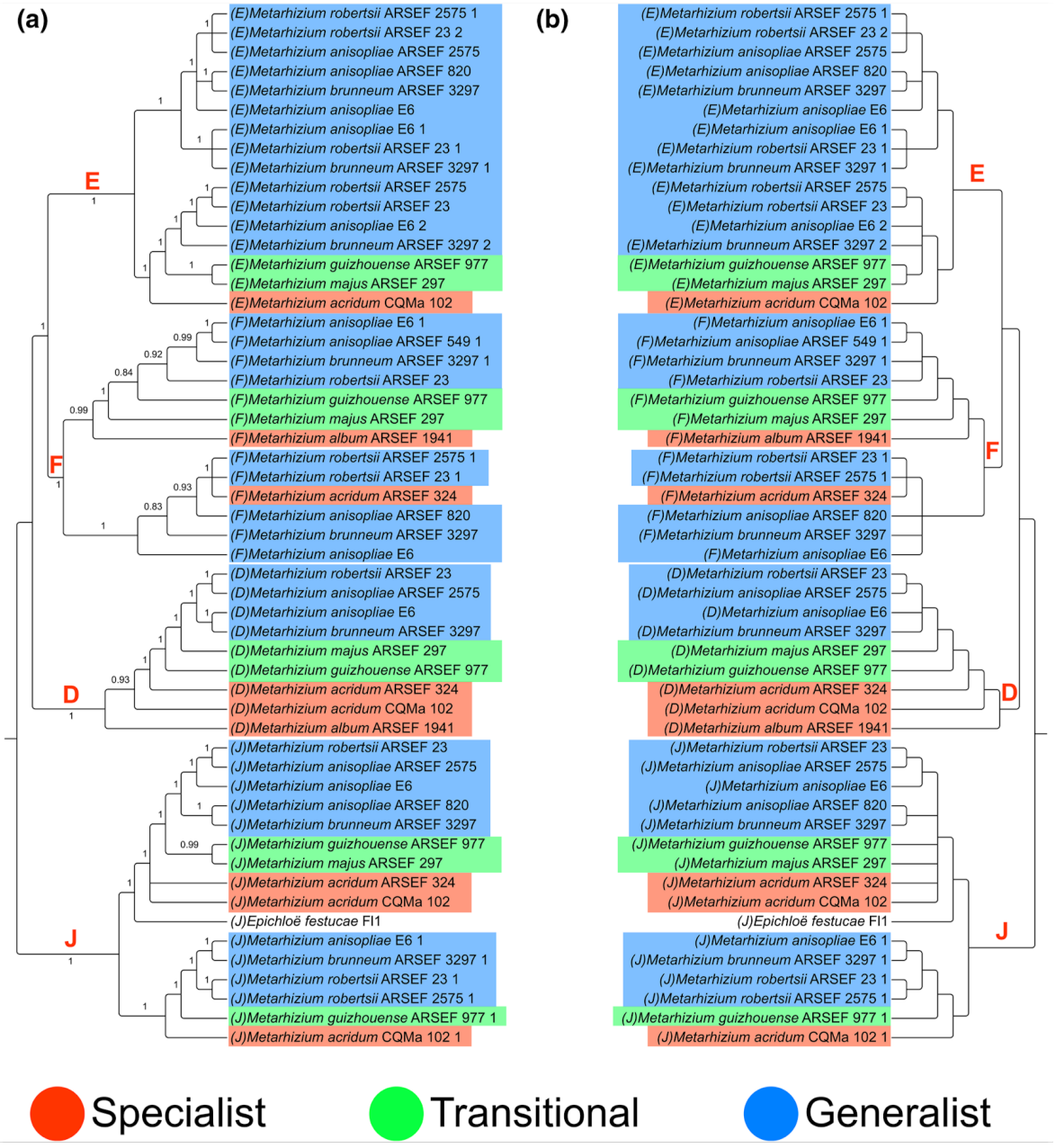


Fig. 2 Subfamily 2 joint phylogenetic reconstructions. a Phylogenetic reconstruction using Bayesian Inference with the nucleotide dataset, following the GTR + I + G evolutionary model, sampling 152,000 generations. Estimated posterior probabilities are adjacent to their corresponding branches; **b** Meta-tree combining all constructed topologies (Maximum Likelihood and Bayesian Inference methods using both amino acid and nucleotide datasets). Component groups are identified in red labels above their corresponding branches. Host range information for *Metarhizium* spp. is depicted in red, blue, and green for specialist, generalist, or transitional OTUs, respectively

E227, S229, T271, A323 (S8 domain), and A374 (C-terminal portion without associated domains) in *M. anisopliae* E6. Positive selection was not identified for isoform K.

Two out of four subfamily 2 members appeared to display positive selection. For Pr1D, lacking a metatree, sites 223, 291, and 317 were identified, corresponding to positions

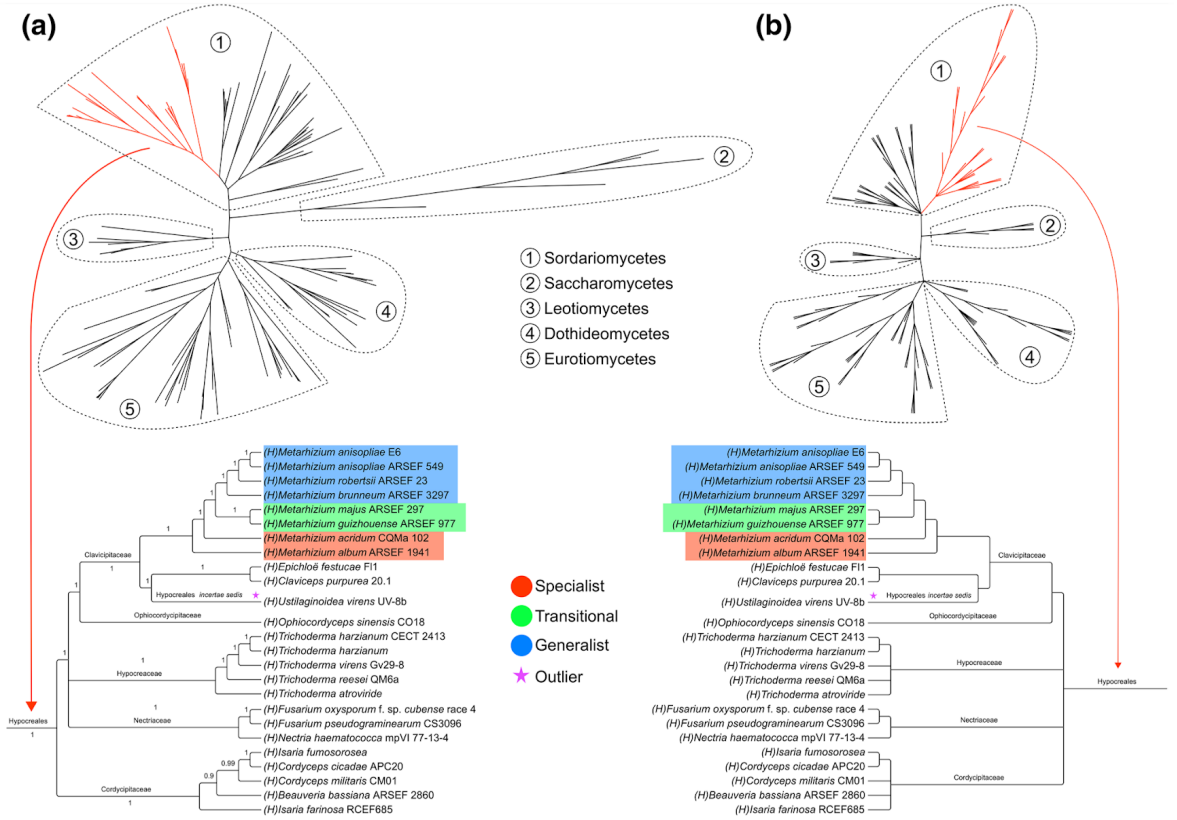


Fig. 3 Subfamily 3/Pr1H phylogenetic reconstructions. a Phylogenetic reconstruction using Bayesian Inference with the nucleotide dataset, following the GTR+I+G evolutionary model, sampling 296,000 generations. Estimated posterior probabilities are adjacent to their corresponding branches; **b** Meta-tree combining all constructed topologies (Maximum Likelihood and Bayesian Inference methods using both amino acid and nucleotide datasets). Taxonomical phylum, class, order, and family information for sampled OTUs are described above each branch, also valid for upper levels. Host range information for *Metarhizium* spp. is depicted in red, blue, and green for specialist, generalist, or transitional OTUs, respectively

Q218, V286, and L321 in *M. anisopliae* E6, located in the S8 domain. Regarding Pr1J, Pr1J-PS-MT and Pr1J-PS-Sel displayed the same positively selected sites: 364, 365, and 367 (notation: copy1/copy2 in *M. anisopliae* E6; positions K337/T337, T338/V338, and S340/T340, all in the proteolytic domain). Data for Pr1E were conflicting: we rejected the null hypothesis for Pr1E-PS-Sel in the M1a vs M2a test, but not for test M7 vs M8, while not detecting positive selection for Pr1E-PS-MT at all. Although above our statistical threshold, we were unable to locate sites with above 0.50 posterior probabilities, suggesting it to be a false positive. We did not observe positive selection for isoform F.

Our data did not point toward positive selection in Subfamily 3/Pr1H. Neither Pr1H-PS-MT nor Pr1H-PS-Sel displayed statistical significance, having LRTs of 0.00 ($p > 0.99$) and 0.01 ($p > 0.99$) in tests M1a vs M2a and M7 vs M8, respectively, for both datasets. It is worth noticing that both LRTs and p values were rounded to two decimal places, rendering values for Pr1H-PS-Sel and Pr1H-PS-MT appearing to be identical, while, in reality, differing on scales as low as 10^{-6}.

Functional divergence (FD)

Estimation of site-specific rate difference due to Type I functional divergence (T1-FD) in subfamilies 1 and 2 suggested different functional constraints after gene duplication between the majority of isoform pairs. An overview of estimated type I functional divergence coefficients (θ_I), LRTs, estimated False Discovery Rate (FDR), and number of identified sites is available in Table 3, while detailed information regarding site position and composition can be found in Online Resource 5.

LRTs for sf1 pointed towards 8 out of 10 isoform pairs displaying signs of T1-FD: A/B, A/G, A/K, B/G, B/I, B/K, G/I, G/K. FDRs ranged from 0.29% (A/B pair) to 11.61%

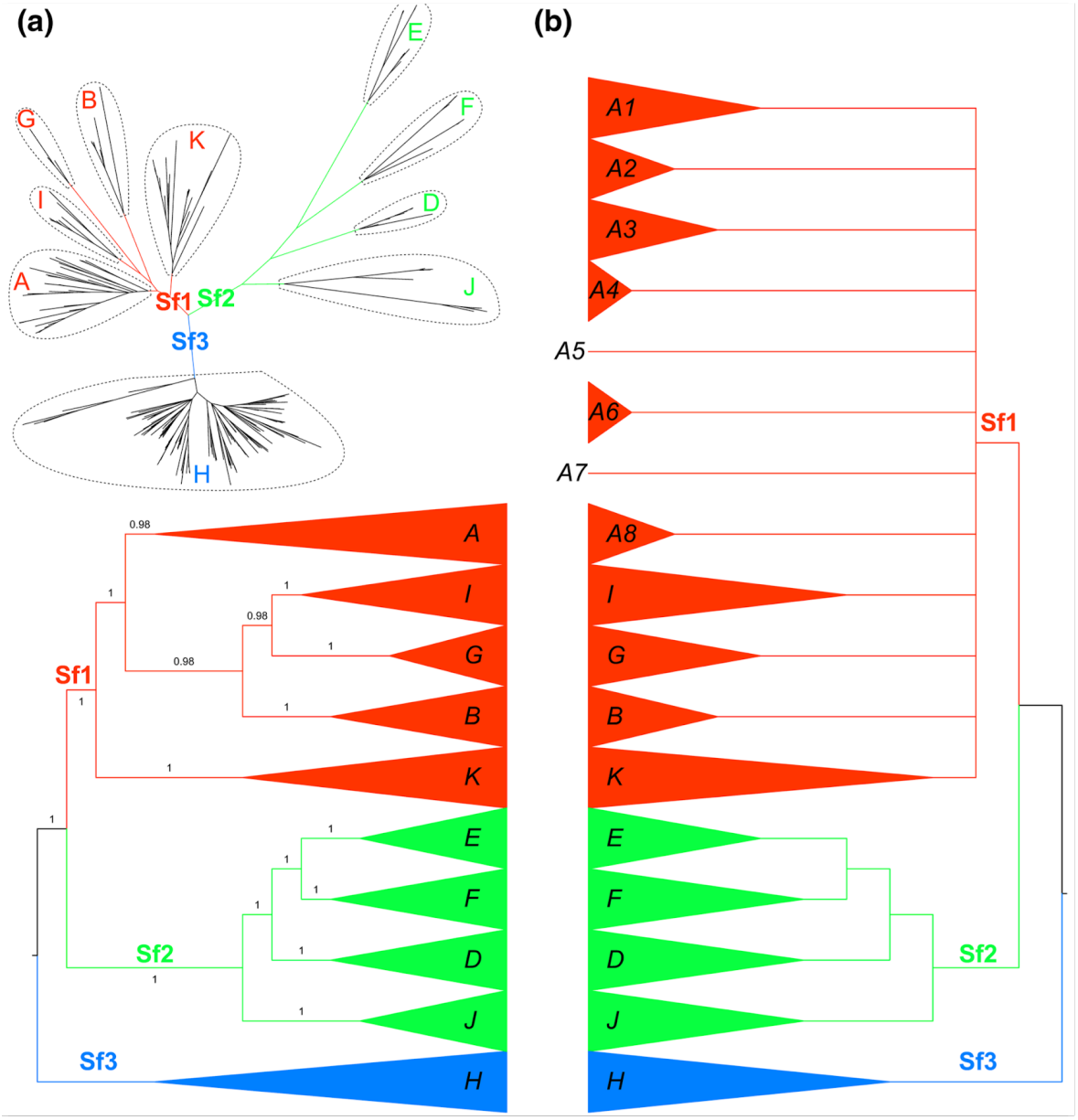


Fig. 4 Global Class II Pr1 phylogenetic reconstructions. a Above: Schematic representation of the relationships among subfamilies. Internal branches, in black, are to reflect OTU abundance; Below: Phylogenetic reconstruction using Bayesian Inference with the GUIDANCE filtered nucleotide dataset, following the GTR + I + G evolutionary model, sampling 10,000,000 generations. Estimated posterior probabilities are adjacent to their corresponding branches; **b** Metatree combining all constructed topologies (Maximum Likelihood and Bayesian Inference methods using both amino acid and nucleotide datasets). Monophyletic clades corresponding to groups (or subsets of them) of OTUs for each isoform are represented as triangles. The red, green, and blue colors relate to subfamily 1, 2, and 3, respectively (branches are labeled accordingly)

(G/K) and number of FD-related sites, when any, were as small as 1 (A/B) and as high as 24 (A/G). Isoform pairs B/I and G/I, despite showing statistical significance, displayed zero sites above our θ_I threshold (and, unsurprisingly, 0% FDR).

Sf2 FD analyses were twofold: regarding Pr1J as a single cluster (named 1J), and considering it to be formed

Table 2 Qualitative view of positive selection tests for all analyzed proteins

Subfamily	Isoform	MetaTree		Selected Topology		
		M1a vs M2a	M7 vs M8	M1a vs M2a	M7 vs M8	#PSS[a]
Sf1	A	N	Y	N	Y	3
	B	Y	Y	N	N	1
	G	N	Y	N	Y	1
	I	Y	Y	Y	Y	5
	K	N	N	N	N	N/A
Sf2	D	N/A	N/A	Y	Y	3
	E	N	N	Y	N	N/A
	F	N	N	N	N	N/A
	J	Y	Y	Y	Y	3
Sf3	H	N	N	N	N	N/A

[a]Number of positively selected sites, *Y* null hypothesis rejection, *N* null hypothesis acceptance, *N/A* non-applicable

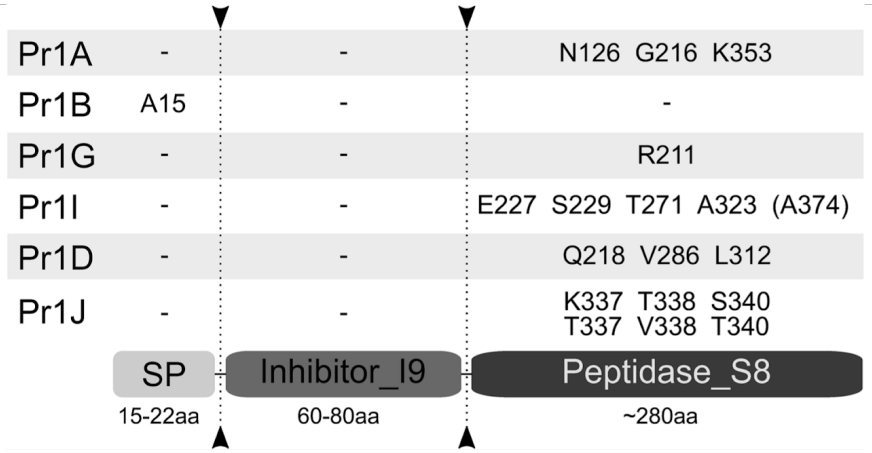


Fig. 5 Location of identified positively selected sites regarding domain architecture. The typical architecture of Pr1 proteases is depicted in the lowermost part. Usually, the polypeptide contains a signal-peptide (SP), an inhibitory domain (Inhibitor_I9; PF05922), and a subtilase-like proteolytic domain (Peptidase_S8; PF00082). Lines are ordered according to isoforms, and hyphens represent the absence of positively selected sites in that region. The site in parentheses (Pr1I) is located in the C-terminal portion, with no identifiable domain by sequence similarity, but still a part of the mature polypeptide. Proteins Pr1E, Pr1F, Pr1H, and Pr1K did not display signs of positive selection

by two independent branches (2J). Since our phylogenies indicate two monophyletic clades following a specialist-transitional-generalist trend, there might be in fact two functionally distinct isoforms that were considered as Pr1J. Supporting this hypothesis, comparisons between clusters J1 and J2 have shown statistical signs of T1-FD, depicting 45 FD-related sites with 9.38% FDR. Additionally, while D/J portrayed a similar scenario to B/I and G/I from sf1 (i.e., statistical significance and no above-threshold sites), both D/J1 and D/J2 comparisons were statistically significant and indicated 7 and 2 FD-related sites with 8.25% and 7.38% FDR, respectively. Comparing 1J to 2J, both E/J and F/J displayed an overall decrease in FDR (12.13–8.56% average for Pr1E; 10.05–2.92% average for Pr1F) and an increase in FD-related sites (1 to 32 and 6 - E/J1 and E/J2, respectively; 2 to 132 and 1-F/J1 and F/J2, respectively). Finally, E/F yielded the same trend as D/J (rejection of the null, no sites) and the alternate hypothesis of FD was rejected for D/F. It is worth mentioning that D/E, D/F, and E/F comparisons for 1J are exactly the same in 2J due to using the same alignment and topologies, yielding the exact same results and, therefore, they are not depicted in the 2J rows in Table 3.

Table 3 Type I functional divergence analysis for Pr1 Class II subtilisin-like serine proteases

Clades	$\theta_I \pm SE_\theta$	LRT θ (p)	FDR[a] (%)	#Sites[a]
Sf1				
A/B	0.310 ± 0.070	19.718 ($< 10^{-5}$)	0.29	1
A/G	0.700 ± 0.100	48.525 ($< 10^{-11}$)	8.22	24
A/I	0.098 ± 0.064	2.390 (> 0.10)	N/A	N/A
A/K	0.190 ± 0.039	23.521 ($< 10^{-5}$)	6.04	4
B/G	0.726 ± 0.133	29.891 ($< 10^{-7}$)	8.55	11
B/I	0.318 ± 0.105	9.200 ($< 10^{-2}$)	0.00	0
B/K	0.314 ± 0.067	21.892 ($< 10^{-5}$)	10.51	6
G/I	0.387 ± 0.161	5.765 (< 0.02)	0.00	0
G/K	0.654 ± 0.117	31.328 ($< 10^{-7}$)	11.61	17
I/K	0.092 ± 0.049	3.560 (> 0.05)	N/A	N/A
Sf2 (1J)				
D/E	0.654 ± 0.152	18.505 ($< 10^{-4}$)	8.93	7
D/F	0.235 ± 0.159	2.188 (> 0.10)	N/A	N/A
D/J	0.374 ± 0.183	4.186 (< 0.05)	0.00	0
E/F	0.443 ± 0.153	8.373 ($< 10^{-2}$)	0.00	0
E/J	0.569 ± 0.196	8.397 ($< 10^{-2}$)	12.13	1
F/J	0.398 ± 0.128	9.667 ($< 10^{-2}$)	10.05	2
zSf2 (2J)				
D/J1	0.594 ± 0.126	22.259 ($< 10^{-5}$)	8.25	7
D/J2	0.564 ± 0.141	16.034 ($< 10^{-4}$)	7.38	2
E/J1	0.778 ± 0.121	41.397 ($< 10^{-9}$)	8.61	32
E/J2	0.670 ± 0.162	17.043 ($< 10^{-4}$)	8.50	6
F/J1	0.982 ± 0.126	60.412 ($< 10^{-14}$)	1.84	132
F/J2	0.512 ± 0.146	12.327 ($< 10^{-3}$)	4.00	1
J1/J2	0.810 ± 0.129	39.527 ($< 10^{-9}$)	9.38	45

θ_I Type I functional divergence coefficient, SE_θ standard error of the estimate, *LRT* θ Likelihood Ratio Test of functional divergence, which is compared to a χ^2 distribution with 1 *df*

[a] 0.86 posterior probability cut-off, *N/A* non-applicable

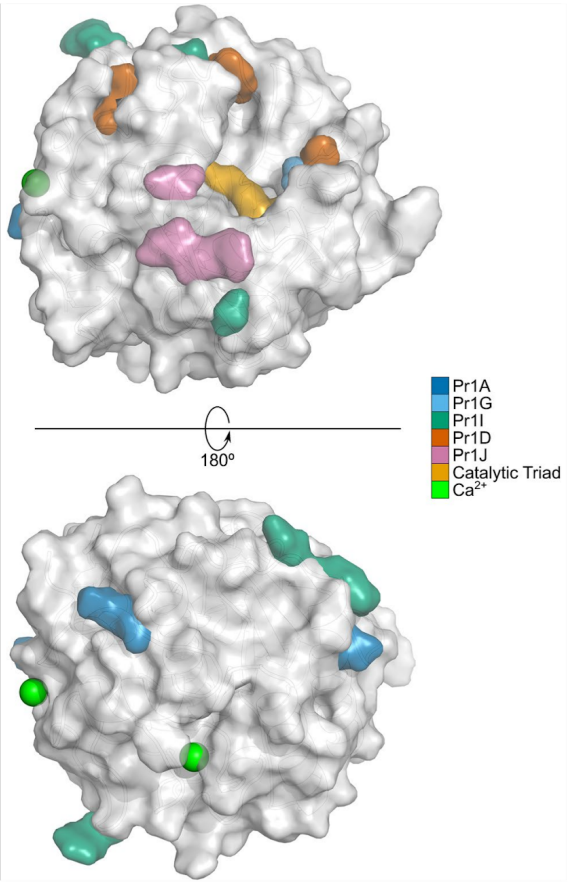


Fig. 6 Location of positively selected sites regarding protein structure. Site positions are depicted as colored sections on the surface representation of a Proteinase K crystallographic structure from *Parengyodontium album* at a resolution of 0.98 Å (PDB ID 1IC6), with the underlying backbone structure rendered in black lines. The lower structure represents a 180° rotation on the horizontal axis of the upper one

Structural projection

Knowing the position of PS sites in primary structures offers little information regarding potential modulation of function at a molecular level other than an eventual physical–chemical change in an amino acid sidechain. As structure begets function, a structure-based sequence alignment in conjunction with crystallographic information may provide a glimpse into how evolution is shaping Pr1 proteins. Our *phmmer* searches indicated PDB ID 1IC6, determined by X-ray crystallography at a 0.98-Å resolution (Betzel et al. 2001), as a suitable homologous structure for all our representative sequence, ranging from as high as 69% identity and 86.1% similarity (Pr1A from *M. guizhouense*) to as low as 32.6% identity and 52.2% similarity (Pr1J2 from *M. anisopliae*), and its residue numbers will be used as reference. As this structure only comprises the mature peptide, we also retrieved the full protein sequence from Uniprot (P06837) and added it to our dataset for sequence alignment (Online Resource 6). Regarding residue conservation: the catalytic triad (D144, H174, and S329 on P06837) was preserved on all sequences; disulfide bonds cysteines (C139–C228 and C283–C354) were present in all Sf1 members except for Pr1G from *M. acridum*, which displayed a deletion of residue 283; and calcium binding sites appeared in variable levels of conservation throughout all isoforms. Amino acid mutations on structurally relevant regions can potentially modulate protein function. By projecting alignment information on to 1IC6 (Fig. 6), we were able to observe four PS residues in proximity to the catalytic cleft: G100 for Pr1G (arginine in *M. anisopliae* and *M. acridum*; glutamine in *M. guizhouense*), S101 (glutamine in *M. anisopliae;* glutamic acid in *M. guizhouense* and *M. acridum*), and Y169

(valine in *M. anisopliae;* tyrosine in *M. guizhouense* and *M. acridum*) for Pr1D, and S221 for Pr1J (J1: serine for all representatives; J2: threonine for *M. anisopliae* and serine for *M. guizhouense* and *M. acridum*). Additionally, residue S17 for Pr1A (asparagine in *M. anisopliae* and *M. guizhouense*, glutamine in *M. acridum*) appeared close to a calcium atom and adjacent to residue T16, which composes calcium binding site 2 (Online Resource 6). Remaining PS residues are distributed on the protein surface with no apparent potential influence on functionally related regions.

Discussion

In order to test the hypothesis that Pr1 serine-proteases belonging to *M. anisopliae* are subject to differential selective pressures, we conducted extensive phylogenetic analyses encompassing all Class II members. Additionally, we identified positively selected sites in six—Pr1A, B, D, G, I, and J—out of ten analyzed datasets, suggesting an increase in non-synonymous substitution rates over synonymous ones, which leads towards additional amino acid variability in those sequences. We have also found evidence for Type I functional divergence in most pairwise comparisons within subfamilies, further supporting that Pr1 isoforms are not functionally redundant and pointing to a potential novel protease isoform. This modification or innovation tendency and functional shifts in the Pr1 family have potential impacts over arthropod host ranges in this fungal genus.

Among sequences that were identified as homologous to those of *M. anisopliae* E6, proteins Pr1E, Pr1F, and Pr1J appeared in multiple non-identical copies. This multiplicity seems to be associated with the host range expansion process in the *Metarhizium* genus, since the increase in host numbers is associated to protein families expansion (Hu et al. 2014). Phylogenetic analyses for Pr1E, which originated from an *in tandem* duplication of a Pr1F-like gene (Bagga et al. 2004), support this hypothesis due to the presence of monophyletic clades related to the "additional" copies, exclusively identified in generalist *Metarhizium* species (*M. anisopliae, M. brunneum*, and *M. robertsii*; Online Resource 3–3 and 3–4). Analyses portrayed in this work, regarding these isoforms, differ from previous studies, where only one copy of each protein was employed per species (Bagga et al. 2004; Hu and St. Leger 2004; Li et al. 2010). Overall, our sampling of homologous sequences to Pr1 proteases was more extensive than those of earlier studies, mainly due to the availability of newly sequenced genomes encompassing fungi of different host ranges (Hu et al. 2014).

As a general trend observed in our phylogenies, proteins belonging to members of the *Metarhizium* genus clustered according to host ranges, placing host-specialty as an ancestral character to wider ranges (such as those of generalist or transitional species), in accordance with the evolutionary relationships established using genomes (Hu et al. 2014). Furthermore, observed branchings at the family level resemble those observed by Hu et al. (2014), in which fungal family evolution in the Hypocreales order is described as [Nectriaceae, (Hypocreaceae, (Cordycipitaceae, (Ophiocordycipitaceae, Clavicipitaceae)))]. Analyzing those datasets comprising multiple families (Pr1A, Pr1B, Pr1H, and Pr1K), the same pattern emerged albeit with minor discrepancies. Noteworthy, no external groups were included in our datasets, rendering it difficult to attribute temporal meaning to topologies, while allowing a near-equivalent representation to previous studies that included such groups by direct branch reordering. The presence of discrepantly grouped orders, like those in the internal branches of Hypocreales (Glomerellales in Pr1K and Xylariales in Pr1A), and polyphyletic groups such as Cordycipitaceae and Ophiocordycipitaceae in Pr1A may be related to suboptimal sampling of amino acid sequences or a different evolutionary process, as the phylogenies of specific proteins do not need to necessarily reflect their species evolution. As pointed out by Pearson (2013), sequences sharing less than 30% identical residues may still be homologous. In this work, we applied a 60% identity threshold, possibly rendering our search overly stringent.

Identified relationships among isoforms from subfamilies 1 and 2 are better supported than previous ones. Holder and Lewis (2003) point towards a 70% bootstrap test value as an indicator of strong support for any given group in phylogenetic analyses, and such value will be used as reference for the following discussion. By employing Pr1 sequences belonging to three *M. anisopliae* lineages, Bagga et al. (2004) have established a [G, (A, B, I)] relationship for subfamily 1, not including Pr1K, and a (D, E, F, J) pattern for subfamily 2 members, with sf3/Pr1H at a basal position. When including proteins from other species, branching was shown as [H, (J, D, (E, F)), (K, (A, B, G, I))], depicting three polytomically related subfamilies. Also in 2004, Hu and St. Leger have inferred a [K, G, I, (A, B)] relationship of sf1 and [D, J, (E, F)] for sf2, also positioning sf3/Pr1H as root. Li et al. (2010) related subtilisin-like protease sequences belonging to the *Pezizomycotina* subphylum, which encompass several pathogenic fungal classes, including various Pr1 isoforms in *M. anisopliae*. Their work separated all three subfamilies although internal relationships in these clades lack statistical support (posterior probabilities estimated using MrBayes, in that case). A more recent study by the same group (Li et al. 2017) did not rescue the monophyly of Pr1 isoforms and subfamilies, depicting the evolution of already described Class II proteases (i.e., not considering putative entries) as [H, A, B, G, I, K J, D, (E, F)]. Global phylogenetic reconstructions in our work (Fig. 4, Online

Resource 3–6) pointed towards a [K, (A, (B, (G, I)))] for sf1 and [J, (D, (E, F))] for sf2, rooted in sf3/Pr1H, where internal branchings for these clades corresponded, in general, to those visualized in the individual phylogenies for each isoform. This work introduces a fully resolved phylogeny at the subfamily and isoform levels, with high statistical support, providing a more solid basis on the classification of Class II subtilisin-like Pr1 proteins, while consistent with previous knowledge on the subject.

Positive selection inferences here presented differ from those identified in previous works. Both Bagga et al. (2004) and Hu and St. Leger (2004) included exclusively information of generalist *Metarhizium* species, being unable to find evidence of positive selection in average d_N/d_S comparisons throughout whole alignments. As it was also pointed out by Bagga and coworkers (2004), averaging ω over all codons may mask Pr1 regions under differential selective pressures. Estimations using our datasets, employing individual sets for each protein belonging to Pr1 subfamilies 1, 2, and 3—including data from generalist, specialist, and transitional *Metarhizium* species, as well as from closely related fungal species—pointed towards positive selection in six out of ten Class II isoforms, specifically in Pr1A, Pr1B, Pr1D, Pr1G, Pr1I, and Pr1J. This suggests an active process of amino acid diversification similar to that observed by Li et al. (2010) for nematode-trapping fungi, which are closely related to entomopathogens. Location of positively selected sites inside the proteolytic domain region is consistent with the hypothesis that each isoform has diverged (and is still diverging) to perform different functions in entomopathogenic fungi (Bagga et al. 2004). Structural information (Fig. 6; Online Resource 6) further supports this claim: residues G100 and S101 (PS in Pr1G and Pr1D, respectively) are both involved in substrate recognition by the formation of a triple-stranded beta-sheet with the peptide substrate (Betzel et al. 1988). Beyond cuticle penetration, these functions can comprise affinities for different substrates, host cellular defense suppression, defense molecules' degradation, and contribute to an increase in host ranges (Vilcinskas 2010). Additionally, the proximity of some PS residues to the catalytic cleft, such as Y169 (in Pr1D) and S221 (Pr1J), might modulate pocket size and physical–chemical properties. A positively selected site inside the signal peptide (pre region) of Pr1B may specify different modes of targeting and membrane insertion (Martoglio et al. 1998). Elevated and increasing expression of this isoform in *D. peruvianus* in the first 96 h of *M. anisopliae* infection (Beys-da-Silva et al. 2014) suggests that this isoform is relevant in the infective process of this specific host and secretion using different pathways may reflect the necessities of infecting a particular host class. This hypothesis, however, should be taken carefully due to its inference being solely based on Pr1B-MT (Online Resource 4–3), which is largely unresolved.

Opposed to what (Hu et al. 2014) have identified, our analyses did not indicate positive selection acting on Pr1K, while agreeing on Pr1G, for the *M. acridum* lineage. Conceptually, the 2014 study has evaluated selective pressures through branch models on seven *Metarhizium* genomes exclusively, while in the present work ω was estimated using site models, also including sequences from other fungal families. For this discordant case in particular, the presence of evolutionarily farther sequences [i.e., from Nectriaceae, which diverged approximately 395 million years ago, against nearly 117 million years for *Metarhizium* spp. (Hu et al. 2014)] may have mitigated selection signals in this dataset. From a different perspective, the presence of highly similar sequences from *Fusarium* spp. (Nectriaceae) may overestimate d_S over d_N, reducing ω in such a way that positive selection signals are somehow masked. Reevaluation of these results by further reducing the dataset through identity criteria, analogously to Pr1A and Pr1H scenarios, besides a less stringent resampling (as discussed above), may be beneficial in addressing this apparent contradiction.

Previous FD studies involving the proteinase K family of subtilisin-like serine proteases in fungi, which included a subset of known Pr1 isoforms, did not display signs of T1-FD (Varshney et al. 2016). Opposed to those results, incorporating an expanded sample of Pr1 proteins, we identified most of the analyzed pairs of proteins under statistically significant T1-FD (Table 3). Four of those, despite rejection of the null hypothesis, did not possess FD-related sites with above-threshold posterior probabilities, suggesting them to be false positives under our criteria. Under this scenario, Pr1I would not be functionally divergent to any sf1 isoform, or would be a functional intermediate in this subfamily; the functionalities of Pr1D would not significantly diverge from Pr1F and Pr1J, but would from Pr1E, placing it in a somewhat similar hypothetical situation than isoform I; and Pr1E and Pr1F would not be functionally divergent or their activities would be very similar. Supporting these hypotheses, in *M. anisopliae* E6, Pr1I has the highest average sequence identity in all of sf1 (compared with itself), Pr1D shares on average more identical residues with Pr1F and Pr1J than it does with isoform E, and Pr1E and Pr1F share more average identical sites between them than every other sf2 isoform.

Large numbers of Pr1 isoforms in generalist *Metarhizium* species in comparison to specialists or even other fungal species may have allowed pathogenicity towards a large number of hosts. Consistent with our results, Vilcinskas (2010) suggests that adaptation to a larger number of hosts should be accompanied by rapid diversification of genes involved in multiple host interactions. Additionally, it is pointed out that adaptation to particular host species should promote loss of genes that are unaffected by selective pressures due to lack of function in the pathogenesis of reduced host ranges. For sf2 (Fig. 2), particularly, the presence of

multiple groups related to Pr1E and Pr1F, solely associated with generalist species (or predominantly generalist, as it is for Pr1F), may indicate a gain in enzymatic capacity by generalists or subtilisin-like gene loss by specialists with probable effects in virulence. Additionally, the presence of two highly supported subclades in clade J, which contains isoforms regarded as second in abundance among the Pr1 family (Freimoser et al. 2003), following the same specialist-transitional-generalist pattern, suggests that they may be different subtypes. Indeed, when splitting those clades and testing for T1-FD, 45 FD-related sites were identified and both Pr1J1 and Pr1J2 were depicted as functionally divergent to all other sf2 isoforms. One thing to notice is that the F/J1 comparison yielded near-one θ_I (Table 3) and all sites displayed posterior probabilities > 0.96 (Online Resource 5). Even though deemed a normal software behavior, some site compositions were homogeneous and incompatible with a FD-related site; thus, data for F/J1 are to be disregarded. Gu (2003) points out that "except for a large number of sequences, one should be cautious about the result when all pairwise sequence identities are > 90%, because of the lacking of statistical power". Although such levels of identity are not observed in every pairwise comparison, many proteins from *Metarhizium* spp. or *Fusarium* spp. in our individual datasets share high sequence identity and could potentially impact T1-FD inferences. Subfamily 2 is particularly susceptible due to almost exclusively containing *Metarhizium* sequences. In order to either confirm or refute these hypotheses, a more thorough phylogenetic analysis with increased sampling of the Pr1 family may be beneficial, as well as a biochemical recharacterization of Pr1J proteins found in the genomes of generalists such as *M. anisopliae*.

Evolution of parasitic lifestyles depends on the availability of enzymatic virulence factors when interacting with hosts, be it on nutrient acquisition, on infection by itself, or by weakening host defenses (Vilcinskas 2010). Presumably, pathogen virulence coevolves as a result of reciprocal selection with the host, there being positive selection towards the evolution of new types of proteases or isoforms that can overcome host defenses, such as protease inhibitors (Vilcinskas 2010). Vilcinskas (2010) has illustrated, in a simplified way, a few possible scenarios and results here presented point towards a combination of scenarios 3 and 4 in *M. anisopliae*, where proteases are either partially or non-inhibited by the host, allowing for a shift in importance of virulence factors depending on hosts, despite absolute concentration or activity, due to a strong diversifying selection of pathogen-associated proteases. The large number of expressed isoforms by this generalist fungus is determinant of host ranges, and its differential expression on distinct hosts is a potential physiological adaptation to the inhibitory mechanisms of its host, while enabling hydrolysis of different cuticle compositions (Vilcinskas 2010). Additionally, isoform amounts and expression patterns for Pr1 proteases may also be pre-adaptive features to a parasitic lifestyle (Hu and St. Leger 2004) or even to different environments outside the host, such as plant rhizospheres (St Leger 2008).

To date, this is the most comprehensive study on the molecular evolution of Class II subtilisin-like serine proteases in *Metarhizium* spp., unveiling patterns of protein diversification that were not addressed before on this family of virulence factors. Identification of positive selection and functional divergence in sf1 and sf2 is consistent with the expected increase in evolutionary rates for duplicated genes involved in infecting multiple hosts (Vilcinskas 2010; Li et al. 2017). Further work is required to confirm the importance of T1-FD-related and positively selected sites in these proteins, as to identify the effects of these sites in protein structure and function, although functionally relevant regions appear to be affected. It would also be of benefit to apply a similar approach to expanded datasets with more permissive criteria, including Pr1C and the additional Pr1H copy that were not included in this study. The employed methodology is meant to be the least subjective as possible, and we enforce the use of collapsed (or condensed) topologies for phylogeny representation to avoid misinterpretation of results by the non-specialist community and to remove unsupported inferences. We expect our work to expand current knowledge of Pr1 evolution, paving the way for future studies on *Metarhizium* spp. virulence and pathogenicity, as well as to provide a reliable analysis framework for molecular evolution studies.

Acknowledgements The authors would like to thank BV, CLF, CLM, JFRB, GPP, MT, RKC, RLB, and NSO for insightful discussions, suggestions and overall support. This work was supported by the following Brazilian agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [Grant: Biocomputacional 23038.010041/2013-13, Rede Avançada em Biologia Computacional (RABICÓ)], Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Grant: Universal 2014 458160/2014-8].

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

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

Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38:W7–W13

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA (2012) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol 61:170–173

Bagga S, Hu G, Screen SE, St. Leger RJ (2004) Reconstructing the diversification of subtilisins in the pathogenic fungus *Metarhizium anisopliae*. Gene 324:159–169

Barelli L, Moonjely S, Behie SW, Bidochka MJ (2016) Fungi with multifunctional lifestyles: endophytic insect pathogenic fungi. Plant Mol Biol 90:657–664

Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol 10:980–980

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303

Betzel C, Pal GP, Saenger W (1988) Three-dimensional structure of proteinase K at 0.15-nm resolution. Eur J Biochem 178:155–171

Betzel C, Gourinath S, Kumar P, Kaur P, Perbandt M, Eschenburg S, Singh TP (2001) Structure of a serine protease proteinase K from *Tritirachium album limber* at 0.98 Å resolution. Biochemistry 40:3080–3088

Beys-da-Silva WO, Santi L, Berger M, Calzolari D, Passos DO, Guimarães JA, Moresco JJ, Yates JR (2014) Secretome of the biocontrol agent *Metarhizium anisopliae* induced by the cuticle of the cotton pest *Dysdercus peruvianus* Reveals new insights into infection. J Proteome Res 13:2282–2296

Butt TM, Coates CJ, Dubovskiy IM, Ratcliffe NA (2016) Entomopathogenic fungi: new insights into host–pathogen interactions. In: Advances in genetics, vol 94. Elsevier, Amsterdam, p 307–364

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772–772

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Faria MR de, Wraight SP (2007) Mycoinsecticides and mycoacaricides: a comprehensive list with worldwide coverage and international classification of formulation types. Biol Control 43:237–256

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37

Finn RD, Bateman A, Clements J, Coggil P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Freimoser FM, Screen S, Bagga S, Hu G, St. Leger RJ (2003) Expressed sequence tag (EST) analysis of two subspecies of *Metarhizium anisopliae* reveals a plethora of secreted proteins with potential activity in insect hosts. Microbiology 149:239–247

Freimoser FM, Hu G, St. Leger RJ (2005) Variation in gene expression patterns as the insect pathogen *Metarhizium anisopliae* adapts to different host cuticles or nutrient deprivation in vitro. Microbiology 151:361–371

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224

Greenfield BPJ, Peace A, Evans H, Dudley E, Ansari MA, Butt TM (2015) Identification of *Metarhizium* strains highly efficacious against *Aedes, Anopheles* and *Culex* larvae. Biocontrol Sci Technol 25:487–502

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Gu X (2003) Functional divergence in protein (family) sequence evolution. In: Long M (ed) Origin and evolution of new gene functions. Contemporary issues in genetics and evolution, vol 10. Springer, Dordrecht

Gu X, Wang Y, Gu J, Velden KV, Xu D (2006) Predicting type-I (rate-shift) functional divergence of protein sequences and applications in functional genomics. Curr Genomics 7:87–96

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y (2013) An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol 30:1713–1719

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284

Hu G, St. Leger RJ (2004) A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. J Evol Biol 17:1204–1214

Hu X, Xiao G, Zheng P, Shang Y, Su Y, Zhang X, Liu X, Zhan S, St. Leger RJ, Wang C (2014) Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. Proc Natl Acad Sci USA 111:1–6

Javar S, Mohamed R, Sajap AS, Lau W-H (2015) Expression of pathogenesis-related genes in *Metarhizium anisopliae* when infecting *Spodoptera exigua*. Biol Control 85:30–36

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282

Khan FI, Wei D-Q, Gu K-R, Hassan MI, Tabrez S (2016) Current updates on computer aided protein modeling and designing. Int J Biol Macromol 85:48–62

Landan G, Graur D (2008) Local reliability measures from sets of co-optimal multiple sequence alignments. Pac Symp Biocomput 15–24

Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278

Le SQ, Gascuel O (2008) An Improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320

Li J, Yu L, Yang J, Dong L, Tian B, Yu Z, Liang L, Zhang Y, Wang X, Zhang K (2010) New insights into the evolution of subtilisin-like serine protease genes in *Pezizomycotina*. BMC Evol Biol 10:68

Li J, Gu F, Wu R, Yang J, Zhang K-Q (2017) Phylogenomic evolutionary surveys of subtilase superfamily genes in fungi. Sci Rep 7:45456

Löytynoja A (2014) Phylogeny-aware alignment with PRANK. In: Russell DJ (ed) Multiple sequence alignment methods. Humana Press, New Yprk, p 155–170

Martoglio B, Dobberstein B, Carafoli E (1998) Signal sequences: more than just greasy peptides. Trends Cell Biol 8:410–415

Nye T (2008) Trees of trees: an approach to comparing multiple alternative phylogenies. Syst Biol 57:785–794
Pearson WR (2013) An Introduction to sequence similarity ("homology") searching. Curr Protoc Bioinforma 42:3.1.1–3.1.8
Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295–2300
Rambaut A (2016) FigTree. http://tree.bio.ed.ac.uk/software/figtree/. Accessed 15 Mar 2015
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542
Rosas-García NM, Avalos-de-León O, Villegas-Mendoza JM, Mireles-Martínez M, Barboza-Corona JE, Castañeda-Ramírez JC (2014) Correlation between Pr1 and Pr2 gene content and virulence in *Metarhizium anisopliae* strains. J Microbiol Biotechnol 24:1495–1502
Sánchez-Pérez L, de C, Barranco-Florido, Rodríguez-Navarro JE, Cervantes-Mayagoitia S, Ramos-López JF MÁ (2014) Enzymes of entomopathogenic fungi, advances and insights. Adv Enzym Res 02:65–76
Santi L, Silva WOB, Pinto AFM, Schrank A, Vainstein MH (2010) *Metarhizium anisopliae* host-pathogen interaction: differential immunoproteomics reveals proteins involved in the infection process of arthropods. Fungal Biol 114:312–319
Schrank A, Vainstein MH (2010) *Metarhizium anisopliae* enzymes and toxins. Toxicon 56:1267–1274
Schrödinger L (2015) The PyMOL Molecular Graphics System, Version 2.2
Sela I, Ashkenazy H, Katoh K, Pupko T (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res 43:W7–W14
Small C-LN, Bidochka MJ (2005) Up-regulation of Pr1, a subtilisin-like protease, during conidiation in the insect pathogen *Metarhizium anisopliae*. Mycol Res 109:307–313
St Leger RJ (2008) Studies on adaptations of *Metarhizium anisopliae* to life in the soil. J Invertebr Pathol 98:271–276
St Leger RJ, Bidochka MJ, Roberts DW (1994) Isoforms of the cuticle-degrading Pr1 proteinase and production of a metalloproteinase by *Metarhizium anisopliae*. Arch Biochem Biophys 313:1–7
Staats CC, Junges A, Guedes RLM, Thompson CE, Morais GL, Boldo JT, Almeida LGP, Andreis FC, Gerber AL, Sbaraini N, Paixão RLA, Broetto L, Landell M, Santi L, Beys-da-Silva WO, Silveira CP, Serrano TR, Oliveira ES, Kmetzsch L, Vainstein MH, Vasconcelos ATR, Schrank A (2014) Comparative genome analysis of entomopathogenic fungi reveals a complex set of secreted proteins. BMC Genom 15:822
Stöver BC, Müller KF (2010) TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:7
Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Am Math Soc Lect Math Life Sci 17:57–86
Varshney D, Jaiswar A, Adholeya A, Prasad P (2016) Phylogenetic analyses reveal molecular signatures associated with functional divergence among Subtilisin like Serine Proteases are linked to lifestyle transitions in Hypocreales. BMC Evol Biol 16:220
Vilcinskas A (2010) Coevolution between pathogen-derived proteinases and proteinase inhibitors of host insects. Virulence 1:206–214
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
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591
Zimmermann G, Papierok B, Glare T (1995) Elias Metschnikoff, Elie Metchnikoff or ilya ilich mechnikov (1845–1916): a pioneer in insect pathology, the first describer of the entomopathogenic fungus *Metarhizium anisopliae* and how to translate a Russian Name. Biocontrol Sci Technol 5:527–530

Discussão

Visando testar se as serino-proteases do gênero *Metarhizium* estão sujeitas a pressões seletivas diferenciais, realizamos um estudo filogenético amplo. Foram observados indícios de seleção positiva na maioria dos casos analisados. Além disso, há evidência de divergência funcional Tipo I dentro das subfamílias 1 e 2, oferecendo suporte à hipótese de que não há redundância funcional das proteases Pr1, bem como sugerindo a existência de uma nova proteoforma de Pr1J até então não descrita. Essa tendência à variabilidade, juntamente com desvios funcionais, tem potencial impacto sobre o leque de hospedeiros artrópodes nesse gênero fúngico.

Em geral, as filogenias mostram entradas de *Metarhizium* ramificadas conforme a abrangência de hospedeiros, posicionando a especialização de hospedeiros como característica plesiomórfica aos alcances maiores[23], em consonância com dados genômicos (HU et al., 2014). Ademais, as reconstruções baseadas em sequências nucleotídicas e aminoacídicas, em geral, recuperaram a filogenia das espécies que as contém. Essa observação implica que a transmissão dos genes *pr1* ocorre de maneira vertical, sem indícios de transferência interespecífica ou horizontal.

As relações filogenéticas observadas entre parálogos das famílias 1 e 2 apresentam melhor suporte estatístico que as previamente disponíveis no momento de publicação deste artigo[24]. As reconstruções de filogenia global (alinhando todas as Pr1 amostradas) apontam para ramificações [K,(A,(B,(G,I)))], para a sf1, e [J,(D,(E,F))] ,para a sf2, enraizadas na sf3/Pr1H onde, em geral, as ramificações internas corresponderam àquelas observadas nas reconstruções individuais por proteoforma. Nosso trabalho introduz uma filogenia totalmente resolvida nos níveis de subfamília e proteoforma, com alto suporte estatístico, provendo uma base mais sólida na classificação das proteínas Pr1 da Classe II,

[23] Transicionais e generalistas.
[24] 28 de março de 2019.

enquanto mantendo a consistência com conhecimento prévio (BAGGA et al., 2004; LI et al., 2010, 2017).

As inferências de seleção positiva e divergência funcional deste trabalho introduzem uma nova perspectiva na evolução das proteases Pr1 Classe II. A presença de pressão seletiva positiva em seis de dez genes *pr1*[25], cujos sítios localizam-se no domínio proteolítico, sugerem processo ativo de diversificação. Adicionalmente, observou-se a presença de divergência funcional estatística na maioria das comparações par a par realizadas em cada subfamília, com múltiplos resíduos-chave associados. Em conjunto, esses dados são consistentes com a hipótese de que a família de proteases Pr1 divergiu (e continua divergindo) de modo a desempenhar atividades não-redundantes em fungos entomopatogênicos, podendo influenciar diferentes afinidades por substratos, supressão ou degradação de respostas do hospedeiro e aumento no alcance de hospedeiros.

Números maiores de proteoformas de Pr1 em espécies generalistas de *Metarhizium* em comparação com especialistas ou outras espécies fúngicas podem ter resultado na capacidade de infectar hospedeiros. Consistente com nossos resultados, Vilcinskas (2010) sugere que a adaptação a um maior número de hospedeiros deve ser acompanhada de rápida diversificação dos genes envolvidos na interação com múltiplos hospedeiros. Adicionalmente, a adaptação a uma espécie particular de hospedeiros deve promover perda de genes que não são afetados por pressões seletivas devido à falta de função na patogênese de alcances reduzidos de hospedeiros. No caso da sf2, em particular, a presença de múltiplos grupos[26] associados apenas, ou em maioria, com espécies generalistas pode indicar um ganho de capacidade enzimática nos especialistas ou perda de genes por parte dos especialistas, com efeitos prováveis na virulência.

A evolução de estilos de vida parasíticos depende da disponibilidade de fatores enzimáticos de virulência na interação com os hospedeiros, seja na aquisição de nutrientes, na infecção propriamente dita ou pelo enfraquecimento das defesas do hospedeiro. Presumidamente, a virulência de um patógeno

[25] Pr1A, Pr1B, Pr1D, Pr1G, Pr1I e Pr1J.
[26] Em Pr1E e Pr1F.

coevolui como resultado de seleção recíproca com seu hospedeiro, havendo seleção positiva no sentido da evolução de novos tipos de proteases ou proteoformas que possam sobrepujar as defesas do alvo, tais como inibidores de proteases. Vilcinskas (2010) ilustra, de maneira simplificada, alguns cenários possíveis. Os dados aqui apresentados apontam na direção de uma combinação dos cenários 3 e 4 em *M. anisopliae*, onde proteases são ou parcialmente, ou não-inibidas pelo hospedeiro, permitindo uma mudança na importância de fatores de virulência de acordo com os hospedeiros, independente de concentração absoluta ou atividade, em virtude de uma forte seleção diversificadora em proteases associadas a patógenos. O número de parálogos expressos por esse fungo generalista é determinante no alcance de hospedeiros e sua expressão diferenciada em hospedeiros distintos é uma adaptação fisiológica em potencial aos mecanismos inibitórios de seu hospedeiro, enquanto permitindo a hidrólise de diferentes composições cuticulares (VILCINSKAS, 2010). Adicionalmente, a quantidade de proteoformas e os padrões de expressão das proteases Pr1 podem ser características pré-adaptativas a um estilo de vida parasítico (HU; ST. LEGER, 2004), ou mesmo a diferentes ambientes fora do hospedeiro, como a rizosfera de plantas (ST LEGER, 2008).

Até o momento de publicação, esse trabalho foi o mais abrangente da evolução das serino-proteases tipo-subtilisina Classe II em *Metarhizium* spp., evidenciando padrões de diversificação previamente desconhecidos nessa família de fatores de virulência. A identificação de seleção positiva e divergência funcional na sf1 e na sf2 é consistente com o aumento esperado nas taxas evolutivas para genes duplicados envolvidos na infecção de múltiplos hospedeiros (LI et al., 2017; VILCINSKAS, 2010). Enquanto regiões funcionalmente relevantes aparentam estar sendo afetadas, mais estudos são necessários para confirmar a importância de resíduos positivamente selecionados e associados à divergência funcional nessas proteínas, de modo a identificar os efeitos desses sítios na estrutura e função proteica. Dessa maneira, os resultados deste Capítulo 1 dão base teórica aos Capítulos 2 (Estrutura) e 3 (Função), a seguir.

Capítulo 2: Estrutura

Com base nas avaliações filogenéticas e de pressão seletiva na família gênica Pr1, buscamos descrever *in silico* o comportamento molecular da potencial nova proteoforma de protease Pr1J, visando analisar diferenças conformacionais e físico-químicas desta em relação à forma "canônica". Para tal, foram construídos os modelos teóricos de estrutura tridimensional das proteases Pr1J1 e Pr1J2 por modelagem comparativa[27] com proteínas cuja estrutura já é conhecida. Após, realizaram-se simulações por dinâmica molecular visando caracterizar as conformações mais prevalentes e evidenciar detalhes que permitam a diferenciação das proteoformas de protease Pr1J.

[27] Ou por homologia

Procedimentos Metodológicos

Modelagem Comparativa

A partir da projeção estrutural descrita no Capítulo 1[28], utilizou-se como molde a estrutura cristalográfica da Proteinase K de *Parengyodontium album* (PDB ID 1IC6; (BETZEL et al., 2001)), determinada por difração de raios-X a uma resolução de 0.98 Å. Uma avaliação pelo servidor PDBSum (LASKOWSKI et al., 2018)[29] indicou que o peptídeo maduro inicia imediatamente após o domínio I9, sendo que apenas as regiões correspondentes nas proteases Pr1J1 (NCBI KFG80219.1, aminoácidos (aa) 117-398) e Pr1J2 (NCBI KFG85392.1, aa 111-398) foram utilizadas. As sequências foram alinhadas com a sequência aminoacídica do molde utilizando o alinhador *online* PROMALS3D (PEI; KIM; GRISHIN, 2008), que foram utilizados como entrada para o *software* Modeller 9.19 (SALI; BLUNDELL, 1993). Foram construídos 10 modelos teóricos, com otimização estrutural por dinâmica molecular, posteriormente avaliados comparativamente pelo escore DOPE (*Discrete Optimized Potential Energy*). O *script* completo encontra-se no Anexo 1. A estrutura de menor DOPE foi submetida à ferramenta PROCHECK (LASKOWSKI et al., 1993), hospedada no servidor PDBSum, para aquisição dos perfis de estrutura secundária e avaliação de perfil de ângulos diedrais pelo gráfico de Ramachandran-Ramakrishnan-Sasisekharan, doravante apenas tratado como gráfico de Ramachandran (RAMACHANDRAN; RAMAKRISHNAN; SASISEKHARAN, 1963). Foram consideradas aceitáveis as estruturas pela inequação:

$$RMF + RAF \geq 95\% \qquad (1)$$

onde RMF é o número de resíduos em regiões mais favoráveis no gráfico e RAF, em regiões adicionalmente favoráveis.

[28] (ANDREIS; SCHRANK; THOMPSON, 2019)*Online Resource 6*
[29] https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1IC6

Dinâmica Molecular

As estruturas obtidas foram submetidas a simulações por Dinâmica Molecular (DM) em triplicata técnica utilizando o pacote GROMACS 2020.1 (LINDAHL et al., 2020a, 2020b), seguindo as etapas aqui descritas (baseadas nos protocolos disponíveis no manual do *software* (LINDAHL et al., 2020a)). Inicialmente, foi construída a topologia da proteína[30] e suas interações descritas segundo o campo de força AMBER99SB-ILDN (LINDORFF-LARSEN et al., 2010), bem como a configuração de solvente explícito (água) do tipo TIP3P (JORGENSEN et al., 1983). A proteína foi inserida em uma caixa dodecaédrica, mantendo uma distância mínima de 1,0 Å das bordas, aplicando-se condições periódicas de contorno nas três dimensões espaciais. O sistema foi solvatado[31] e foram adicionados contra-íons Na^+ e Cl^-, conforme necessário, para a neutralização da carga total do sistema. Em seguida, o sistema foi submetido à minimização de energia potencial segundo o algoritmo *Steepest Descent*, com passo de minimização de 0,01 ps, em um máximo de 50.000 passos ou quando o sistema atingir força máxima inferior a 1000,0 kJ mol^{-1} nm^{-1}. Após, foi realizada a geração de velocidades iniciais e acomodação do solvente em torno da proteína (com restrição de posição) permitindo variação da pressão do sistema, mantendo volume e temperatura[32] constantes (NVT; Anexo 2).

A partir deste ponto, comum a todas as réplicas, iniciou-se a amostragem em triplicata técnica do sistema. Seguiu-se de forma análoga a equilibração NVT, porém fixando a pressão e temperatura e permitindo variação de volume (NPT; Anexo 3), mantendo as restrições de posição na proteína, até a equilibração do volume do sistema. Por fim, executou-se a fase de produção, simulando o sistema (agora sem restrição de posições) por 200 ns, utilizando o integrador *leap-frog* com passo de integração de 2 fs (Anexo 4). As simulações foram avaliadas[33] em relação a RMSD (*Root Mean Square Deviation*), quantificando variação de

[30] 1IC6, Pr1J1 ou Pr1J2
[31] Adição de moléculas de água na caixa teórica
[32] 300,15K ou 27°C
[33] Através das médias±erro padrão das triplicatas.

coordenadas em relação à estrutura inicial (NPT), e RMSF (*Root Mean Square Fluctuation*) por resíduo, quantificando a variação posicional das cadeias laterais de cada resíduo ao longo de toda a simulação. Ambos foram calculados com o pacote GROMACS. Resíduos associados à seleção positiva e divergência funcional serão avaliados em relação às ligações de hidrogênio em pontos-chave da simulação.

Resultados e Discussão

Modelagem Comparativa

Dos dez modelos construídos para as proteínas Pr1J1 e Pr1J2, foram escolhidos os modelos de números 9 e 10 (Tabela 3), respectivamente. Ambos, além da estrutura cristalina 1IC6[34], foram avaliados quanto ao impedimento estérico (gráfico de Ramachandran), acessibilidade a solvente e progressão de estruturas secundárias observadas. A estrutura-molde deve ser inspecionada por efeitos induzidos no processo de cristalização (empacotamento cristalino) e demais anormalidades que podem influenciar a predição das estruturas. Pelas estatísticas fornecidas pelo PDBSum/PROCHECK (Figura 6; Anexos 5, 6 e 7): a estrutura 1IC6 apresenta 100% de resíduos (não-glicina e não-prolina) em RMFs e RAFs; Pr1J1 apresenta 98,8%; Pr1J2 apresenta 97,9%. Os resíduos em regiões adicionais de Pr1J1 correspondem às posições T106, R129 e K250, estando localizados em zonas limítrofes de estrutura secundária, não aparentando, em princípio, serem de grande influência nas análises subsequentes. Com relação à Pr1J2, os resíduos D31, P56, Y177, T174 e S173 encontram-se em regiões menos favoráveis, também localizadas em zonas limítrofes de estrutura secundária, ou desprovidas de elementos secundários.

Em primeira vista (pré-simulação), a estrutura-molde e os modelos construídos apresentam convergência na progressão de estruturas secundárias. Em particular, verifica-se maior semelhança entre o cristal 1IC6 e o modelo estrutural de Pr1J1 que com Pr1J2 (Figura 6). Ademais, quando realizadas equivalências posicionais com base em alinhamentos, verifica-se deleção em Pr1J2 na região compreendida entre os resíduos 30 e 43, um conjunto de inserções e deleções nas regiões 71-76 e 115-117, inserção em 190-193 e 240-241 em Pr1J2 (com potencial influência na estrutura secundária). Com relação a ligações dissulfeto, anotadas na estrutura-molde, observou-se que não

[34] https://www.ebi.ac.uk/pdbe/entry/pdb/1ic6

há conservação de cisteínas nas posições homólogas, tanto para o par C41-C138 quanto para C201-C275. Isso sinaliza que é possível esperar menos restrição dinâmica nas estruturas terciárias quando em solução para as estruturas-modelo. Sítios de ligação a íons cálcio (Ca^{+2}) não aparentam estar conservados. Por fim, a tríade catalítica composta por D46, H82 e S248 encontra-se conservada nas três estruturas.

Com base nas avaliações filogenéticas e de pressão seletiva na família gênica Pr1, buscamos descrever *in silico* o comportamento molecular da potencial nova proteoforma de protease Pr1J, visando analisar diferenças conformacionais e físico-químicas desta em relação à forma "canônica". Para tal, foram construídos os modelos teóricos de estrutura tridimensional das proteases Pr1J1 e Pr1J2 por modelagem comparativa com proteínas cuja estrutura já é conhecida. Após, realizaram-se simulações por dinâmica molecular visando caracterizar as conformações mais prevalentes e evidenciar detalhes que permitam a diferenciação das proteoformas de protease Pr1J.

Tabela 3: Funções de pontuação para avaliação de modelos teóricos, conforme fornecidas pelo MODELLER. O sombreado indica o modelo selecionado para simulação.

	Pr1J1		Pr1J2	
Número	**molpdf**	**DOPE**	**molpdf**	**DOPE**
1	1.435,25	-30.315,06	5.838,83	-26.373,41
2	1.509,97	-30.296,72	6.173,85	-25.427,11
3	1.568,99	-29.992,55	5.912,45	-25.574,84
4	1.459,64	-30.475,02	5.847,91	-26.858,13
5	1.440,01	-30.199,05	5.852,84	-26.115,52
6	1.444,12	-30.157,10	5.926,18	-26.330,27
7	1.441,64	-30.488,49	5.848,53	-26.504,05
8	1.541,57	-30.439,34	6.154,08	-26.245,43
9	1.412,49	-30.580,62	5.866,88	-26.571,08
10	1.469,79	-30.478,91	5.986,23	-26.945,24

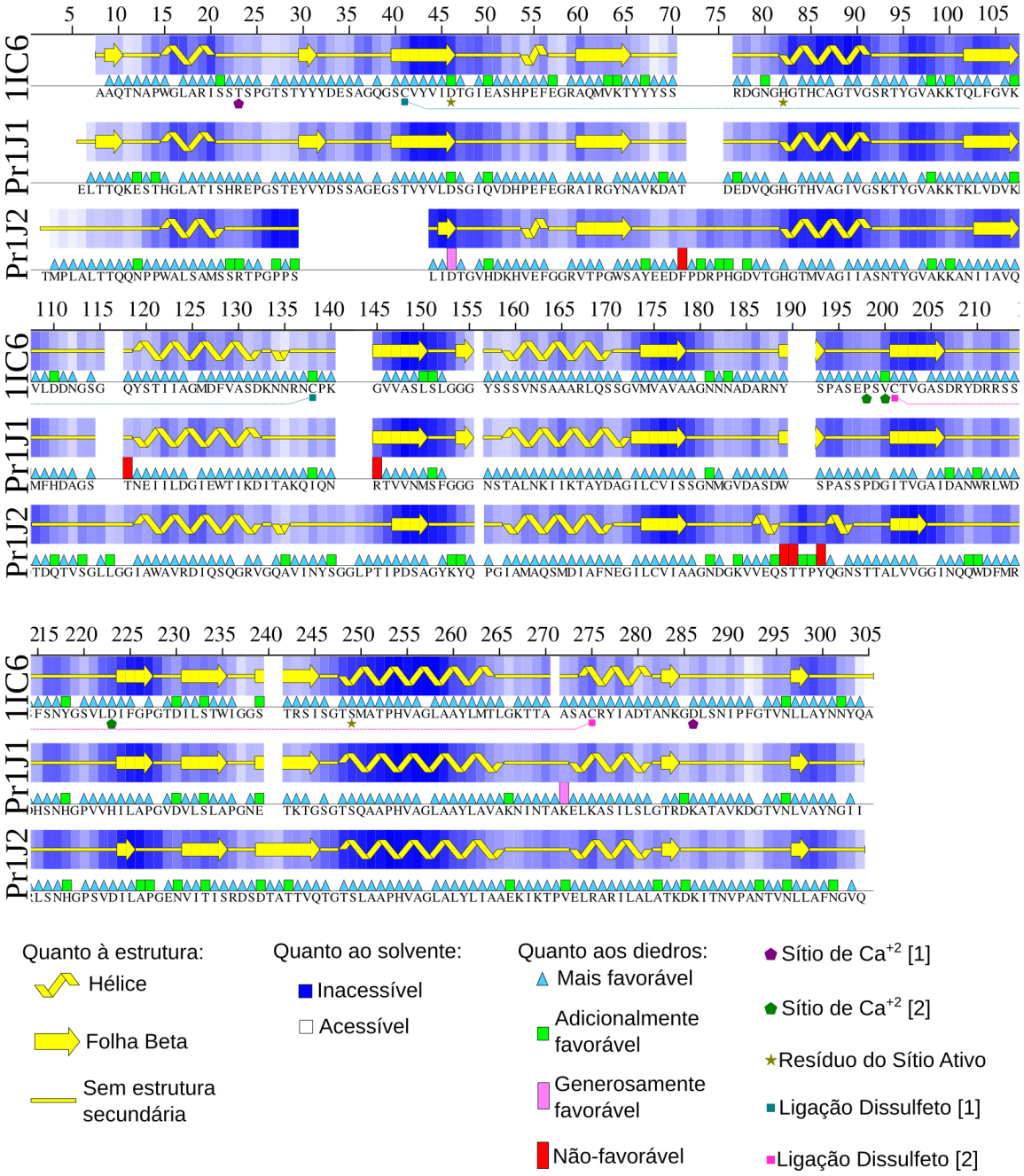


Figura 6: Comparativo de propriedades da estrutura-molde (1IC6) com modelos teóricos (Pr1J1 e J2). Apresentam-se, de cima para baixo, informações de progressão de estrutura secundária (diferentes formas em amarelo), acessibilidade ao solvente (gradientes de azul a branco, em retângulos correspondendo a cada resíduo), posicionamento no gráfico de Ramachandran e propriedades de resíduos do molde. Posições de resíduos homólogos estão ajustadas conforme alinhamento do PROMALS3D.

Dinâmica Molecular

Para avaliação do *ensemble* conformacional das simulações em triplicata (Anexos 8, 9, 10) para estruturas terciárias teóricas de Pr1J1 e Pr1J2, bem como da estrutura cristalográfica de código PDB 1IC6 (Figura 7), foi necessário tomar estatísticas das três simulações. Dados de RMSD de cada simulação foram conjuntamente representados por $\overline{RMSD} \pm s$, onde $\overline{RMSD}$ representa a média das amostras e o desvio padrão amostral de cada passo temporal da simulação (Anexo 11), suavizando a curva através da tomada dos valores médios $\overline{RMSD}$ e s em janela deslizante de 500 ps (Anexo 12). Para o RMSF, valores conjuntos de flutuação por resíduo foram representados pelos seus desvios-padrão. *Scripts* de construção dos gráficos com o *software* Gnuplot ("gnuplot homepage", [s.d.]) encontram-se nos Anexos 13 e 14. Gráficos com réplicas individualizadas de RMSD e RMSF para 1IC6, Pr1J1 e Pr1J2 encontram-se, respectivamente, nos Anexos 15, 16 e 17.

Analisando os conjuntos de simulações individualmente, verifica-se uma diferença intrínseca da flexibilidade de cadeias laterais, descrita pelo desvio-padrão de RMSF por resíduo (Figura 7, direita), entre as proteínas. Na simulação da Proteinase K 1IC6, observa-se uma amplitude de desvio-padrão de $0{,}99 \times 10^{-3}$ (V155) a 0,93 (A279) Å, enquanto para os modelos teóricos Pr1J1 e Pr1J2, esses valores vão de $2{,}52 \times 10^{-3}$ (A206) a 1,92 (V70) Å e de $9{,}03 \times 10^{-3}$ (V170) a 4,4 (T1) Å, respectivamente. Para contexto, o raio covalente de um átomo de hidrogênio é de aproximadamente 0,31 Å ("Periodic Table", 2019) e o comprimento de uma ligação de hidrogênio varia de 1,6 a 2,0 Å, aproximadamente (LEGON; MILLEN, 1987). A correlação com as amplitudes de RMSF, somente, para os *ensembles* implica que provavelmente não ocorreram rearranjos significativos da rede de ligações de 1IC6 durante as simulações, enquanto são mais prováveis nas estruturas teóricas (J1 < J2). Naturalmente,

essas suposições apenas podem ser confirmadas analisando o comportamento das cadeias laterais em seu contexto molecular ao longo da simulação. Se analisarmos a região compreendida entre os resíduos do sítio ativo no início das simulações (expandidas para compreender elementos de estrutura secundária apresentados na Figura 6)[35], fica evidente que rearranjos de cadeias laterais são mais abundantes e amplos nas estruturas teóricas e, dentre elas, são mais pronunciados em Pr1J2.

Com relação à manutenção geral do enovelamento proteico, indiretamente representada pelo RMSD ao longo do tempo (Figura 7, gráfico inferior), verifica-se um padrão similar ao observado nos dados de RMSF. Tomando a estrutura resultante da acomodação NVT como ponto de referência para as coordenadas, observa-se a menor variação em 1IC6 (0,3052±0,0075 Å [$t = 0\,ns$] a 0,9434±0,1879 Å [$t = 115,87\,ns$]), variação intermediária em Pr1J1 (0,5137±0,0084 Å [$t = 0\,ns$] a 2,2895±0,2921 Å [$t = 118,82\,ns$]) e a maior variação em Pr1J2 (0,8260±0,0126 Å [$t = 0\,ns$] a 5,9472±1,3290 Å [$t = 192,59\,ns$]). Da mesma forma, essas diferenças se refletem nas conformações assumidas em $t = 200\,ns$, observando as posições relativas dos resíduos da tríade catalítica e dos aminoácidos sob seleção positiva (Figura 7, colunas 1, 2 e 3). Para melhor observar essas variações conformacionais, seria benéfico avaliar como os padrões de estrutura secundária variam ao longo do tempo de simulação, empregando conjuntamente o pacote GROMACS com o programa DSSP (KABSCH; SANDER, 1983). Dessa forma, os elementos de estrutura secundária são definidos para cada passo da simulação, permitindo uma melhor visão de sua manutenção.

[35] Figura 6, posições 39 a 264

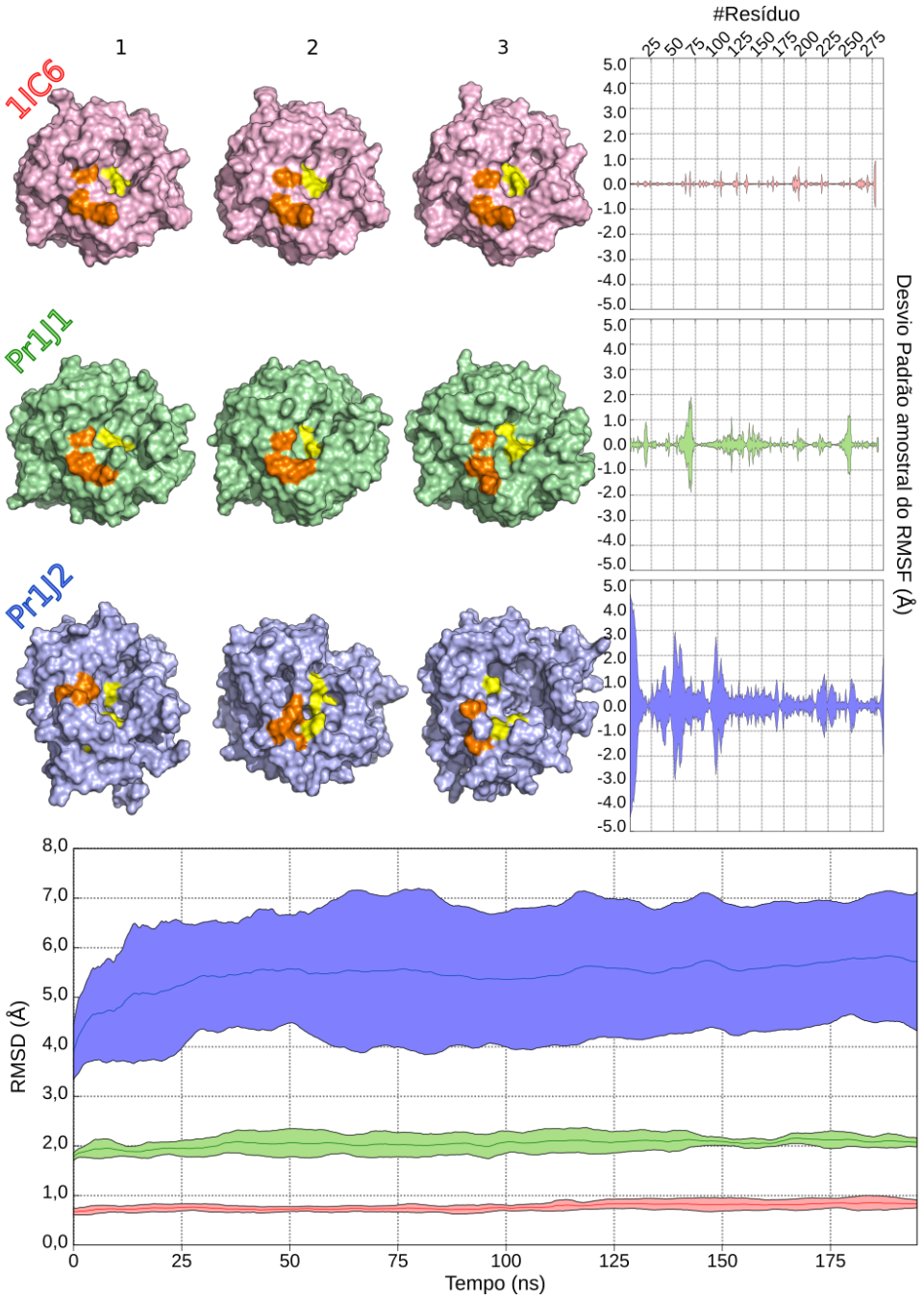


Figura 7: Simulações por Dinâmica Molecular. Apresentam-se as estruturas assumidas no quadro final da simulação ($t=200\ ns$) da simulação controle (PDB ID 1IC6), Pr1J1 e Pr1J2, com cada coluna representando a superfície de cada réplica (1, 2 e 3). A superfície em amarelo refere-se aos resíduos da tríade catalítica. A superfície em laranja representa resíduos positivamente selecionados em Pr1J. Os gráficos laterais apresentam os desvios padrão amostral dos valores de *Root Mean Square Fluctuation* (RMSF) por resíduo. O gráfico inferior apresenta média±desvio padrão (linha e área, respectivamente) de 3 simulações para cada sistema.

Em conjunto, essas observações parecem reforçar a maior similaridade estrutural e, por conseguinte, funcional, entre 1IC6 e Pr1J1. Contudo, ressalta-se a não-conservação de sítios de cisteína e dos resíduos de coordenação com Ca^{+2} das estruturas-modelo em relação à estrutura-molde. Isso tem como efeito prático estruturas menos restritas em solução, em particular em relação à coordenação dos íons cálcio, tidos como necessários ao enovelamento e estabilidade de proteases tipo-subtilisina (GALLAGHER; BRYAN; GILLILAND, 1993; SIEZEN; LEUNISSEN, 1997). Adicionalmente, na Proteinase K, íons Ca^{+2} propriamente ligados parecem estar associados com aumento na afinidade de ligação a substratos (LIU et al., 2011). Convém apontar que as simulações não contaram com forças adicionais para restrição das posições de Ca^{+2}, utilizando apenas os parâmetros do campo de força AMBER99SB-ILDN para interação. Visto que a simulação adequada de complexos metálicos por DM clássica (i.e. sem empregar Mecânica Quântica) é um desafio corrente no campo (VERLI, 2014), a própria escolha metodológica pode ter enviesado os resultados nesse sentido. Como alternativa, as simulações podem ser refeitas empregando métodos híbridos QM-MM (*Quantum Mechanics-Molecular Mechanics*), que são capazes de calcular o comportamento de nuvens eletrônicas desse tipo de interação (SENN; THIEL, 2009).

Do que foi possível constatar, estudos *in silico* de estrutura molecular envolvendo proteases Pr1 fúngicas são escassos, especialmente com proteoformas diferentes da “canônica” (i.e. Pr1A). Liu e colaboradores (2007) estudaram proteases envolvidas na degradação de cutícula de FAPs (Pr1; apresentando similaridades com Pr1A, conforme entrada Uniprot P29138) e de fungos nematófagos (Ver112 e VCP1) utilizando modelagem comparativa contra a Proteinase K 1IC6. Observou-se compartilhamento de elementos estruturais[36], condições de reação, além de alta identidade de sequência. Apesar das similaridades, verificou-se, à época, resíduos variáveis dentro dos sítios de ligação a substratos e diferentes flexibilidades conformacionais em cavidades da

[36] Especificamente: pontes dissulfeto e sítios de ligação a cálcio

região catalítica[37], com efeitos esperados em especificidade de substrato e atividade catalítica. Apesar da importância das observações, não é esperado que estruturas estáticas reflitam adequadamente as propriedades dinâmicas de proteínas em solução, visto que a função biológica é intrínseca aos movimentos físicos das biomoléculas (HENZLER-WILDMAN; KERN, 2007). Recentemente, estudos de DM da proteína Ver112 derivada de *Lecanicillium psalliotae* (YANG et al., 2019), um fungo nematófago, reforçam essa discrepância. A presença de um núcleo estrutural rígido provê aumento da termoestabilidade e resistência contra autólise ou proteólise, enquanto uma superfície vizinha flexível (especialmente em alças na superfície) é pré-requisito para o desempenho da função enzimática, permitindo a acomodação e orientação do substrato, orientação, catálise e liberação do produto. Ressalta-se, porém, que essas alterações conformacionais não estão restritas às adjacências da região catalítica, visto que flutuações distantes dessa zona podem afetar a dinâmica do sítio de ligação ao substrato por modos concertados, mecanismos de dobradiça e outras interações de elementos de estrutura terciária.

Comparando o presente estudo com estudos prévios, verificam-se diferenças notáveis na construção dos experimentos. Especificamente, Liu e seus colaboradores (2007) apontaram como características ótimas para a protease Pr1A um pH alcalino entre 8 e 10 e temperaturas de 50 a 60 ℃. Yang e associados (2019) aplicaram o campo de forças do tipo *united-atom* GROMOS 43a1 a uma temperatura de 300 K (≈27°C). Nossas simulações para Pr1J foram realizadas sob o campo de forças *all-atom* AMBER99SB-ILDN a uma temperatura de 300,15 K (27°C) em pH 7. Em vista dessas diferenças, pode ser benéfico refazer as simulações com pH entre 8 e 10, mantendo a temperatura em 27 °C, visto que essa faixa alcalina pode trazer alterações nos estados de protonação de cadeias laterais, em particular para cisteínas ($pK_{\text{cadeia lateral}} = 8{,}18$; "Amino Acids Reference Chart", [s.d.]). Notadamente, os dados aqui apresentados são deveras iniciais e mais análises são necessárias para melhor avaliar os comportamentos

[37] Indiretamente, medindo a variação de ligações de hidrogênio e pontes salinas.

das proteínas simuladas em solução. Uma descrição completa de proteínas requer uma superfície multidimensional de energia que define as probabilidades dos estados conformacionais e barreiras energéticas entre eles, sendo que muitos processos biológicos são regidos por alterações em taxas e populações conformacionais, em vez de um simples padrão liga-e-desliga (HENZLER-WILDMAN; KERN, 2007). Métricas ainda não avaliadas incluem o raio de giração[38], avaliação de ligações de hidrogênio ao longo do tempo para cadeias laterais específicas, tempo de retenção de íons cálcio, variação do volume da cavidade catalítica no tempo, medição da área de superfície acessível ao solvente (SASA), dentre outras. Adicionalmente, a simulação de um complexo proteína-substrato pode fornecer detalhes relevantes do processo catalítico. Substratos sintéticos utilizados para ensaios de cinética enzimática como Succinil-Ala(x3)-*p*-nitroanilina, Succinil-Ala(x2)-Pro-Phe-*p*-nitroanilina e caseína são candidatos promissores, visto que, além de auxiliarem na avaliação dinâmica do complexo proteína-substrato, servem como ponto de ligação com os experimentos propostos no Capítulo 3.

[38] Ou raio de giro; R_g

Capítulo 3: Função

Em paralelo à caracterização *in silico* das proteases Pr1J de *M. anisopliae*, buscamos caracterizar e diferenciar *in vitro* as atividades das proteoformas Pr1J1 (“canônica”) e J2 (“teórica”). Para tal, construímos vetores plasmidiais visando induzir a expressão dessas proteínas em *Escherichia coli*, de modo a obter quantidades suficientes para ensaios de atividade enzimática que permitam evidenciar diferenças entre Pr1J1 e Pr1J2 em nível funcional.

Procedimentos Metodológicos

Esses experimentos foram realizados em colaboração com o Dr. Nicolau Sbaraini e com a Dra. Julia Catarina Vieira Reuwsaat e encontram-se em desenvolvimento, sendo que o que segue reflete o estado corrente dos resultados parciais obtidos. A Figura 8 provê um panorama da metodologia empregada.

Linhagens, plasmídeos, meios de cultura e outros preparos

Para as etapas de clonagem e manutenção de plasmídeos foram utilizadas as linhagens DH5α e 10 de *E. coli*. Para expressão heteróloga, a linhagem BL21(DE3)pLysS de *E. coli* foi utilizada. Utilizou-se o plasmídeo pET23d(+) (Figura 9, superior), contendo marca de seleção para ampicilina, como vetor de expressão. Para cultivo, utilizou-se o meio Luria-Bertani (LB) (BERTANI, 1951) acrescido de ágar (15 g/L) quando necessário cultivo sólido, e/ou ampicilina a 100 µg/mL quando necessário meio seletivo. Procedimentos de eletroforese foram realizados utilizando gel de agarose 0,9% contendo 0,05% de brometo de etídeo.

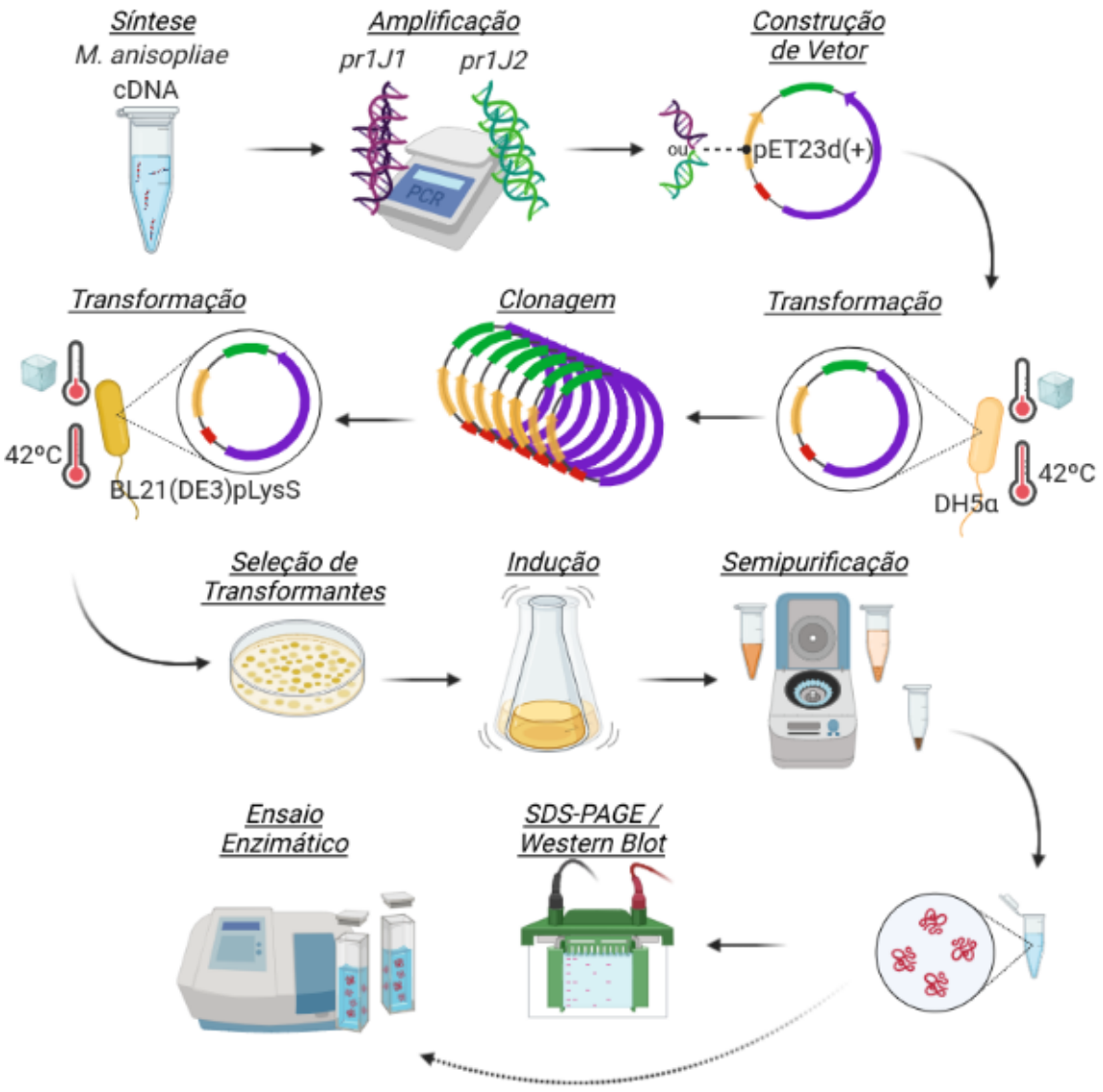


Figura 8: Panorama metodológico do Capítulo 3. Construído com BioRender.com

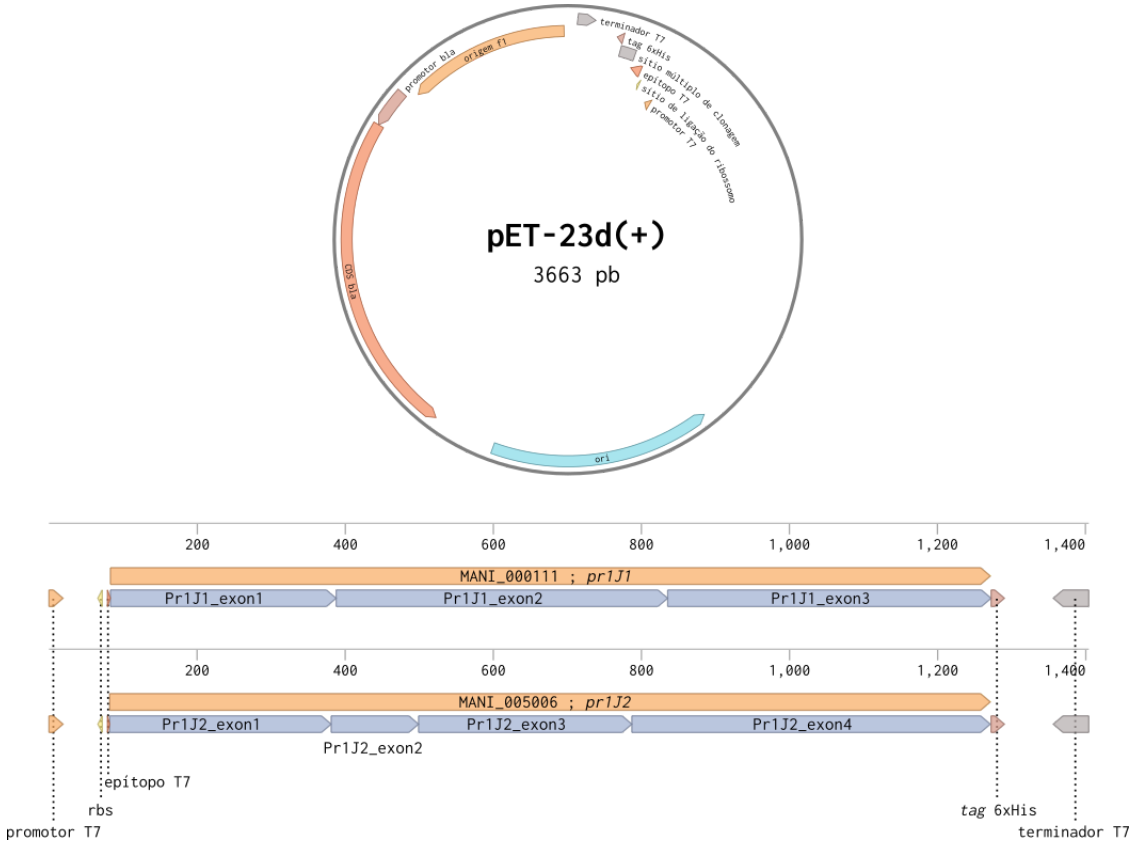


Figura 9: Mapa gráfico dos vetores plasmidiais utilizados. Superior: vetor pET23d(+). **Inferior:** localização dos genes *pr1J1* e *pr1J2* em relação ao plasmídeo. *bla*: gene codificante de β-lactamase, conferindo resistência a ampicilina. *ori*: origem de replicação. *rbs:* sítio de ligação do ribossomo

Material biológico e síntese de cDNA

Foi utilizado como molde o RNA extraído de culturas de *M. anisopliae,* linhagem E6, gentilmente cedido pelo Dr. Nicolau Sbaraini (ARRUDA et al., 2005; SBARAINI et al., 2019). Alíquotas de RNA extraído (2 µg) foram tratadas com DNAse RQ1 (Promega), segundo protocolo especificado pela fabricante, sendo posteriormente quantificadas e diluídas a uma concentração de 100 ng/µL. A fim de validar se o tratamento com DNAse foi efetivo na digestão de DNA genômico contaminante, realizou-se PCR (Taq DNA Polimerase, Ludwig Biotecnologia) com oligonucleotídeos específicos para o gene da tubulina[39]. A ausência de

[39] Par de oligonucleotídeos 1 (Anexo 18)

amplificação indica que a amostra está livre de DNA genômico em quantidade detectável.

Confirmada a eficácia do tratamento, realizou-se a síntese de cDNA. Inicialmente, adicionaram-se 700 ng de ácido nucleico a 9,6 µL de água DEPC e 2 µL do oligonucleotídeo CDS (10 µM)[40]. A solução foi incubada a 70 °C por 5 min e em gelo por 5 min para anelamento de oligonucleotídeos. As amostras foram então incubadas com a enzima transcritase reversa (ImProm-II; Promega), segundo especificação da fabricante, à temperatura ambiente por 5 min e a 42°C por 1 h. O produto final foi diluído à concentração de 10 ng/µL. Foram então realizadas três PCRs (Taq DNA Polimerase, Ludwig Biotecnologia) com: (1) oligonucleotídeos específicos para o gene da tubulina[41], para confirmação de integridade; (2) oligonucleotídeos específicos para o gene da Pr1J1 (JNNZ01000177.1)[42], para identificação; (3) oligonucleotídeos específicos para o gene da Pr1J2 (JNNZ01000033.1)[43,44], também para identificação.

Clonagem

Uma vez confirmada a síntese de cDNA referente aos genes *pr1J1 e pr1J2*, realizamos novas PCRs utilizando a DNA polimerase de alta fidelidade Q5 (New England Biolabs) com os oligonucleotídeos para anelamento nas regiões de *pr1J1* e *pr1J2*. As amostras foram fracionadas em gel de agarose e as bandas referentes aos produtos de amplificação foram excisadas, solubilizadas e purificadas utilizando o *kit* comercial *NucleoSpin Gel and PCR Clean-up* (Macherey-Nagel), seguindo protocolo recomendado pela fabricante. A quantidade de DNA na solução resultante foi quantificada usando *NanoDrop*[45].

[40] Par de oligonucleotídeos 2 (Anexo 18)
[41] Par de oligonucleotídeos 1 (Anexo 18)
[42] Par de oligonucleotídeos 3 (Anexo 18)
[43] Par de oligonucleotídeos 4 (Anexo 18)
[44] Ambos os códigos de acesso referem-se aos contigs do genoma sequenciado. Referir à anotação para posicionamento exato.
[45] NanoDrop Lite (Thermo Scientific)

A construção dos plasmídeos pET23D(+)-*pr1J1* e pET23D(+)-*pr1J2* (Figura 9, inferior) foi realizada pela reação de *Hot Fusion* (FU et al., 2014), adicionando em tubo de 0,2 mL: 0,5 µL de DNA do vetor pET23D(+) linearizado com as enzimas de restrição NcoI e XhoI (50 ng/ul); 4,5 µL do amplicon *pr1J1* ou pr1*J2*; 5 µL de solução *Hot Fusion* 2x (Tris pH 7,5 a 0.2 M; $MgCl_2$ a 20 mM; dNTPs a 0.4 mM; DTT a 20 mM; 10% PEG-8000; exonuclease T5 a 0,0075 U/µL; DNA polimerase *Phusion Hot Start* a 0,05 U/µL). Os tubos foram incubados em termociclador por 1 h a 50 °C e resfriados até 20 °C em decréscimos de 0,1 °C por segundo para amplificação.

Os produtos foram empregados para a transformação de células termocompetentes de *E. coli*, realizada por choque térmico. Em tubo de 1,5 mL contendo 50 µL de células de *E. coli* DH5α ou 10 , adicionaram-se 5 µL dos produtos da reação de *Hot Fusion*. As células foram incubadas por 20 min em gelo, seguido de choque térmico a 42 ℃ por 90 s. Após o choque térmico, adicionaram-se 400 µL de LB líquido ao tubo contendo as células. As culturas foram, então, incubadas em estufa a 37 ℃ por 1 h, homogeneizadas por inversão a cada 30 min, e transferidas para placas de petri[46] contendo meio seletivo sólido. As placas foram incubadas por 14~16 h em estufa a 37 ℃. Por fim, tendo sido obtidas colônias potencialmente abrigando os plasmídeos de interesse, reações de PCR de até 10 colônias foram realizadas, a fim de identificar positivos.

As colônias de *E. coli* potencialmente contendo o plasmídeo pET23d(+)-*pr1J1* ou -*pr1J2* foram inoculadas em tubos de ensaio contendo 5 mL de meio líquido em presença de ampicilina à concentração de 50 µg/mL e incubadas em plataforma orbital a 37 ℃ com agitação de 200 rpm por 14~16 h. Os cultivos que apresentaram crescimento celular[47] foram submetidos à extração de DNA plasmidial (*miniprep*).O DNA extraído foi solubilizado em 30 µL de água, sendo tratado com RNAse (1 µl de RNAse 100 mg/mL por 1 h a 37℃). A fim de confirmar a correta montagem dos vetores de expressão (pET23d(+)-*pr1J1* e

[46] Aproximadamente 250µL em cada
[47] Determinado pela turbidez da cultura

pET23d(+)-*pr1J2*) clivagens com enzimas de restrição e sequenciamento dos potenciais insertos foram realizados[48].

Expressão Heteróloga

Para expressão das proteases Pr1J de *M. anisopliae* em *E. coli*, transformou-se o plasmídeo pET23D(+)-*pr1J1* ou *pr1J2* na linhagem BL21(DE3)pLysS por choque térmico, confirmando os transformantes por PCR de colônia, conforme previamente descrito. Após, foi realizado pré-inóculo de colônias de transformantes em tubos de ensaio contendo 1 mL de meio LB suplementado com 1% de glicose e 100 µg/mL de ampicilina, incubados em banho-maria a 37°C e 200 rpm *overnight*. Em seguida, adicionaram-se 50 µL do pré-inóculo em 50 mL de meio líquido em erlenmeyers de 500 mL. A cultura foi incubada da mesma maneira até atingir OD_{600nm}[49] entre 0,4 e 0,6[50], momento no qual acrescentou-se IPTG à concentração de 1 mM para induzir a expressão da proteína recombinante e coletou-se alíquota de 1 mL em tubo de 3,5 mL (0h de indução). Em seguida, o inóculo foi incubado nas mesmas condições por 3h, realizando coletas a cada hora (tempo 1h, 2h, 3h). Ao final, o remanescente da cultura foi armazenado em tubo para estoque. Todas as alíquotas foram, no ato da coleta, centrifugadas a 13 krpm por 3 min e descartou-se o sobrenadante, congelando-se o *pellet*.

SDS-PAGE

O precipitado resultante da etapa de Expressão Heteróloga foi ressuspenso em 100 µL de tampão de amostra de SDS-PAGE (Tabela 4) com auxílio de vórtex e, após, fervidos por 10 min em banho-maria. As amostras foram centrifugadas a 14 krpm por 1 min à temperatura ambiente, separou-se o

[48] ACTGene Análises Moleculares
[49] Densidade óptica a 600 nm de comprimento de onda; aferida no espectrofotômetro BioMate 3S (Thermo Scientific)
[50] Aferida no espectrofotômetro NanoDrop Lite (Thermo Scientific)

sobrenadante e se ressuspendeu o *pellet* em 80 µL de tampão de amostra. Realizou-se SDS-PAGE[51] com 10 µL do sobrenadante para avaliação de solubilidade da proteína recombinante.

Tabela 4: Preparos para SDS-PAGE.

Tampão de amostra (10mL)	Glicerol	10% (v/v)
	β-mercaptoetanol	0,05% (v/v)
	SDS	2,3% (m/v)
	Tris-HCl 0,0625 M (pH 6,8)	0,125% (v/v)
	Azul de bromofenol	0,2% (m/v)
Tampão de corrida 5x (1 L)	Tris Base	15 g
	Glicina	94 g
	SDS	5 g
	Água destilada	1,496 mL
Gel 12%	Tris-HcL 1,5M (pH 8,8)	1,25 mL
	SDS 10%	50 µL
	Bis-Acrilamida	2 mL
	TEMED	4 µL
	APS 10%	100 µL
Gel 4%	Água MilliQ	1,525 mL
	Tris-HCl 0,5M (pH 6,8)	0,675 mL
	SDS 10%	25 µL
	Bis-Acrilamida	0,33 mL
	TEMED	5 µL
	APS 10%	50 µL

Western Blot

As amostras semipurificadas de proteína Pr1J1[52] recombinante foram separadas por SDS-PAGE e transferidas por método semi-seco para membrana de fluoreto de polivinildieno, posteriormente bloqueada com solução de leite. Foi

[51] Voltagem variando de 100 V (10 min) a 150 V (1h30min), até a corrida completa das amostras.
[52] Amostras de Pr1J2 não foram analisadas por ausência de indução de expressão gênica ()

executada sondagem indireta com anticorpos primários para hexa-histidina e anticorpos secundários conjugados a peroxidase. Foi realizada detecção de quimioluminescência utilizando substrato de *Western Blot* Pierce ECL (Thermo Scientific), conforme instruções do fabricante.

Resultados e Discussão

Vetores de Expressão

Após sequenciamento dos insertos *pr1J1* e *pr1J2* nos vetores pET23d(+) (Anexos 19 e 20), verifica-se a integridade da CDS de *pr1J1*, sem íntrons conforme anotação do genoma. Já para *pr1J2*, é possível observar a presença do intron 1[53], consenso do sequenciamento direto e reverso, e a potencial presença do intron 3[54], apontado pelo sequenciamento reverso e em parte pelo sequenciamento direto. Considerando a origem de mRNA/cDNA dos insertos, é possível que se trate de uma isoforma de *splicing* induzida pelas condições originais de cultivo. Alternativamente, pode se tratar de forma canônica do mRNA de *pr1J2* que não pôde ser inteiramente identificada no processo de anotação. As diferenças identificadas entre as CDSs montam a 129 pb, traduzidas para 43 aa, e podem explicar a variância conformacional observada nas simulações do Capítulo 2. Mais experimentos são necessários para confirmar ou refutar essas hipóteses.

Expressão Heteróloga

Pr1J1

Das colônias potencialmente transformadas com pET23d(+)-*pr1J1*, quatro foram selecionadas subjetivamente para expressão heteróloga. Corridas de SDS-PAGE sugerem indução bem-sucedida em *E. coli* BL21(DE3)pLysS (Figura 10, linha superior), evidenciada pelo aumento expressivo na espessura das bandas do controle positivo (proteína Krp1, de aproximadamente 43 kDa; (REUWSAAT et al., 2018)) conforme aumento do tempo de experimento. Quanto às colônias transformantes (Figura 10, linhas central e inferior), é possível

[53] Localizado entre os pares de bases 490 e 550 do *locus* MANI_005006
[54] Entre 1000 e 1070 pb de MANI_05006

identificar o mesmo padrão em 2, 3 e 4, ausente em 1. Variações na densidade de bandas nas canaletas é provavelmente decorrente da manipulação das amostras.

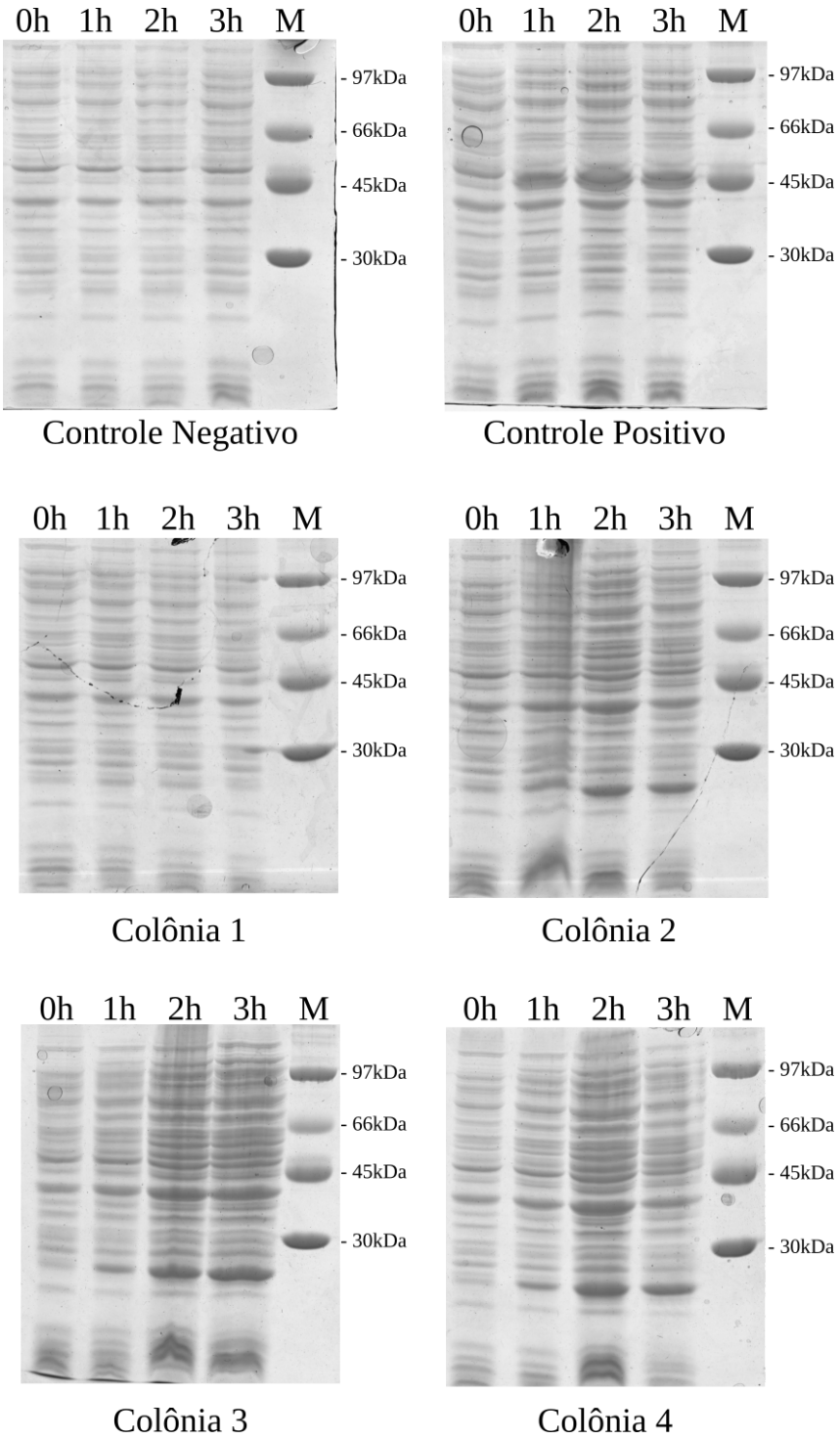


Figura 10: Indução de expressão de Pr1J1. Apresentam-se géis de SDS-PAGE para controles positivo e negativo (linha superior), colônias transformantes 1, 2, 3 e 4 (linhas intermediária e inferior) em coletas no início do processo de indução de expressão (0h) e 1h, 2h e 3h após indução. M: marcador de peso molecular.

Para confirmação da expressão da proteína Pr1J1 recombinante, utilizaram-se as coletas de 0h e 2h das colônias 2, 3 e 4 para avaliação via *Western Blot* (Figura 11). O padrão de migração das proteínas apresentou-se consistente com as corridas anteriores (Figura 10). Observa-se na revelação final (Figura 11, canto inferior direito) um sinal fortemente demarcado no controle positivo, em aproximadamente 43 kDa, indicando a presença da proteína Krp1 recombinante contendo cauda de histidina. Em contraste, é possível visualizar duas bandas sutis entre os marcadores de peso molecular 30 e 45 kDa, consistentes com o peptídeo Pr1J1 maduro (30,31 kDa; marcador ①) e o pró-peptídeo Pr1J1 (39,94 kDa; marcador ②). Contudo, a baixa intensidade das bandas e a discrepância quase ausente entre 0h e 2h sugere que devem se tratar de ligações inespecíficas do anticorpo anti-cauda de histidina. Dessa maneira, não é possível afirmar que a expressão da Pr1J1 recombinante foi bem-sucedida. Para confirmar ou refutar essa conclusão preliminar ainda é necessário analisar o precipitado das amostras de maneira similar.

Pr1J2

De forma análoga à realizada para Pr1J1, quatro colônias de bactérias potencialmente contendo pET23d(+)-*pr1J2* foram selecionadas para indução da expressão gênica (Figura 12). Pelas características dos controles negativo (ausência de alteração ao longo do tempo) e positivo (aumento na espessura das bandas em ~43 kDa, correspondendo à proteína Krp1), constata-se que o processo de indução foi bem-sucedido. Contudo, não foi possível expressar o plasmídeo recombinante pET23d(+)-*pr1J2,* visto que não há alterações nos padrões de bandas em SDS-PAGEs das quatro colônias selecionadas. É necessário reavaliar o processo inteiramente, visto que a presença de introns em *pr1J2* não havia sido considerada de antemão e impactam diretamente as etapas de amplificação do material genético.

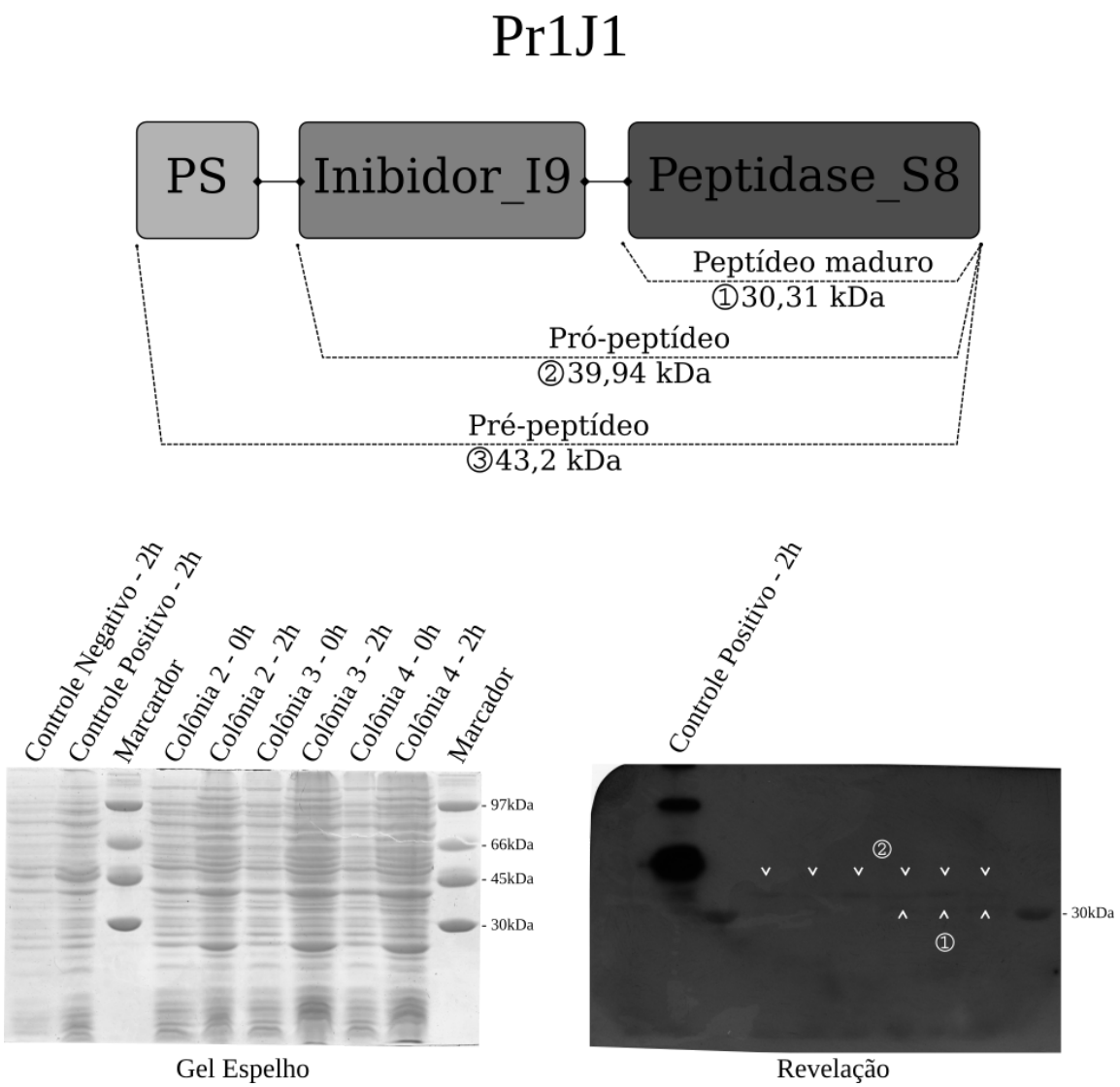


Figura 11: *Western Blot* de Pr1J1. Acima, apresenta-se esquema de composição do polipeptídeo Pr1J1, contendo massas moleculares correspondentes a cada segmento (pré-peptídeo, pró-peptídeo e peptídeo maduro). Abaixo e à esquerda apresenta-se gel espelho da corrida de SDS-PAGE com amostras selecionadas após indução de expressão. Abaixo e à esquerda apresenta-se a revelação final em filme fotográfico após tratamento com anticorpos marcados.

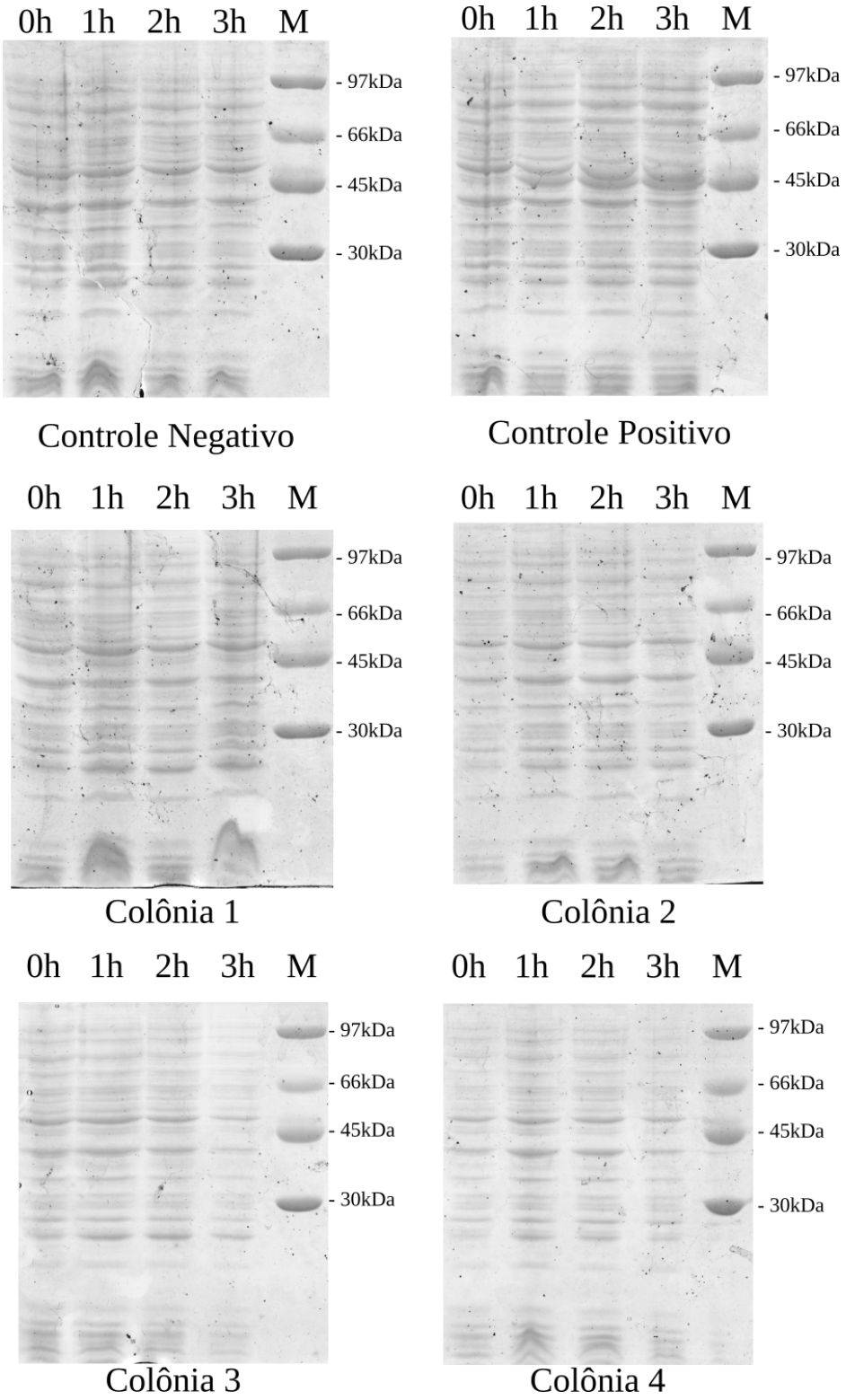


Figura 12: Indução de expressão de Pr1J2. Apresentam-se géis de SDS-PAGE para controles positivo e negativo (linha superior), colônias transformantes 1, 2, 3 e 4 (linhas intermediária e inferior) em coletas no início do processo de indução de expressão (0h) e 1h, 2h e 3h após indução. M: marcador de peso molecular.

Considerações Finais

Avaliando todos os dados apresentados, há suporte à tese de que a pressão seletiva nas proteases Pr1 pode ser quantificada e alterações decorrentes desse processo causam efeitos conformacionais e funcionais observáveis e detectáveis na estrutura proteica?

As análises filogenéticas e de pressão seletiva do Capítulo 1 sugerem que existe pressão seletiva positiva atuando nos genes *pr1*, promovendo variação nas proteínas codificadas e com potencial efeito funcional. Adicionalmente, é possível identificar resíduos associados à divergência funcional entre as duas proteoformas, dando suporte à não-redundância das proteínas. Existem indícios em projeções por homologia de que regiões funcionalmente relevantes estão sendo afetadas, requerendo experimentos mais detalhados para confirmação. Portanto, foi possível quantificar estatisticamente a pressão seletiva sobre os genes da família Pr1 de proteases em *M. anisopliae* e fungos relacionados. Adicionalmente, observando os padrões de ramificação filogenética, há evidência para a existência de duas proteoformas associadas à Pr1J, indicada pela prevalência do padrão evolutivo observado para as espécies que as contém, reforçando a prevalência de diversificação nessa família gênica.

Em sequência, buscamos descrever *in silico* as diferenças estruturais entre as proteoformas Pr1J1 e Pr1J2. As modelagens estruturais e simulações do Capítulo 2 apontam para diferenças teóricas em termos de composição de estrutura secundária[55] e variações conformacionais expressivas, incluindo na manutenção do sítio catalítico. Ressalta-se que a flutuação de coordenadas mostrou-se mais prevalente na proteína Pr1J2, sugerindo conformações potencialmente instáveis. Em contraste, o comportamento de Pr1J1 apresentou-se compatível com a estrutura de referência (PDB 1IC6). Entretanto, o comportamento dinâmico de resíduos identificados como alvo de pressão seletiva positiva no Capítulo 1 ainda requer avaliação mais detalhada, de modo a

[55] Padrões de inserções e deleções apresentados na Figura 6

evidenciar suas importâncias na manutenção de estrutura e/ou função. Logo, os efeitos da pressão seletiva sobre a estrutura de proteínas Pr1J permanecem inconclusivos.

Ensaios enzimáticos comparativos entre as proteoformas Pr1J1 e Pr1J2 podem fornecer detalhes cinéticos que permitiriam diferenciar ambas no âmbito funcional. Os experimentos retratados no Capítulo 3 foram uma tentativa de expressão heteróloga dos genes *pr1J1* e *pr1J2* de *M. anisopliae* em vetor bacteriano. Embora a construção dos vetores plasmidiais tenha sido aparentemente bem-sucedida, confirmada por sequenciamento, a efetiva expressão dos genes não foi possível. Alguns fatores complicadores se apresentaram, como a presença de elementos intrônicos em pET23d(+)-*pr1J2*, que exige avaliação detalhada de sua origem (canônica ou induzida). Entretanto, ainda é necessária a avaliação da fração insolúvel das alíquotas de indução para que se descarte completamente o sucesso do experimento. Até o momento, ainda são inconclusivas as diferenças na atividade enzimática das proteoformas Pr1J1 e Pr1J2 de *M. anisopliae*.

Tendo em vista o caráter parcial das atividades desenvolvidas nos capítulos 2 e 3, alguns caminhos podem ser vislumbrados. Quanto às simulações de estrutura e dinâmica proteica, pretende-se avaliar o comportamento dos resíduos apontados no Capítulo 1 como fonte de variabilidade, bem como a avaliação dos benefícios de novas simulações de Pr1J2 incorporando regiões supostamente intrônicas observadas no Capítulo 3. Em relação à expressão heteróloga, pretende-se reavaliar as condições experimentais para que a transformação e expressão seja possível, incluindo a substituição de linhagens bacterianas ou troca de vetores plasmidiais. Ademais, a realização de SDS-PAGE com a fração insolúvel das induções trará respostas quanto ao (in)sucesso do procedimento. Uma vez confirmada a expressão dos genes *pr1J1* e *pr1J2*, pretende-se purificar as proteínas por coluna de afinidade e, após, realizar ensaio enzimático com substratos cromogênicos, tais como Succinil-Ala(x3)-p-nitroanilina

e Succinil-Ala(x2)-Pro-Phe-p-nitroanilina, fornecendo informações comparativas diretas quanto às funções das proteases Pr1J1 e Pr1J2.

A proposta dessa tese de doutorado é explorar de forma abrangente o processo de diversificação em famílias gênicas de FAPs, desde características genético-evolutivas, passando pelos efeitos estruturais, dinâmicos e mecanísticos dessas variações, chegando a ensaios enzimáticos de atividade com proteínas purificadas. Essa abordagem mista teórico-prática/computacional-experimental apresenta-se sinergística, com potencial descritivo de escala celular à molecular maior que os experimentos isolados. As análises evolutivas aqui apresentadas acrescentam ao conhecimento sobre a evolução de famílias gênicas associadas à virulência de FAPs como *M. anisopliae*, bem como potenciais caminhos trilhados ao longo do processo evolutivo. Em conjunto, as avaliações de estrutura e atividade enzimática, quando concluídas, podem fornecer perspectivas únicas quanto à forma como evoluíram essas proteínas e como suas atividades vem se alterando até o tempo presente. Por fim, espera-se que as humildes contribuições desse autor sejam fonte de novas ideias e novos caminhos a serem construídos em conjunto com a comunidade científica.

Caro leitor, cara leitora,

Um forte abraço.

Bibliografia

ADL, S. M. et al. The revised classification of eukaryotes. **The Journal of eukaryotic microbiology**, v. 59, n. 5, p. 429–493, set. 2012.

ADL, S. M. et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. **The Journal of eukaryotic microbiology**, v. 66, n. 1, p. 4–119, jan. 2019.

Amino Acids Reference Chart. Disponível em: <https://www.sigmaaldrich.com/BR/pt/technical-documents/technical-article/protein-biology/protein-structural-analysis/amino-acid-reference-chart>. Acesso em: 29 set. 2021.

ANDREIS, F. C.; SCHRANK, A.; THOMPSON, C. E. Molecular evolution of Pr1 proteases depicts ongoing diversification in *Metarhizium* spp. **Molecular genetics and genomics: MGG**, v. 294, n. 4, p. 901–917, ago. 2019.

ARRUDA, W. et al. Morphological alterations of *Metarhizium anisopliae* during penetration of *Boophilus microplus* ticks. **Experimental & applied acarology**, v. 37, n. 3-4, p. 231–244, 2005.

BAGGA, S. et al. Reconstructing the diversification of subtilisins in the pathogenic fungus *Metarhizium anisopliae*. **Gene**, v. 324, n. 1-2, p. 159–169, jan. 2004.

BERTANI, G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. **Journal of bacteriology**, v. 62, n. 3, p. 293–300, set. 1951.

BETZEL, C. et al. Structure of a serine protease proteinase K from *Tritirachium album* limber at 0.98 A resolution. **Biochemistry**, v. 40, n. 10, p. 3080–3088, 13 mar. 2001.

BEYS-DA-SILVA, W. O. et al. Secretome of the Biocontrol Agent *Metarhizium anisopliae* Induced by the Cuticle of the Cotton Pest *Dysdercus peruvianus* Reveals New Insights into Infection. **Journal of proteome research**, v. 13, n. 5, p. 2282–2296, 2 maio 2014.

BEYS-DA-SILVA, W. O. et al. Updating the application of *Metarhizium anisopliae* to control cattle tick *Rhipicephalus microplus* (Acari: Ixodidae). **Experimental parasitology**, v. 208, p. 107812, jan. 2020.

BOOMSMA, J. J. et al. Evolutionary Interaction Networks of Insect Pathogenic Fungi. **Annual review of entomology**, v. 59, n. 1, p. 467–485, 7 jan. 2014.

BUTT, T. M. et al. Entomopathogenic Fungi: New Insights into Host–Pathogen Interactions. In: [s.l: s.n.].

BUTT, T. M.; JACKSON, C.; MAGAN, N. Introduction-fungal biological control agents: Progress, problems and potentials. **Fungi as Biocontrol Agents Progress, Problems and Potential; Butt, TM, Jackson, C. , Magan, N. , Eds**, p. 1–7, 2001.

FARIA, M. R. DE; WRAIGHT, S. P. Mycoinsecticides and Mycoacaricides: A comprehensive list with worldwide coverage and international classification of formulation types. **Biological control: theory and applications in pest management**, v. 43, n. 3, p.

237–256, dez. 2007.

FREIMOSER, F. M. et al. Expressed sequence tag (EST) analysis of two subspecies of *Metarhizium anisopliae* reveals a plethora of secreted proteins with potential activity in insect hosts. **Microbiology** , v. 149, n. Pt 1, p. 239–247, 1 jan. 2003.

FREIMOSER, F. M.; HU, G.; ST. LEGER, R. J. Variation in gene expression patterns as the insect pathogen *Metarhizium anisopliae* adapts to different host cuticles or nutrient deprivation in vitro. **Microbiology**, v. 151, n. 2, p. 361–371, 1 fev. 2005.

FU, C. et al. Hot Fusion: an efficient method to clone multiple DNA fragments as well as inverted repeats without ligase. **PloS one**, v. 9, n. 12, p. e115318, 31 dez. 2014.

GALLAGHER, T.; BRYAN, P.; GILLILAND, G. L. Calcium-independent subtilisin by design. **Proteins**, v. 16, n. 2, p. 205–213, jun. 1993.

gnuplot homepage. Disponível em: <http://www.gnuplot.info/>. Acesso em: 17 set. 2021.

HANE, J. K. et al. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. **Genome biology**, v. 12, n. 5, p. R45, 24 maio 2011.

HEALY, R. A. et al. Functional and phylogenetic implications of septal pore ultrastructure in the ascoma of *Neolecta vitellina*. **Mycologia**, v. 105, n. 4, p. 802–813, jul. 2013.

HENZLER-WILDMAN, K.; KERN, D. Dynamic personalities of proteins. **Nature**, v. 450, n. 7172, p. 964–972, 13 dez. 2007.

HU, G.; ST. LEGER, R. J. A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. **Journal of evolutionary biology**, v. 17, n. 6, p. 1204–1214, nov. 2004.

HU, X. et al. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 47, p. 1–6, 25 nov. 2014.

IWANICKI, N. S. A. et al. Monitoring of the field application of *Metarhizium anisopliae* in Brazil revealed high molecular diversity of *Metarhizium* spp in insects, soil and sugarcane roots. **Scientific reports**, v. 9, n. 1, p. 4443, 14 mar. 2019.

JAVAR, S. et al. Expression of pathogenesis-related genes in *Metarhizium anisopliae* when infecting *Spodoptera exigua*. **Biological control: theory and applications in pest management**, v. 85, p. 30–36, jun. 2015.

JORGENSEN, W. L. et al. Comparison of simple potential functions for simulating liquid water. **The Journal of chemical physics**, v. 79, n. 2, p. 926–935, 15 jul. 1983.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v. 22, n. 12, p. 2577–2637, dez. 1983.

KEPLER, R. M. et al. New teleomorph combinations in the entomopathogenic genus *Metacordyceps*. **Mycologia**, v. 104, n. 1, p. 182–197, jan. 2012.

KEPLER, R. M. et al. Clarification of generic and species boundaries for *Metarhizium* and related fungi through multigene phylogenetics. **Mycologia**, v. 106, n. 4, p. 811–829, 1 jul. 2014.

KOJIMA, S.; MINAGAWA, T.; MIURA, K. The propeptide of subtilisin BPN' as a temporary inhibitor and effect of an amino acid replacement on its inhibitory activity. **FEBS letters**, v. 411, n. 1, p. 128–132, 7 jul. 1997.

LASKOWSKI, R. A. et al. PROCHECK: a program to check the stereochemical quality of protein structures. **Journal of applied crystallography**, v. 26, n. 2, p. 283–291, 1 abr. 1993.

LASKOWSKI, R. A. et al. PDBsum: Structural summaries of PDB entries. **Protein science: a publication of the Protein Society**, v. 27, n. 1, p. 129–134, jan. 2018.

LEGON, A. C.; MILLEN, D. J. Angular geometries and other properties of hydrogen-bonded dimers: a simple electrostatic interpretation of the success of the electron-pair model. **Chemical Society reviews**, v. 16, n. 0, p. 467–498, 1 jan. 1987.

LI, J. et al. New insights into the evolution of subtilisin-like serine protease genes in Pezizomycotina. **BMC evolutionary biology**, v. 10, n. 1, p. 68, jan. 2010.

LI, J. et al. Phylogenomic evolutionary surveys of subtilase superfamily genes in fungi. **Scientific reports**, v. 7, p. 45456, 30 mar. 2017.

LINDAHL et al. **GROMACS 2020.1 Manual**. [s.l: s.n.].

LINDAHL et al. **GROMACS 2020.1 Source code**. [s.l: s.n.].

LINDORFF-LARSEN, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. **Proteins**, v. 78, n. 8, p. 1950–1958, jun. 2010.

LIU, F. et al. Making two organelles from one: Woronin body biogenesis by peroxisomal protein sorting. **The Journal of cell biology**, v. 180, n. 2, p. 325–339, 28 jan. 2008.

LIU, S.-Q. et al. Characterizing structural features of cuticle-degrading proteases from fungi by molecular modeling. **BMC structural biology**, v. 7, p. 33, 18 maio 2007.

LIU, S.-Q. et al. The effect of calciums on molecular motions of proteinase K. **Journal of molecular modeling**, v. 17, n. 2, p. 289–300, fev. 2011.

LI, Y. et al. Functional analysis of the propeptide of subtilisin E as an intramolecular chaperone for protein folding. Refolding and inhibitory abilities of propeptide mutants. **The Journal of biological chemistry**, v. 270, n. 42, p. 25127–25132, 20 out. 1995.

LUTZONI, F. et al. Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. **American journal of botany**, v. 91, n. 10, p. 1446–1480, out. 2004.

NARANJO-ORTIZ, M. A.; GABALDÓN, T. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. **Biological reviews of the Cambridge Philosophical Society**, v. 94, n. 6, p. 2101–2137, 2019.

PEI, J.; KIM, B.-H.; GRISHIN, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. **Nucleic acids research**, v. 36, n. 7, p. 2295–2300, abr. 2008.

Periodic Table. Disponível em: <https://www.periodic-table.org/>. Acesso em: 17 set. 2021.

RAMACHANDRAN, G. N.; RAMAKRISHNAN, C.; SASISEKHARAN, V. Stereochemistry of polypeptide chain configurations. **Journal of molecular biology**, v. 7, p. 95–99, jul. 1963.

REUWSAAT, J. C. V. et al. A Predicted Mannoprotein Participates in *Cryptococcus gattii* Capsular Structure. **mSphere**, v. 3, n. 2, 25 abr. 2018.

ROSAS-GARCÍA, N. M. et al. Correlation between Pr1 and Pr2 gene content and virulence in *Metarhizium anisopliae* strains. **Journal of microbiology and biotechnology**, v. 24, n. 11, p. 1495–1502, 28 nov. 2014.

SALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. **Journal of molecular biology**, v. 234, n. 3, p. 779–815, 5 dez. 1993.

SÁNCHEZ-PÉREZ, L. D. C. et al. Enzymes of Entomopathogenic Fungi, Advances and Insights. **Advances in enzyme regulation**, v. 02, n. June, p. 65–76, 2014.

SANTI, L. et al. *Metarhizium anisopliae* host-pathogen interaction: differential immunoproteomics reveals proteins involved in the infection process of arthropods. **Fungal biology**, v. 114, n. 4, p. 312–319, abr. 2010.

SBARAINI, N. et al. Secondary metabolite gene clusters in the entomopathogen fungus *Metarhizium anisopliae*: genome identification and patterns of expression in a cuticle infection model. **BMC genomics**, v. 17, n. Suppl 8, p. 736, 25 out. 2016.

SBARAINI, N. et al. Genome-wide DNA methylation analysis of *Metarhizium anisopliae* during tick mimicked infection condition. **BMC genomics**, v. 20, n. 1, p. 836, 11 nov. 2019.

SCHMIDT, V. et al. Fungal dermatitis, glossitis and disseminated visceral mycosis caused by different *Metarhizium granulomatis* genotypes in veiled chameleons (Chamaeleo calyptratus) and first isolation in healthy lizards. **Veterinary microbiology**, v. 207, p. 74–82, ago. 2017.

SCHOCH, C. L. et al. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. **Systematic biology**, v. 58, n. 2, p. 224–239, abr. 2009.

SCHRANK, A.; VAINSTEIN, M. H. *Metarhizium anisopliae* enzymes and toxins. **Toxicon: official journal of the International Society on Toxinology**, v. 56, n. 7, p. 1267–1274, 15 dez. 2010.

SENN, H. M.; THIEL, W. QM/MM methods for biomolecular systems. **Angewandte Chemie** , v. 48, n. 7, p. 1198–1229, 2009.

SIEZEN, R. J.; LEUNISSEN, J. A. M. Subtilases: The superfamily of subtilisin-like serine proteases. **Protein science: a publication of the Protein Society**, v. 6, n. 3, p. 501–523,

31 dez. 1997.

SMALL, C.-L. N.; BIDOCHKA, M. J. Up-regulation of Pr1, a subtilisin-like protease, during conidiation in the insect pathogen *Metarhizium anisopliae*. **Mycological research**, v. 109, n. 3, p. 307–313, mar. 2005.

STAATS, C. C. et al. Comparative genome analysis of entomopathogenic fungi reveals a complex set of secreted proteins. **BMC genomics**, v. 15, n. 1, p. 822, 2014.

ST LEGER, R. et al. Construction of an improved mycoinsecticide overexpressing a toxic protease. **Proceedings of the National Academy of Sciences of the United States of America**, v. 93, n. 13, p. 6349–6354, 1996.

ST. LEGER, R. J. The role of cuticle-degrading proteases in fungal pathogenesis of insects. **Canadian journal of botany. Journal canadien de botanique**, v. 73, n. S1, p. 1119–1125, 31 dez. 1995.

ST LEGER, R. J. Studies on adaptations of *Metarhizium anisopliae* to life in the soil. **Journal of invertebrate pathology**, v. 98, n. 3, p. 271–276, jul. 2008.

ST LEGER, R. J.; BIDOCHKA, M. J.; ROBERTS, D. W. Isoforms of the cuticle-degrading Pr1 proteinase and production of a metalloproteinase by *Metarhizium anisopliae*. **Archives of biochemistry and biophysics**, v. 313, n. 1, p. 1–7, ago. 1994.

ST LEGER, R. J.; WANG, J. B. Metarhizium: jack of all trades, master of many. **Open biology**, v. 10, n. 12, p. 200307, dez. 2020.

SUH, S. O.; NODA, H.; BLACKWELL, M. Insect symbiosis: derivation of yeast-like endosymbionts within an entomopathogenic filamentous lineage. **Molecular biology and evolution**, v. 18, n. 6, p. 995–1000, jun. 2001.

VERLI, H. Dinâmica Molecular. In: VERLI, H. (Ed.). . **Bioinformática da Biologia à Flexibilidade Molecular**. [s.l.] Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014. p. 172–186.

VILCINSKAS, A. Coevolution between pathogen-derived proteinases and proteinase inhibitors of host insects. **Virulence**, v. 1, n. 3, p. 206–214, maio 2010.

WANG, C.; TYPAS, M. A.; BUTT, T. M. Detection and characterisation of pr1 virulent gene deficiencies in the insect pathogenic fungus *Metarhizium anisopliae*. **FEMS microbiology letters**, v. 213, n. 2, p. 251–255, ago. 2002.

WISECAVER, J. H.; ROKAS, A. Fungal metabolic gene clusters-caravans traveling across genomes and environments. **Frontiers in microbiology**, v. 6, p. 161, 3 mar. 2015.

WISECAVER, J. H.; SLOT, J. C.; ROKAS, A. The evolution of fungal metabolic pathways. **PLoS genetics**, v. 10, n. 12, p. e1004816, dez. 2014.

YANG, L.-Q. et al. Insight derived from molecular dynamics simulation into dynamics and molecular motions of cuticle-degrading serine protease Ver112. **Journal of biomolecular structure & dynamics**, v. 37, n. 8, p. 2004–2016, maio 2019.

ZIMMERMANN, G.; PAPIEROK, B.; GLARE, T. Elias Metschnikoff, Elie Metchnikoff or Ilya

Ilich Mechnikov (1845-1916): A Pioneer in Insect Pathology, the First Describer of the Entomopathogenic Fungus *Metarhizium anisopliae* and How to Translate a Russian Name. **Biocontrol science and technology**, v. 5, n. 4, p. 527–530, dez. 1995.

Anexos

Anexo 1: Script em linguagem *Python* utilizado para modelagem comparativa com MODELLER, utilizando como exemplo a modelagem para Pr1J1 de *Metarhizium anisopliae*.

```
# Homology modeling by the automodel class
import sys
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.']

a = automodel(env,
              alnfile  = '1ic6_jmani1.pir',     # alignment filename
              knowns   = '1ic6',                # templates
              sequence = 'Jmani1',              # code of the target
              assess_methods=(assess.DOPE))

a.starting_model= 1           # index of the first model
a.ending_model  = 10          # index of the last model
                               # (determines how many models to calculate)
# Thorough MD optimization:
a.md_level = refine.slow

# Repeat the whole cycle 2 times and do not stop unless obj.func. > 1E6
a.repeat_optimization = 2
a.max_molpdf = 1e6

a.make()                          # do the actual homology modeling

# Get a list of all successfully built models from a.outputs
ok_models = [x for x in a.outputs if x['failure'] is None]
```

```
# Rank the models by DOPE score
key = 'DOPE score'
if sys.version_info[:2] == (2,3):
   # Python 2.3's sort doesn't have a 'key' argument
   ok_models.sort(lambda a,b: cmp(a[key], b[key]))
else:
   ok_models.sort(key=lambda a: a[key])

#Rank all models:
print(">> Model Rankings:")
for m in range(len(ok_models)):
      print("  %i\t%s  (DOPE  score  %.3f)"  %  (m+1,ok_models[m]['name'],
ok_models[m][key]))
```

Anexo 2: Arquivo de parâmetros para simulação NVT no *software* GROMACS.

```
define                  = -DPOSRES  ; position restrain the protein
; Run parameters
integrator              = md        ; leap-frog integrator
nsteps                  = 50000     ; 2 * 50000 = 100 ps
dt                      = 0.002     ; 2 fs
; Output control
nstxout                 = 500       ; save coordinates every 1.0 ps
nstvout                 = 500       ; save velocities every 1.0 ps
nstenergy               = 500       ; save energies every 1.0 ps
nstlog                  = 500       ; update log file every 1.0 ps
; Bond parameters
continuation            = no        ; first dynamics run
constraint_algorithm    = lincs     ; holonomic constraints
constraints             = h-bonds   ; bonds involving H are constrained
lincs_iter              = 1         ; accuracy of LINCS
lincs_order             = 4         ; also related to accuracy
; Nonbonded settings
cutoff-scheme           = Verlet    ; Buffered neighbor searching
ns_type                 = grid      ; search neighboring grid cells
nstlist                 = 10        ; 20 fs, largely irrelevant with
Verlet
rcoulomb                = 1.0       ; short-range electrostatic cutoff (in
nm)
rvdw                    = 1.0       ; short-range van der Waals cutoff (in
nm)
DispCorr                = EnerPres  ; account for cut-off vdW scheme
; Electrostatics
coulombtype             = PME       ; Particle Mesh Ewald for long-range
electrostatics
pme_order               = 4         ; cubic interpolation
fourierspacing          = 0.16      ; grid spacing for FFT
; Temperature coupling is on
tcoupl                  = V-rescale ; modified Berendsen thermostat
tc-grps                 = Protein Non-Protein   ; two coupling groups -
more accurate
tau_t                   = 0.1     0.1 ; time constant, in ps
ref_t                   = 300     300 ; reference temperature, one for
each group, in K
```

```
; Pressure coupling is off
pcoupl                   = no        ; no pressure coupling in NVT
; Periodic boundary conditions
pbc                      = xyz       ; 3-D PBC
; Velocity generation
gen_vel                   = yes       ; assign velocities from Maxwell
distribution
gen_temp                 = 300        ; temperature for Maxwell distribution
gen_seed                 = -1        ; generate a random seed
```

Anexo 3: Arquivo de parâmetros para simulação NPT no *software* GROMACS.

```
define                  = -DPOSRES  ; position restrain the protein
; Run parameters
integrator              = md        ; leap-frog integrator
nsteps                  = 50000     ; 2 * 50000 = 100 ps
dt                      = 0.002     ; 2 fs
; Output control
nstxout                 = 500       ; save coordinates every 1.0 ps
nstvout                 = 500       ; save velocities every 1.0 ps
nstenergy               = 500       ; save energies every 1.0 ps
nstlog                  = 500       ; update log file every 1.0 ps
; Bond parameters
continuation            = yes       ; Restarting after NVT
constraint_algorithm    = lincs     ; holonomic constraints
constraints             = h-bonds   ; bonds involving H are constrained
lincs_iter              = 1         ; accuracy of LINCS
lincs_order             = 4         ; also related to accuracy
; Nonbonded settings
cutoff-scheme           = Verlet    ; Buffered neighbor searching
ns_type                 = grid      ; search neighboring grid cells
nstlist                 = 10        ; 20 fs, largely irrelevant with Verlet
scheme
rcoulomb                = 1.0       ; short-range electrostatic cutoff (in
nm)
rvdw                    = 1.0       ; short-range van der Waals cutoff (in
nm)
DispCorr                = EnerPres  ; account for cut-off vdW scheme
; Electrostatics
coulombtype             = PME       ; Particle Mesh Ewald for long-range
electrostatics
pme_order               = 4         ; cubic interpolation
fourierspacing          = 0.16      ; grid spacing for FFT
; Temperature coupling is on
tcoupl                  = V-rescale             ; modified Berendsen
thermostat
tc-grps                 = Protein Non-Protein   ; two coupling groups -
more accurate
tau_t                   = 0.1     0.1           ; time constant, in ps
```

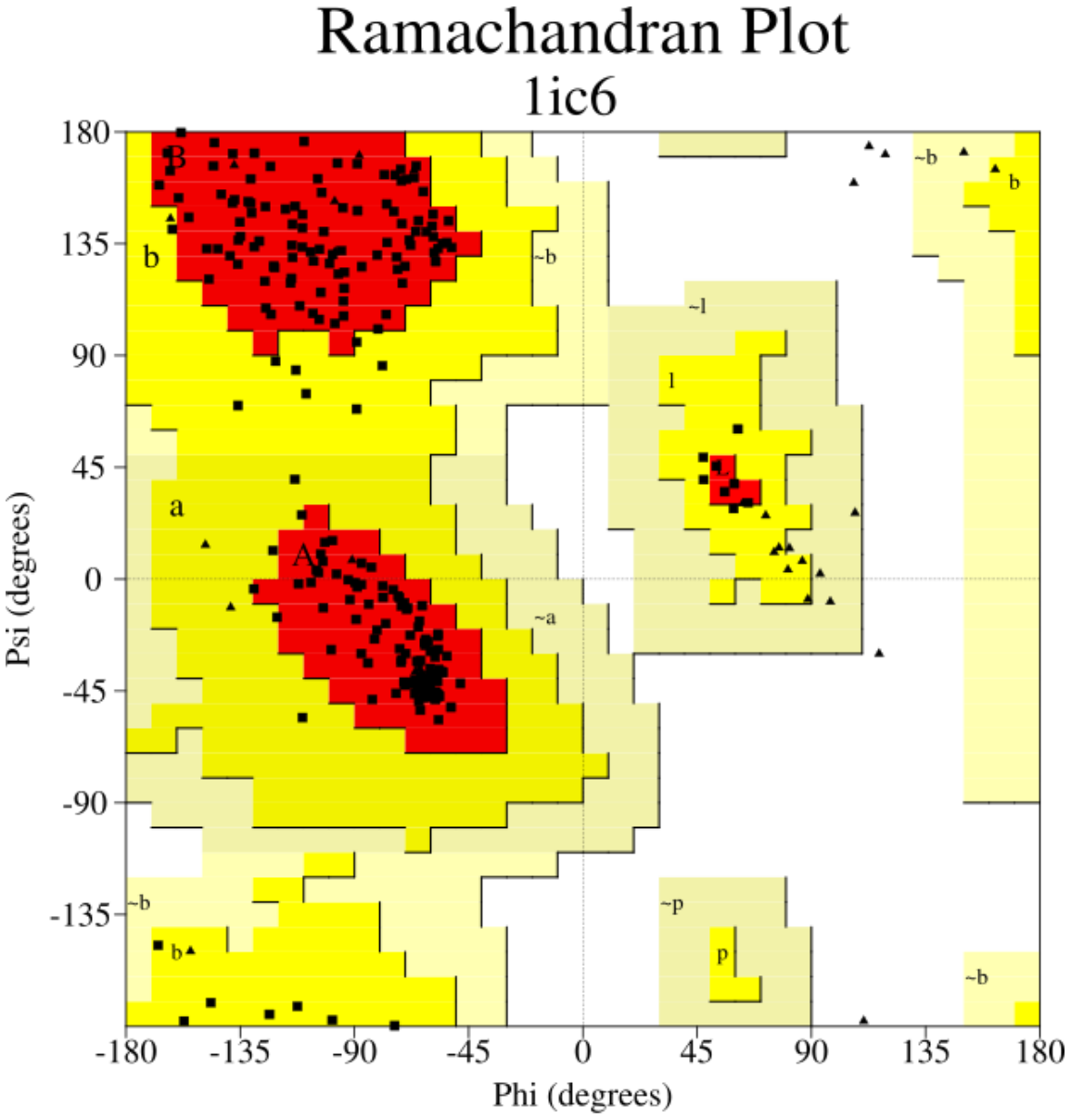
```
ref_t                    = 300     300           ; reference temperature,
one for each group, in K
; Pressure coupling is on
pcoupl                   = Parrinello-Rahman     ; Pressure coupling on in
NPT
pcoupltype               = isotropic             ; uniform scaling of box
vectors
tau_p                    = 2.0                   ; time constant, in ps
ref_p                    = 1.0                   ; reference pressure, in
bar
compressibility          = 4.5e-5                ; isothermal
compressibility of water, bar^-1
refcoord_scaling         = com
; Periodic boundary conditions
pbc                      = xyz       ; 3-D PBC
; Velocity generation
gen_vel                  = no        ; Velocity generation is off
```

Anexo 4: Arquivo de parâmetros para fase de produção por Dinâmica Molecular no *software* GROMACS.

```
; Run parameters
integrator              = md        ; leap-frog integrator
nsteps                  = 100000000    ; 2 * 100000000 = 200000000 fs (200
ns)
dt                      = 0.002     ; 2 fs
; Output control
nstxout                 = 0         ; suppress bulky .trr file by
specifying
nstvout                 = 0         ; 0 for output frequency of nstxout,
nstfout                 = 0         ; nstvout, and nstfout
nstenergy               = 5000      ; save energies every 10.0 ps
nstlog                  = 5000      ; update log file every 10.0 ps
nstxout-compressed      = 5000      ; save compressed coordinates every
10.0 ps
compressed-x-grps       = System    ; save the whole system
; Bond parameters
continuation            = yes       ; Restarting after NPT
constraint_algorithm    = lincs     ; holonomic constraints
constraints             = h-bonds   ; bonds involving H are constrained
lincs_iter              = 1         ; accuracy of LINCS
lincs_order             = 4         ; also related to accuracy
; Neighborsearching
cutoff-scheme           = Verlet    ; Buffered neighbor searching
ns_type                 = grid      ; search neighboring grid cells
nstlist                 = 10        ; 20 fs, largely irrelevant with Verlet
scheme
rcoulomb                = 1.0       ; short-range electrostatic cutoff (in
nm)
rvdw                    = 1.0       ; short-range van der Waals cutoff (in
nm)
; Electrostatics
coulombtype             = PME       ; Particle Mesh Ewald for long-range
electrostatics
pme_order               = 4         ; cubic interpolation
fourierspacing          = 0.16      ; grid spacing for FFT
; Temperature coupling is on
```

```
tcoupl                  = V-rescale             ; modified Berendsen
thermostat
tc-grps                 = Protein Non-Protein   ; two coupling groups -
more accurate
tau_t                   = 0.1     0.1           ; time constant, in ps
ref_t                   = 300     300           ; reference temperature,
one for each group, in K
; Pressure coupling is on
pcoupl                  = Parrinello-Rahman     ; Pressure coupling on in
NPT
pcoupltype              = isotropic             ; uniform scaling of box
vectors
tau_p                   = 2.0                   ; time constant, in ps
ref_p                   = 1.0                   ; reference pressure, in
bar
compressibility         = 4.5e-5                ; isothermal
compressibility of water, bar^-1
; Periodic boundary conditions
pbc                     = xyz       ; 3-D PBC
; Dispersion correction
DispCorr                = EnerPres  ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no        ; Velocity generation is off
```

Anexo 5: Gráfico de Ramachandran para o PDB 1IC6 (molde), conforme análise do PDBSum/PROCHECK.

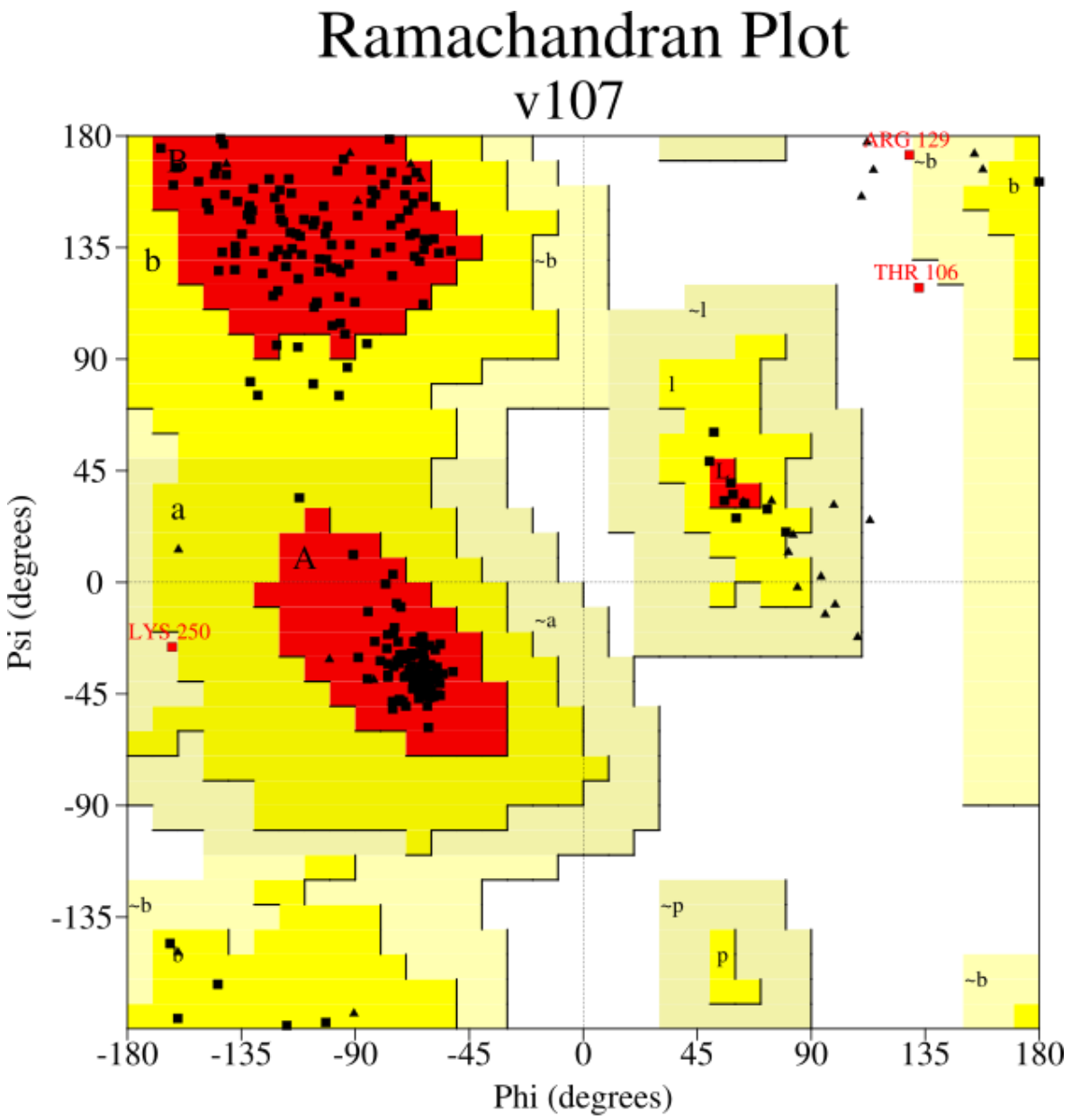


Plot statistics

Residues in most favoured regions [A,B,L]	211	89.8%
Residues in additional allowed regions [a,b,l,p]	24	10.2%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	235	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	33	
Number of proline residues	9	
Total number of residues	279	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Anexo 6: Gráfico de Ramachandran para modelo de Pr1J1, conforme análise do PDBSum/PROCHECK.

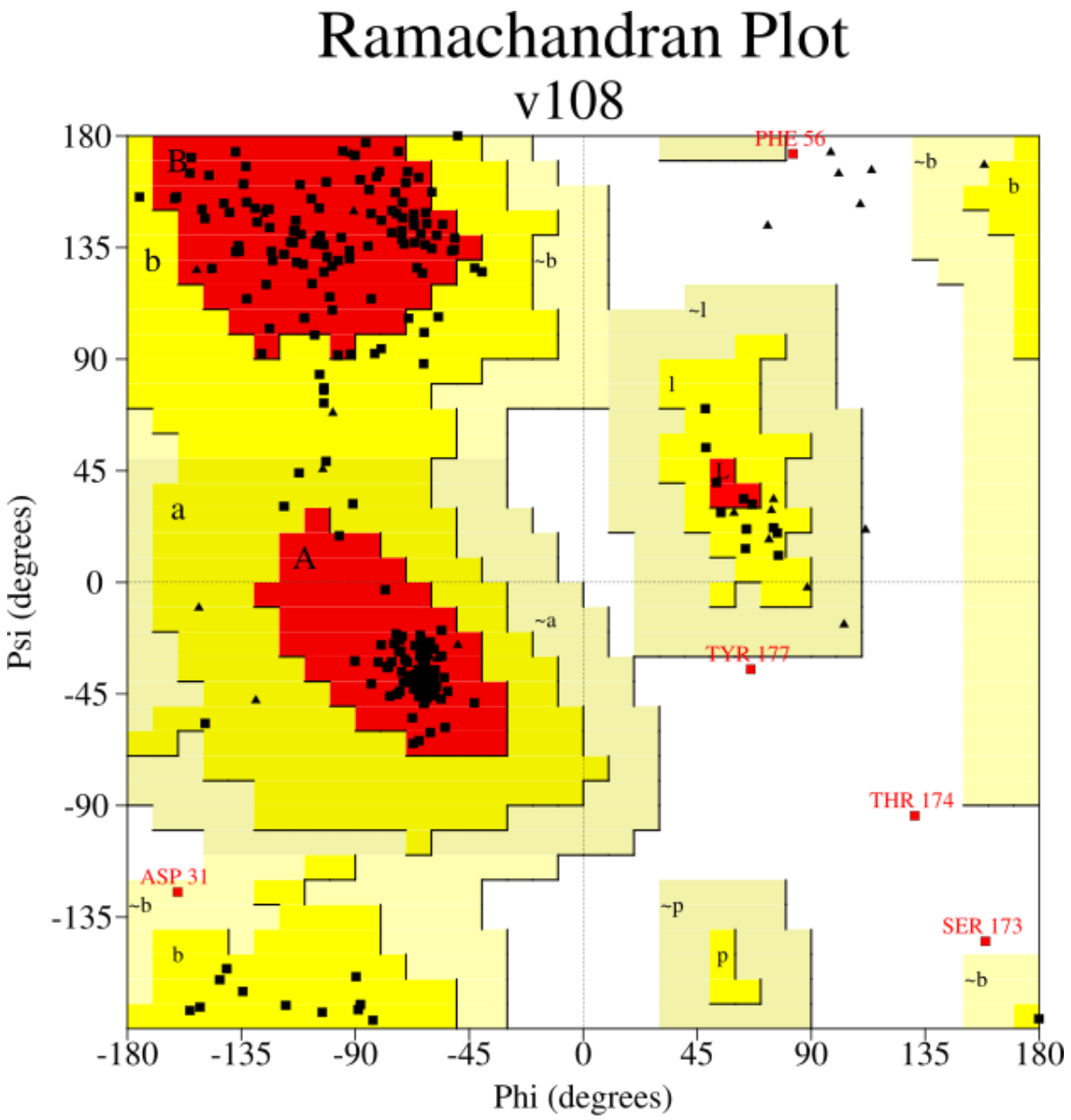


Plot statistics

Residues in most favoured regions [A,B,L]	218	90.5%
Residues in additional allowed regions [a,b,l,p]	20	8.3%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.4%
Residues in disallowed regions	2	0.8%
Number of non-glycine and non-proline residues	241	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	31	
Number of proline residues	8	
Total number of residues	282	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Anexo 7: Gráfico de Ramachandran para modelo de Pr1J2, conforme análise do PDBSum/PROCHECK.



Plot statistics

Residues in most favoured regions [A,B,L]	196	82.4%
Residues in additional allowed regions [a,b,l,p]	37	15.5%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.4%
Residues in disallowed regions	4	1.7%
Number of non-glycine and non-proline residues	238	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	30	
Number of proline residues	18	
Total number of residues	288	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Anexo 8: *Shell script* para simulação por Dinâmica Molecular usando GROMACS. Apresenta-se somente o utilizado para 1IC6, sendo que os demais seguem o mesmo procedimento, apenas alterando caminhos e variáveis.

```
GMXBIN='gmx'
OUTDIR='/home/andreisfc/Documents/5-molecular_dynamics/pr1j_joint'

#passos iniciais para 1ic6
ROOTNAME='1ic6'
cd 1ic6

#preparacao
$GMXBIN pdb2gmx -f $ROOTNAME.ca.pdb -o $ROOTNAME.ca.gro -p $ROOTNAME.ca.top
-ignh -water tip3p -ff amber99sb-ildn

#definicao da caixa
$GMXBIN editconf -f $ROOTNAME.ca.gro -o $ROOTNAME.ca.box.gro -c -d 1.0 -bt
dodecahedron

#solvatacao
$GMXBIN solvate -cp $ROOTNAME.ca.box.gro -p $ROOTNAME.ca.top -cs -o
$ROOTNAME.ca.box.solv.gro

#neutralizacao do sistema
$GMXBIN grompp -f ions.mdp -c $ROOTNAME.ca.box.solv.gro -p $ROOTNAME.ca.top
-o $ROOTNAME.ca.ions.tpr

$GMXBIN genion -s $ROOTNAME.ca.ions.tpr -p $ROOTNAME.ca.top -o
$ROOTNAME.ca.ions.gro -pname NA -nname CL -neutral

#minimizacao
$GMXBIN grompp -f minim.mdp -c $ROOTNAME.ca.ions.gro -p $ROOTNAME.ca.top -o
$ROOTNAME.ca.em.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.em -v

echo '10' | $GMXBIN energy -f $ROOTNAME.ca.em.edr -o
$ROOTNAME.ca.em.potential.xvg
```

```
#equilibracao h2o
$GMXBIN grompp -f nvt.mdp -c $ROOTNAME.ca.em.gro -r $ROOTNAME.ca.em.gro -p
$ROOTNAME.ca.top -o $ROOTNAME.ca.nvt.tpr

# Fase de réplicas
REPNUM=1
## 1ic6 [rep1]
ROOTNAME='1ic6'
cd 1ic6

mkdir rep$REPNUM
cp npt.mdp md.mdp *.top *.itp rep$REPNUM/
cp $ROOTNAME.ca.nvt.tpr rep$REPNUM/$ROOTNAME.ca.nvt-rep$REPNUM.tpr
cd rep$REPNUM

#NVT
$GMXBIN mdrun -deffnm $ROOTNAME.ca.nvt-rep$REPNUM -v

$GMXBIN grompp -f npt.mdp -c $ROOTNAME.ca.nvt-rep$REPNUM.gro -r
$ROOTNAME.ca.nvt-rep$REPNUM.gro -t $ROOTNAME.ca.nvt-rep$REPNUM.cpt -p
$ROOTNAME.ca.top -o $ROOTNAME.ca.npt-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.npt-rep$REPNUM -v

echo 18 | $GMXBIN energy -f $ROOTNAME.ca.npt-rep$REPNUM.edr -o
$ROOTNAME.ca.npt-rep$REPNUM.pressure.xvg

#fase de producao
$GMXBIN grompp -f md.mdp -c $ROOTNAME.ca.npt-rep$REPNUM.gro -t
$ROOTNAME.ca.npt-rep$REPNUM.cpt -p $ROOTNAME.ca.top -o
$ROOTNAME.ca.md-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.md-rep$REPNUM -v

##
REPNUM=2
# 1ic6 [rep2]
ROOTNAME='1ic6'
cd ../../1ic6
```

```
mkdir rep$REPNUM
cp npt.mdp md.mdp *.top *.itp rep$REPNUM/
cp $ROOTNAME.ca.nvt.tpr rep$REPNUM/$ROOTNAME.ca.nvt-rep$REPNUM.tpr
cd rep$REPNUM

#NVT
$GMXBIN mdrun -deffnm $ROOTNAME.ca.nvt-rep$REPNUM -v

$GMXBIN grompp -f npt.mdp -c $ROOTNAME.ca.nvt-rep$REPNUM.gro -r
$ROOTNAME.ca.nvt-rep$REPNUM.gro -t $ROOTNAME.ca.nvt-rep$REPNUM.cpt -p
$ROOTNAME.ca.top -o $ROOTNAME.ca.npt-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.npt-rep$REPNUM -v

echo 18 | $GMXBIN energy -f $ROOTNAME.ca.npt-rep$REPNUM.edr -o
$ROOTNAME.ca.npt-rep$REPNUM.pressure.xvg

#fase de producao
$GMXBIN grompp -f md.mdp -c $ROOTNAME.ca.npt-rep$REPNUM.gro -t
$ROOTNAME.ca.npt-rep$REPNUM.cpt -p $ROOTNAME.ca.top -o
$ROOTNAME.ca.md-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.md-rep$REPNUM -v

##
REPNUM=3

## 1ic6 [rep3]
ROOTNAME='1ic6'
cd ../../1ic6

mkdir rep$REPNUM
cp npt.mdp md.mdp *.top *.itp rep$REPNUM/
cp $ROOTNAME.ca.nvt.tpr rep$REPNUM/$ROOTNAME.ca.nvt-rep$REPNUM.tpr
cd rep$REPNUM

#NVT
$GMXBIN mdrun -deffnm $ROOTNAME.ca.nvt-rep$REPNUM -v
```

```
$GMXBIN grompp -f npt.mdp -c $ROOTNAME.ca.nvt-rep$REPNUM.gro -r
$ROOTNAME.ca.nvt-rep$REPNUM.gro -t $ROOTNAME.ca.nvt-rep$REPNUM.cpt -p
$ROOTNAME.ca.top -o $ROOTNAME.ca.npt-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.npt-rep$REPNUM -v

echo 18 | $GMXBIN energy -f $ROOTNAME.ca.npt-rep$REPNUM.edr -o
$ROOTNAME.ca.npt-rep$REPNUM.pressure.xvg

#fase de producao
$GMXBIN grompp -f md.mdp -c $ROOTNAME.ca.npt-rep$REPNUM.gro -t
$ROOTNAME.ca.npt-rep$REPNUM.cpt -p $ROOTNAME.ca.top -o
$ROOTNAME.ca.md-rep$REPNUM.tpr

$GMXBIN mdrun -deffnm $ROOTNAME.ca.md-rep$REPNUM -v
```

Anexo 9: *Shell script* para centralização das trajetórias calculadas utilizando o *script* do Anexo 8. Apresenta-se somente o utilizado para 1IC6, sendo que os demais seguem o mesmo procedimento, apenas alterando caminhos e variáveis.

```
# 1ic6
echo 'q' | gmx_mpi make_ndx -f 1ic6/1ic6.ca.em.gro -o 1ic6/1ic6.ndx

echo 1 0 | gmx_mpi trjconv -s 1ic6/1ic6.ca.em.tpr -f 1ic6/1ic6.ca.em.trr -o
1ic6/1ic6.ca.em.center.gro -pbc mol -center

echo 1 0 | gmx_mpi trjconv -s 1ic6/1ic6.ca.em.tpr -f
1ic6/rep1/1ic6.ca.md-rep1.xtc -o
1ic6/rep1/1ic6.ca.md-rep1.center-system.xtc -pbc mol -center

echo 1 0 | gmx_mpi trjconv -s 1ic6/1ic6.ca.em.tpr -f
1ic6/rep2/1ic6.ca.md-rep2.xtc -o
1ic6/rep2/1ic6.ca.md-rep2.center-system.xtc -pbc mol -center

echo 1 0 | gmx_mpi trjconv -s 1ic6/1ic6.ca.em.tpr -f
1ic6/rep3/1ic6.ca.md-rep3.xtc -o
1ic6/rep3/1ic6.ca.md-rep3.center-system.xtc -pbc mol -center
```

Anexo 10: *Shell script* para cálculo de RMSD e RMSF das trajetórias obtidas executando o *script* do Anexo 9. Apresenta-se somente o utilizado para 1IC6, sendo que os demais seguem o mesmo procedimento, apenas alterando caminhos e variáveis.

```
#1ic6
root="1ic6"

rep=1
echo 4 4 | gmx_mpi rms -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -n $root/$root.ndx -o
$root/rep$rep/$root.ca.md-rep$rep.center-system.rmsd.xvg

echo 1 | gmx_mpi rmsf -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -res -n $root/$root.ndx
-o $root/rep$rep/$root.ca.md-rep$rep.center-system.rmsf.xvg

rep=2
echo 4 4 | gmx_mpi rms -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -n $root/$root.ndx -o
$root/rep$rep/$root.ca.md-rep$rep.center-system.rmsd.xvg

echo 1 | gmx_mpi rmsf -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -res -n $root/$root.ndx
-o $root/rep$rep/$root.ca.md-rep$rep.center-system.rmsf.xvg

rep=3
echo 4 4 | gmx_mpi rms -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -n $root/$root.ndx -o
$root/rep$rep/$root.ca.md-rep$rep.center-system.rmsd.xvg

echo 1 | gmx_mpi rmsf -s $root/$root.ca.em.tpr -f
$root/rep$rep/$root.ca.md-rep$rep.center-system.xtc -res -n $root/$root.ndx
-o $root/rep$rep/$root.ca.md-rep$rep.center-system.rmsf.xvg
```

Anexo 11: *Script* em linguagem *Julia* utilizado para cálculo de média e desvio-padrão amostral de RMSF e RMSF das simulações por Dinâmica Molecular em triplicata, com base nos cálculos gerados utilizando GROMACS.

```
### Calculates average values for GROMACS RMSD/RMSF files in standard
XMGrace
### format
## Input: n .xvg files
## Output: tab-separated values (STDOUT)
using ArgParse
using Statistics

## CLI argument processing
s = ArgParseSettings()
s.description = "Calculates average values for GROMACS RMSD/RMSF files in
standard XMGrace format"
s.add_help = true
s.add_version = true
s.version = "0.1"
s.commands_are_required = true

@add_arg_table! s begin
   "--input", "-i"
       help = ".xvg file names"
       required = true
       nargs = '+'
   "--output" , "-o"
       help = "output file name"
       required = true
       nargs = 1
end

parsed_args = parse_args(s)

infiles = parsed_args["input"]
outfile = parsed_args["output"]

let
```

```
numcols = length(infiles)+1
counter = 2
fst = true
global data = []
for file in infiles
   tmp1 = readlines(file) # read contents of file to array
   filter!(f->f[1]∉['#','&','@'], tmp1) # remove comments
   if fst
       global numrows = length(tmp1)
       data = Array{Float64}(undef, numrows, numcols)
       fst = false
   end

   for i in 1:numrows
       tmp = split(tmp1[i])
       data[i,1] = parse(Float64, tmp[1])
       data[i,counter] = parse(Float64, tmp[2])
   end
   counter = counter + 1
end
end

fh = open(outfile[1], "w")
for i in 1:numrows
   print(fh,
"$(data[i,1])\t$(mean(data[i,2:end]))\t$(std(data[i,2:end]))\n")
end
close(fh)
```

Anexo 12: *Script* em linguagem *Julia* utilizado para cálculo da média e desvio-padrão dos valores obtidos com o *script* do Anexo 8.

```
### Window-averages GROMACS RMSD/RMSF files created by avg-rms.jl
### format
## Input: .xvg file
## Output: tab-separated values (STDOUT)
using ArgParse
using Statistics

## CLI argument processing
s = ArgParseSettings()
s.description = "Window-averages GROMACS RMSD/RMSF files created by
avg-rms.jl"
s.add_help = true
s.add_version = true
s.version = "0.1"
s.commands_are_required = true

@add_arg_table! s begin
   "--input", "-i"
       help = ".tsv"
       required = true
       nargs = 1
   "--output" , "-o"
       help = "output file name"
       required = true
       nargs = 1
   "--winsize" , "-w"
       help = "window size"
       arg_type = Int
       required = true
       nargs = 1
end

parsed_args = parse_args(s)

infile = parsed_args["input"][1]
outfile = parsed_args["output"]
```

```
winsize = parsed_args["winsize"][1]

let
global data = []
tmp1 = readlines(infile) # read contents of file to array
filter!(f->f[1]∉['#','&','@'], tmp1) # remove comments
global numrows = length(tmp1)
println("Size of input = $(numrows)")
data = Array{Float64}(undef, numrows, 3)

for i in 1:(numrows-winsize)
   tmp = split(tmp1[i])
   data[i,1] = parse(Float64, tmp[1])
   avgs = []
   stds = []
   for j in i:i+winsize
       push!(avgs, parse(Float64, split(tmp1[j])[2]))
       push!(stds, parse(Float64, split(tmp1[j])[3]))
   end
   data[i,2] = mean(avgs)
   data[i,3] = mean(stds)
end

let fh = open(outfile[1], "w")
   for i in 1:(numrows-winsize)
       print(fh, "$(data[i,1])\t$(data[i,2])\t$(data[i,3])\n")
   end
end

end #let
```

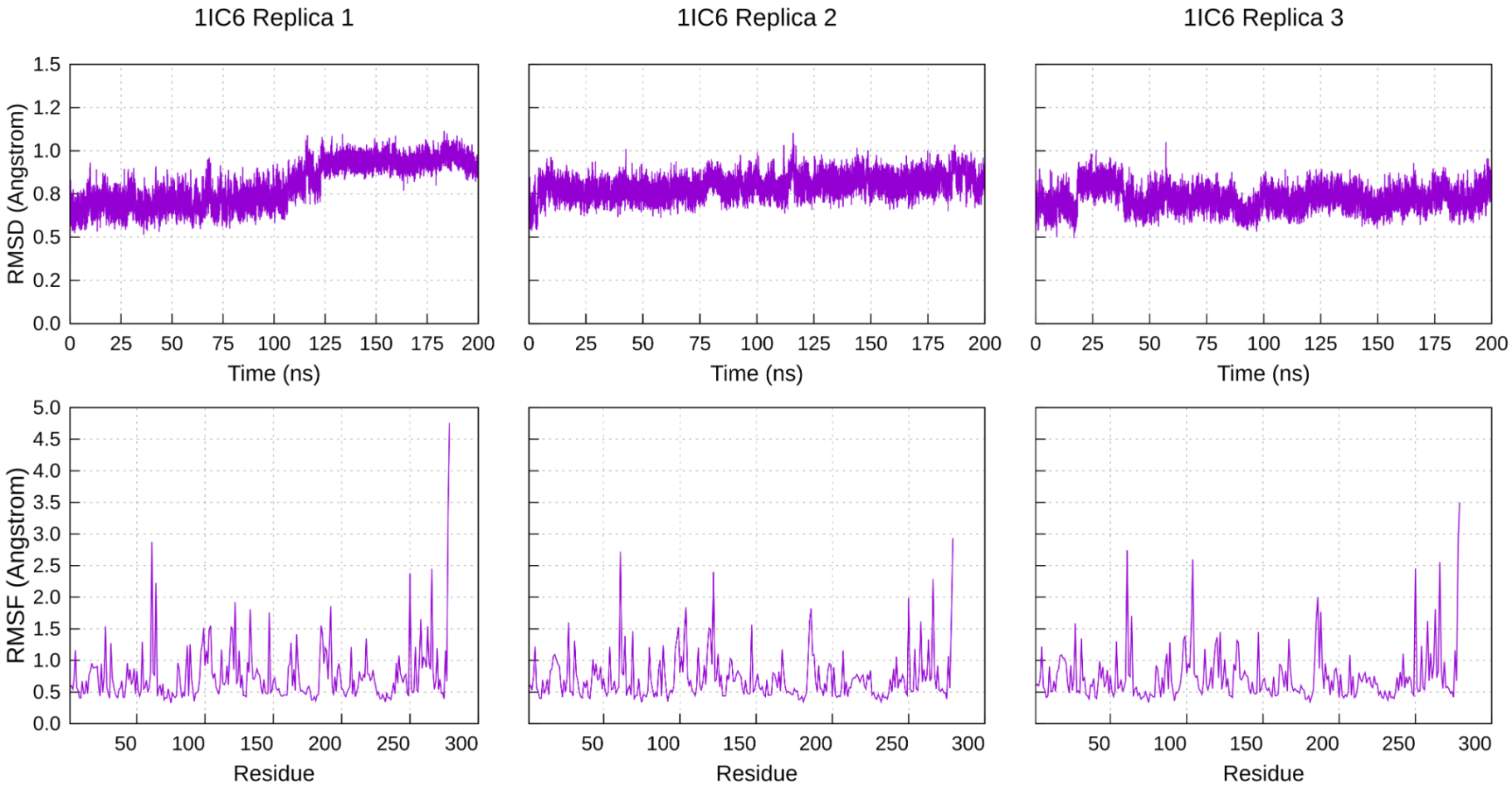
Anexo 13: *Script* para construção dos gráficos de RMSD médio utilizando Gnuplot.

```
set datafile commentschars "#@&"
set terminal svg enhanced size 1366,720
set output '1ic6_J1_J2-rmsdw500.svg'
set style fill noborder
set grid

### RMSD
set xtics 0,25,200 font ",18" nomirror
set xlabel "Time (ns)" font ",20"
set format y "%.1f"
set ytics nomirror
set ylabel "RMSD (Angstrom)" font ",20"
set title "Root Mean Square Deviation" font ",22"

plot "1ic6.ca.md.center-system.rmsd.statistics.win500.xvg" u
($1*0.001):(($2*10)+($3*10)):(($2*10)-($3*10)) w filledcurves notitle,\
'' u ($1*0.001):($2*10) title '1ic6' w lines,\
"Jmani1.ca.md.center-system.rmsd.statistics.win500.xvg" u
($1*0.001):(($2*10)+($3*10)):(($2*10)-($3*10)) w filledcurves notitle,\
'' u ($1*0.001):($2*10) title 'Pr1J1' w lines,\
"Jmani2.ca.md.center-system.rmsd.statistics.win500.xvg" u
($1*0.001):(($2*10)+($3*10)):(($2*10)-($3*10)) w filledcurves notitle,\
'' u ($1*0.001):($2*10) title 'Pr1J2' w lines
```

Anexo 14: *Script* para construção dos gráficos de desvio-padrão de RMSF utilizando Gnuplot.

```
set datafile commentschars "#@&"
set terminal svg enhanced size 1366,720
set output '1ic6_J1_J2-rmsf.svg'
set style fill noborder
set grid

### RMSD
set xtics 0,25,300 font ",18" nomirror
set xlabel "Residue #" font ",20"
set format y "%.1f"
set ytics nomirror
set ylabel "RMSF Sample Standard Deviation (Å)" font ",20"
set title "Root Mean Square Fluctuation by residue" font ",22"

plot "1ic6.ca.md.center-system.rmsf.statistics.xvg" u
($1):(($3*10)):(-($3*10)) w filledcurves title '1IC6',\
"Jmani1.ca.md.center-system.rmsf.statistics.xvg" u
($1):(($3*10)):(-($3*10)) w filledcurves title 'Pr1J1',\
"Jmani2.ca.md.center-system.rmsf.statistics.xvg" u
($1):(($3*10)):(-($3*10)) w filledcurves title 'Pr1J2'
```

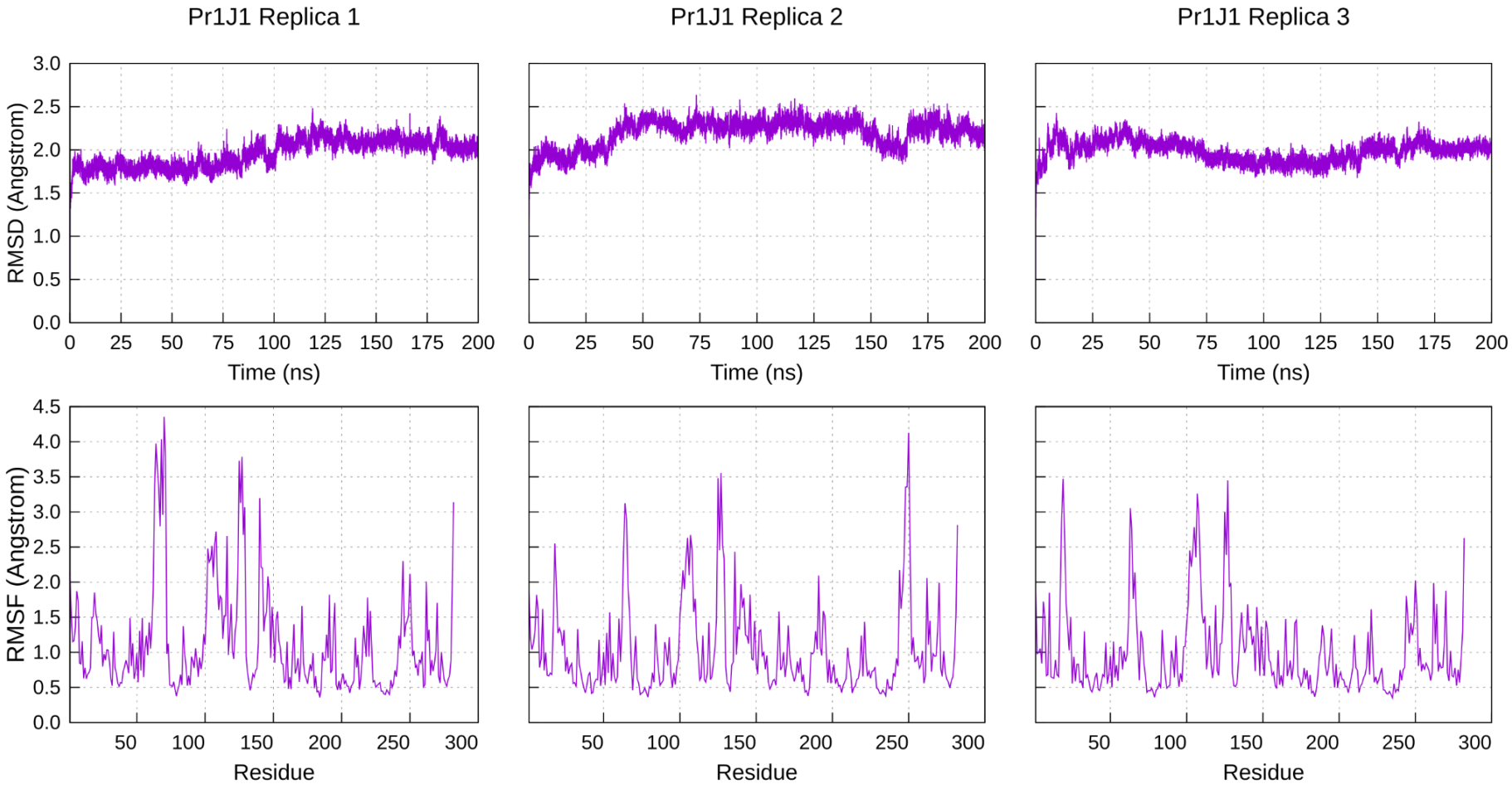
Anexo 15: Gráficos de RMSD e RMSF da triplicada de simulação para 1IC6, bem como script para construção utilizando Gnuplot.



```
set datafile commentschars "#@&"
set terminal svg enhanced size 1366,720
set output 'graphs/1ic6_rmsd-f.svg'
set multiplot layout 2,3 rowsfirst
set grid

### RMSD
set yrange [0.0:1.5]
set xtics 0,25,200 font ",18" nomirror
set xlabel "Time (ns)" font ",20"
set format y "%.1f"
set ytics 0,0.25,1.5 font ",18" nomirror
set ylabel "RMSD (Angstrom)" font ",20"
set title "1IC6 Replica 1" font ",22"
plot "1ic6/rep1/1ic6.ca.md-rep1.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "1IC6 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "1ic6/rep2/1ic6.ca.md-rep2.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "1IC6 Replica 3" font ",22"
plot "1ic6/rep3/1ic6.ca.md-rep3.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle

### RMSF
unset title
set xtics 0,50 right font ",18" nomirror
```

```
set xlabel "Residue" font ",20"
set format y "%.1f"
set yrange [0.0:5.0]
set ytics 0,0.5 font ",18" nomirror
set ylabel "RMSF (Angstrom)" font ",22"
#set title "1IC6 Replica 1" font ",22"
plot "1ic6/rep1/1ic6.ca.md-rep1.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "1IC6 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "1ic6/rep2/1ic6.ca.md-rep2.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "1IC6 Replica 3" font ",22"
plot "1ic6/rep3/1ic6.ca.md-rep3.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
```

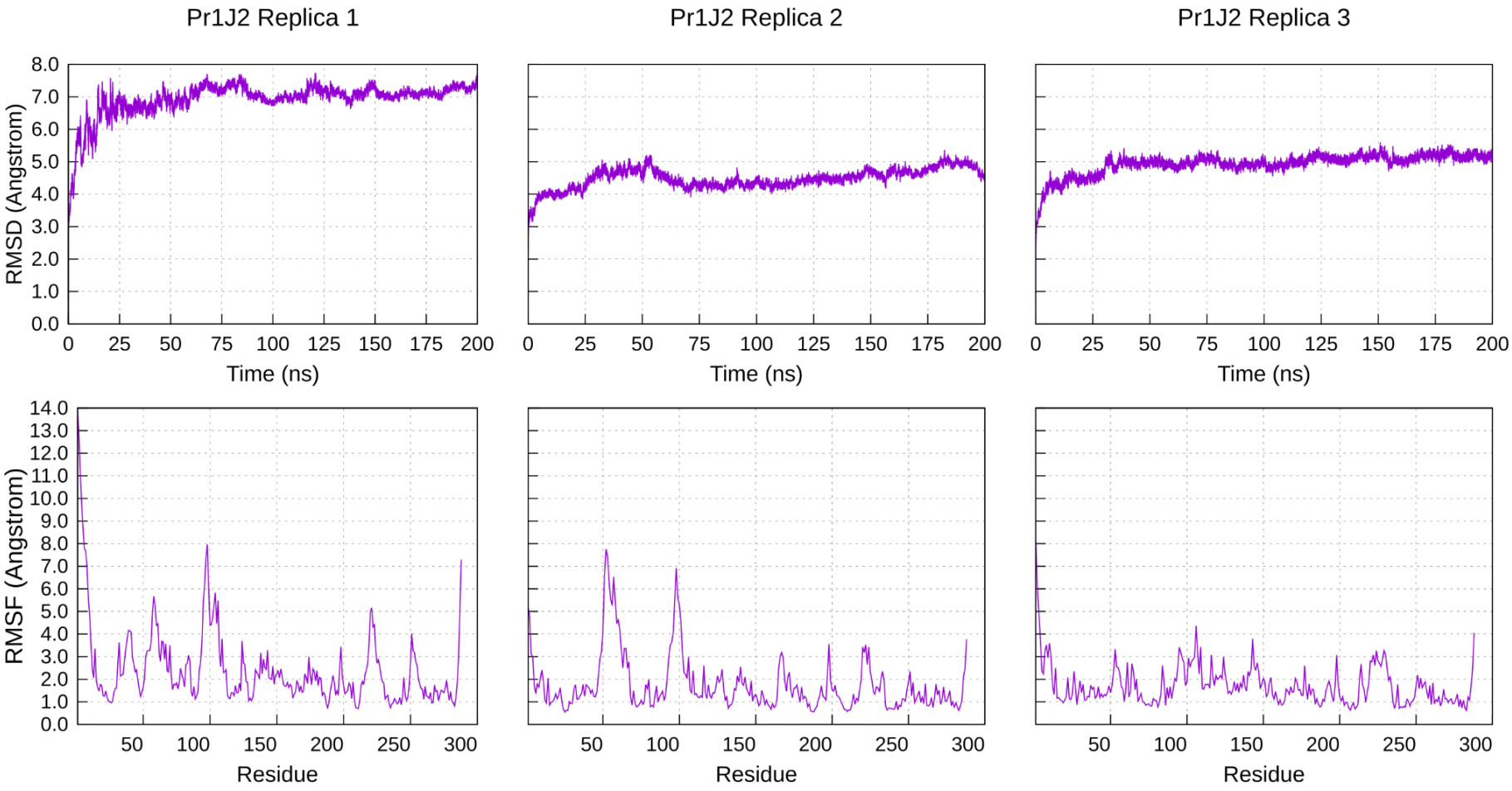
Anexo 16: Gráficos de RMSD e RMSF da triplicada de simulação para Pr1J1, bem como script para construção utilizando Gnuplot.



```
set datafile commentschars "#@&"
set terminal svg enhanced size 1366,720
set output 'graphs/jmani1_rmsd-f.svg'
set multiplot layout 2,3 rowsfirst
set grid

### RMSD
set yrange [0.0:3.0]
set xtics 0,25,200 font ",18" nomirror
set xlabel "Time (ns)" font ",20"
set format y "%.1f"
set ytics 0,0.5,3.0 font ",18" nomirror
set ylabel "RMSD (Angstrom)" font ",20"
set title "Pr1J1 Replica 1" font ",22"
plot "jmani1/rep1/Jmani1.ca.md-rep1.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "Pr1J1 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "jmani1/rep2/Jmani1.ca.md-rep2.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "Pr1J1 Replica 3" font ",22"
plot "jmani1/rep3/Jmani1.ca.md-rep3.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle

### RMSF
unset title
```

```
set xtics 0,50 right font ",18" nomirror
set xlabel "Residue" font ",20"
set format y "%.1f"
set yrange [0.0:4.5]
set ytics 0,0.5 font ",18" nomirror
set ylabel "RMSF (Angstrom)" font ",22"
#set title "Pr1J1 Replica 1" font ",22"
plot "jmani1/rep1/Jmani1.ca.md-rep1.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "Pr1J1 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "jmani1/rep2/Jmani1.ca.md-rep2.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "Pr1J1 Replica 3" font ",22"
plot "jmani1/rep3/Jmani1.ca.md-rep3.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
```

Anexo 17: Gráficos de RMSD e RMSF da triplicada de simulação para Pr1J2, bem como script para construção utilizando Gnuplot.



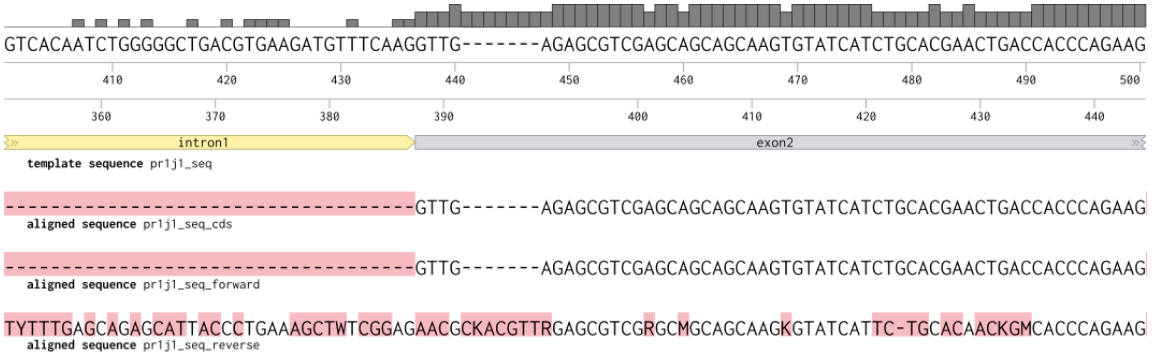
```
set datafile commentschars "#@&"
set terminal svg enhanced size 1366,720
set output 'graphs/jmani2_rmsd-f.svg'
set multiplot layout 2,3 rowsfirst
set grid

### RMSD
set yrange [0.0:8.0]
set xtics 0,25,200 font ",18" nomirror
set xlabel "Time (ns)" font ",20"
set format y "%.1f"
set ytics 0,1,8.0 font ",18" nomirror
set ylabel "RMSD (Angstrom)" font ",20"
set title "Pr1J2 Replica 1" font ",22"
plot "jmani2/rep1/Jmani2.ca.md-rep1.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "Pr1J2 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "jmani2/rep2/Jmani2.ca.md-rep2.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle
set title "Pr1J2 Replica 3" font ",22"
plot "jmani2/rep3/Jmani2.ca.md-rep3.center-system.rmsd.xvg" u ($1*0.001):($2*10) w lines notitle

### RMSF
unset title
set xtics 0,50 right font ",18" nomirror
```

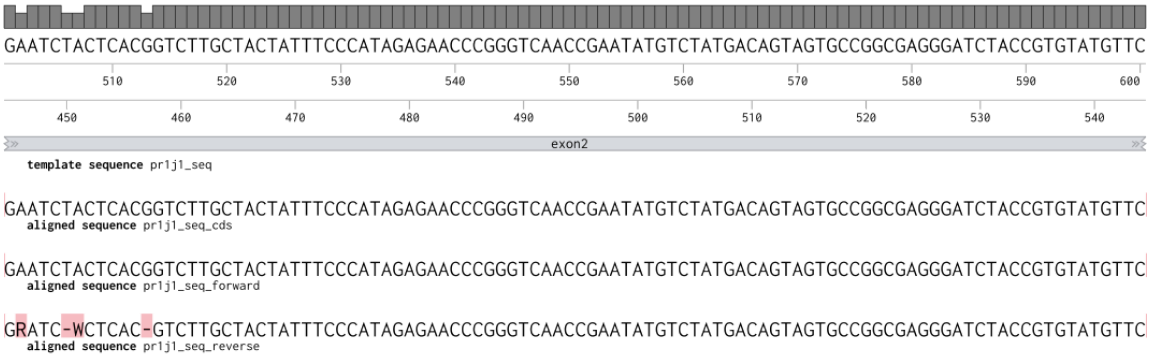
```
set xlabel "Residue" font ",20"
set format y "%.1f"
set yrange [0.0:14]
set ytics 0,1 font ",18" nomirror
set ylabel "RMSF (Angstrom)" font ",22"
#set title "Pr1J2 Replica 1" font ",22"
plot "jmani2/rep1/Jmani2.ca.md-rep1.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "Pr1J2 Replica 2" font ",22"
set format y '' #Remove tic labels
unset ylabel
plot "jmani2/rep2/Jmani2.ca.md-rep2.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
#set title "Pr1J2 Replica 3" font ",22"
plot "jmani2/rep3/Jmani2.ca.md-rep3.center-system.rmsf.xvg" u ($1):($2*10) w lines notitle
```

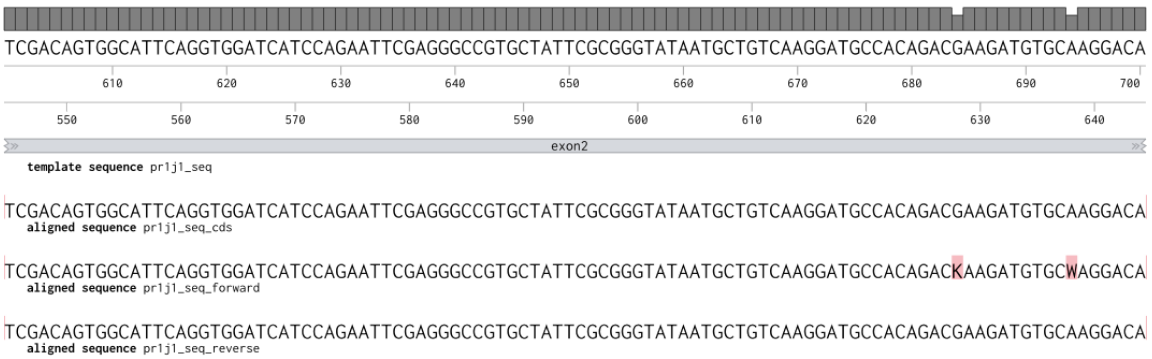
Anexo 18: Relação de oligonucleotídeos utilizados.

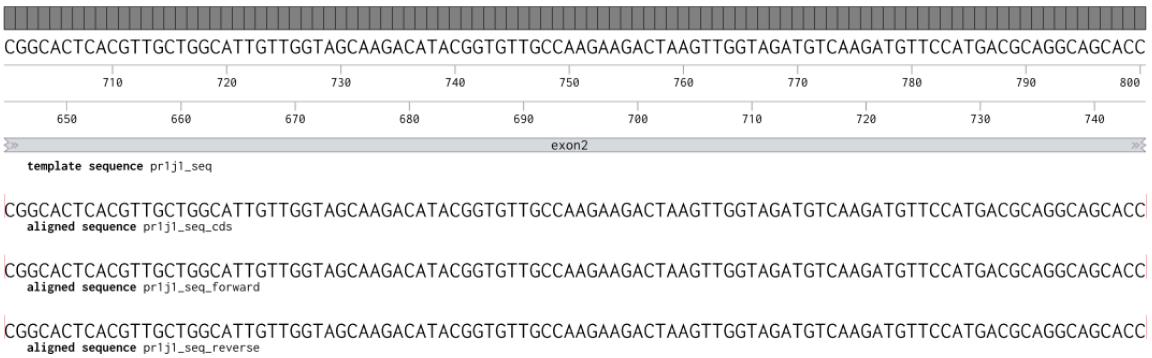
Par	Nome	Sequência (5'→3')	Tm (°C)
1	Tubulina_Direto	GTA ACC AAA TTG GTG CTG CT	60
	Tubulina_Reverso	CGA CGG AGA AAG TGG CCA TC	60
2	CDS	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT NN	60
3	Pr1J1_Direto	TTT AAG AAG GAG ATA TAC CAT GTT TTC GTT CAA AAC TC	60
	Pr1J1_Reverso	GTG GTG GTG GTG GTG GTG tAT GAT GCC GTT GTA TGC	60
4	Pr1J2_Direto	TTT AAG AAG GAG ATA TAC CAT GGC TAT TCT CAA GGC CTT TAC	60
	Pr1J2_Reverso	GTG GTG GTG GTG GTG GTG CTG AAC CCC ATT AAA CGC AAG C	60

Anexo 19: Sequenciamento do inserto do plasmídeo pET23d(+)-*pr1J1.*

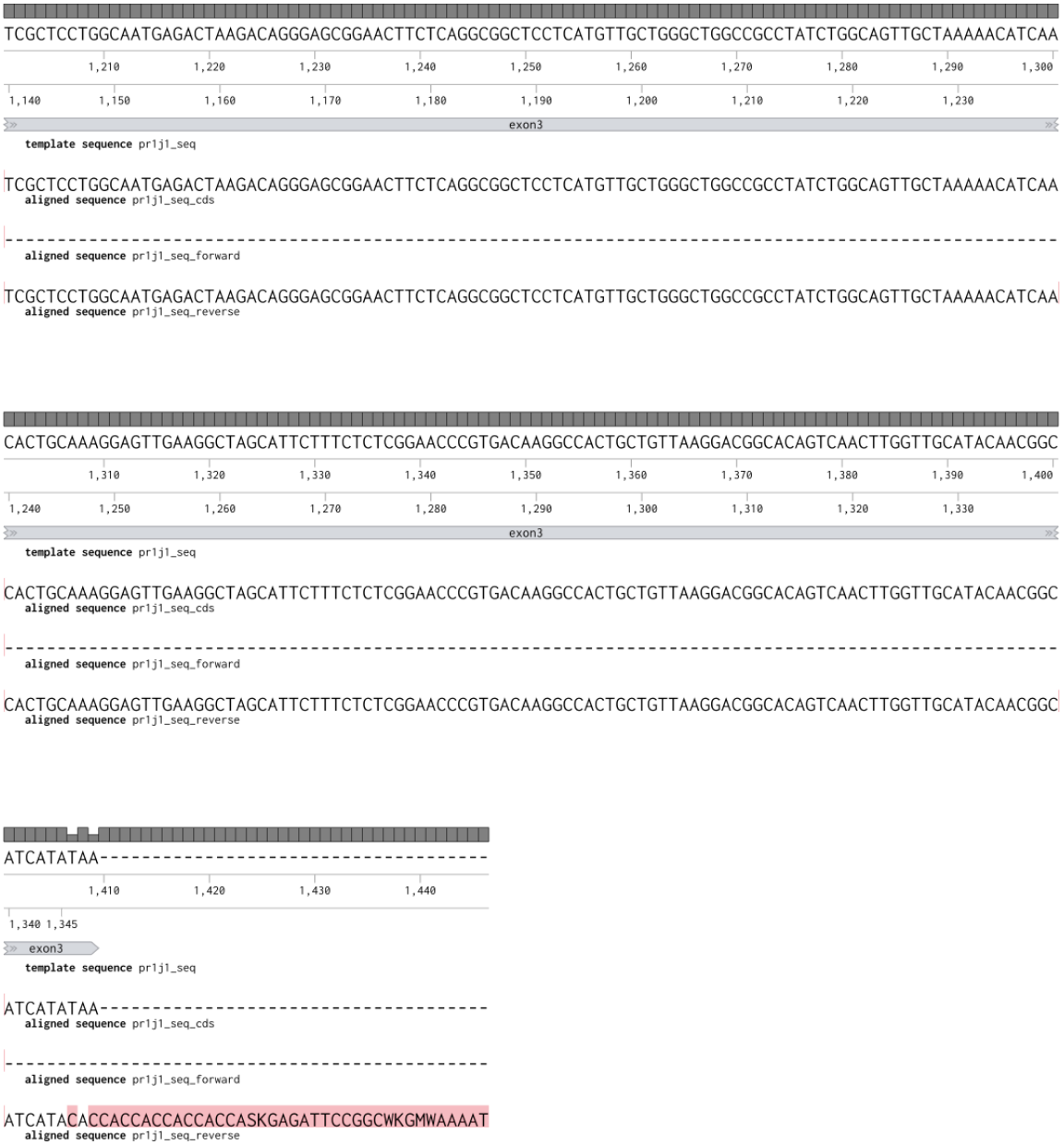

GTCACAATCTGGGGGCTGACGTGAAGATGTTTCAAGGTTG-------AGAGCGTCGAGCAGCAGCAAGTGTATCATCTGCACGAACTGACCACCCAGAAG
410 420 430 440 450 460 470 480 490 500
360 370 380 390 400 410 420 430 440
intron1
exon2
template sequence pr1j1_seq
----------GTTG-------AGAGCGTCGAGCAGCAGCAAGTGTATCATCTGCACGAACTGACCACCCAGAAG
aligned sequence pr1j1_seq_cds
----------GTTG-------AGAGCGTCGAGCAGCAGCAAGTGTATCATCTGCACGAACTGACCACCCAGAAG
aligned sequence pr1j1_seq_forward
TYTTTGAGCAGAGCATTACCCTGAAAGCTWTCGGAGAACGCKACGTTRGAGCGTCGRGCMGCAGCAAGKGTATCATTC-TGCACAACKGMCACCCAGAAG
aligned sequence pr1j1_seq_reverse


GAATCTACTCACGGTCTTGCTACTATTTCCCATAGAGAACCCGGGTCAACCGAATATGTCTATGACAGTAGTGCCGGCGAGGGATCTACCGTGTATGTTC
510 520 530 540 550 560 570 580 590 600
450 460 470 480 490 500 510 520 530 540
exon2
template sequence pr1j1_seq
GAATCTACTCACGGTCTTGCTACTATTTCCCATAGAGAACCCGGGTCAACCGAATATGTCTATGACAGTAGTGCCGGCGAGGGATCTACCGTGTATGTTC
aligned sequence pr1j1_seq_cds
GAATCTACTCACGGTCTTGCTACTATTTCCCATAGAGAACCCGGGTCAACCGAATATGTCTATGACAGTAGTGCCGGCGAGGGATCTACCGTGTATGTTC
aligned sequence pr1j1_seq_forward
GRATC-WCTCAC-GTCTTGCTACTATTTCCCATAGAGAACCCGGGTCAACCGAATATGTCTATGACAGTAGTGCCGGCGAGGGATCTACCGTGTATGTTC
aligned sequence pr1j1_seq_reverse

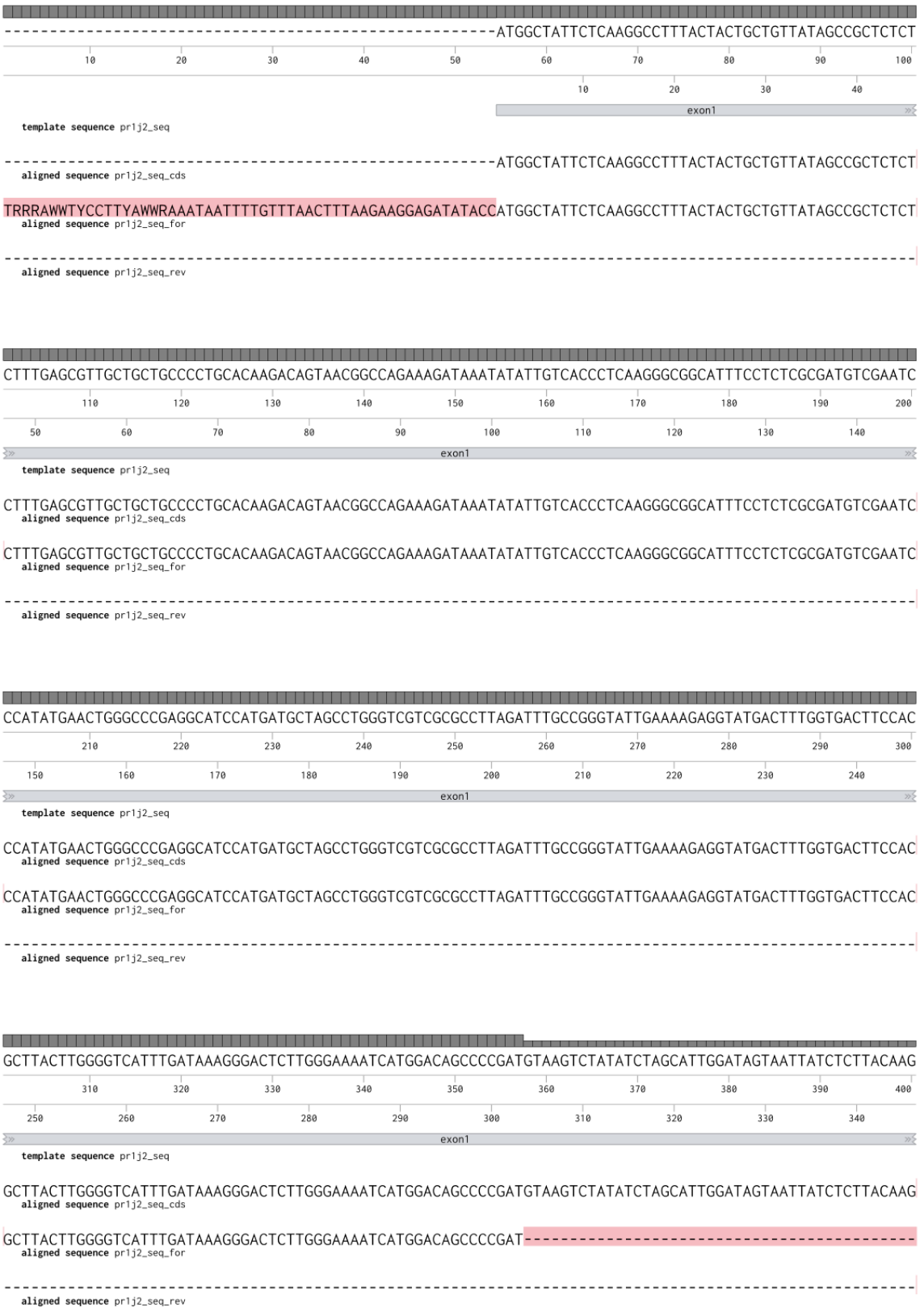

TCGACAGTGGCATTCAGGTGGATCATCCAGAATTCGAGGGCCGTGCTATTCGCGGGTATAATGCTGTCAAGGATGCCACAGACGAAGATGTGCAAGGACA
610 620 630 640 650 660 670 680 690 700
550 560 570 580 590 600 610 620 630 640
exon2
template sequence pr1j1_seq
TCGACAGTGGCATTCAGGTGGATCATCCAGAATTCGAGGGCCGTGCTATTCGCGGGTATAATGCTGTCAAGGATGCCACAGACGAAGATGTGCAAGGACA
aligned sequence pr1j1_seq_cds
TCGACAGTGGCATTCAGGTGGATCATCCAGAATTCGAGGGCCGTGCTATTCGCGGGTATAATGCTGTCAAGGATGCCACAGACKAAGATGTGCWAGGACA
aligned sequence pr1j1_seq_forward
TCGACAGTGGCATTCAGGTGGATCATCCAGAATTCGAGGGCCGTGCTATTCGCGGGTATAATGCTGTCAAGGATGCCACAGACGAAGATGTGCAAGGACA
aligned sequence pr1j1_seq_reverse


CGGCACTCACGTTGCTGGCATTGTTGGTAGCAAGACATACGGTGTTGCCAAGAAGACTAAGTTGGTAGATGTCAAGATGTTCCATGACGCAGGCAGCACC
710 720 730 740 750 760 770 780 790 800
650 660 670 680 690 700 710 720 730 740
exon2
template sequence pr1j1_seq
CGGCACTCACGTTGCTGGCATTGTTGGTAGCAAGACATACGGTGTTGCCAAGAAGACTAAGTTGGTAGATGTCAAGATGTTCCATGACGCAGGCAGCACC
aligned sequence pr1j1_seq_cds
CGGCACTCACGTTGCTGGCATTGTTGGTAGCAAGACATACGGTGTTGCCAAGAAGACTAAGTTGGTAGATGTCAAGATGTTCCATGACGCAGGCAGCACC
aligned sequence pr1j1_seq_forward
CGGCACTCACGTTGCTGGCATTGTTGGTAGCAAGACATACGGTGTTGCCAAGAAGACTAAGTTGGTAGATGTCAAGATGTTCCATGACGCAGGCAGCACC
aligned sequence pr1j1_seq_reverse


AATGAGATTATCCTGGACGGAATCGAGTGGACAATCAA-GGACATTACAGCTAAGCAGATTCAGAACCGGACCGTTGTCAACATGTCTTTTGGTAGGTGT
exon2
intron2
template sequence pr1j1_seq
AATGAGATTATCCTGGACGGAATCGAGTGGACAATCAA-GGACATTACAGCTAAGCAGATTCAGAACCGGACCGTTGTCAACATGTCTTTT---------
aligned sequence pr1j1_seq_cds
AATGAGATTATCCTGGACGGAATCGAGTGGACARTCMACGGACATTACAGCTAAKCAGATTCAGAACCSGACCGTTGTC-ACATGTCTTTTGGCG-----
aligned sequence pr1j1_seq_forward
AATGAGATTATCCTGGACGGAATCGAGTGGACAATCAA-GGACATTACAGCTAAGCAGATTCAGAACCGGACCGTTGTCAACATGTCTTTT---------
aligned sequence pr1j1_seq_reverse
GACAGTTTGTCCGTCGTCGCCCGTCAAAGCTCGCATCACTGACGATATTCTTTCCCCCCCAACAGGCGGCGGAAACTCTACTGCCCTGAACAAGATAATA
intron2
exon3
template sequence pr1j1_seq
----------GGCGGCGGAAACTCTACTGCCCTGAACAAGATAATA
aligned sequence pr1j1_seq_cds
----------CGGAACTCTACTGCCT--GACCAGATAAT-
aligned sequence pr1j1_seq_forward
----------GGCGGCGGAAACTCTACTGCCCTGAACAAGATAATA
aligned sequence pr1j1_seq_reverse
AAGACTGCGTACGACGCAGGTATTCTGTGCGTCATCTCGTCTGGGAACATGGGTGTCGATGCCTCAGACTGGTCTCCCGCCTCGTCCCCTGATGGCATCA
exon3
template sequence pr1j1_seq
AAGACTGCGTACGACGCAGGTATTCTGTGCGTCATCTCGTCTGGGAACATGGGTGTCGATGCCTCAGACTGGTCTCCCGCCTCGTCCCCTGATGGCATCA
aligned sequence pr1j1_seq_cds
AAGACTGC-TACGACGCA-GTATTCTGTGCGTCATCTCGTCW-GGAACATGGGTGTCGATGCCTCGACTKG--TCYCCGCCTACTYCCCTGAT-GCATCA
aligned sequence pr1j1_seq_forward
AAGACTGCGTACGACGCAGGTATTCTGTGCGTCATCTCGTCTGGGAACATGGGTGTCGATGCCTCAGACTGGTCTCCCGCCTCGTCCCCTGATGGCATCA
aligned sequence pr1j1_seq_reverse
CTGTTGGCGCCAT-TGATGCCAACTGGCGGCTATGGGA---TCACTCCAACCACGGCCCCGTTGTTCACATCTTGGCTCCTGGCGTGGACGTCTTGTCCC
exon3
template sequence pr1j1_seq
CTGTTGGCGCCAT-TGATGCCAACTGGCGGCTATGGGA---TCACTCCAACCACGGCCCCGTTGTTCACATCTTGGCTCCTGGCGTGGACGTCTTGTCCC
aligned sequence pr1j1_seq_cds
-TGTTGGCGCCATGCAGCCATAYTTGASYGCTATGKGGGAATCACT----------
aligned sequence pr1j1_seq_forward
CTGTTGGCGCCAT-TGATGCCAACTGGCGGCTATGGGA---TCACTCCAACCACGGCCCCGTTGTTCACATCTTGGCTCCTGGCGTGGACGTCTTGTCCC
aligned sequence pr1j1_seq_reverse

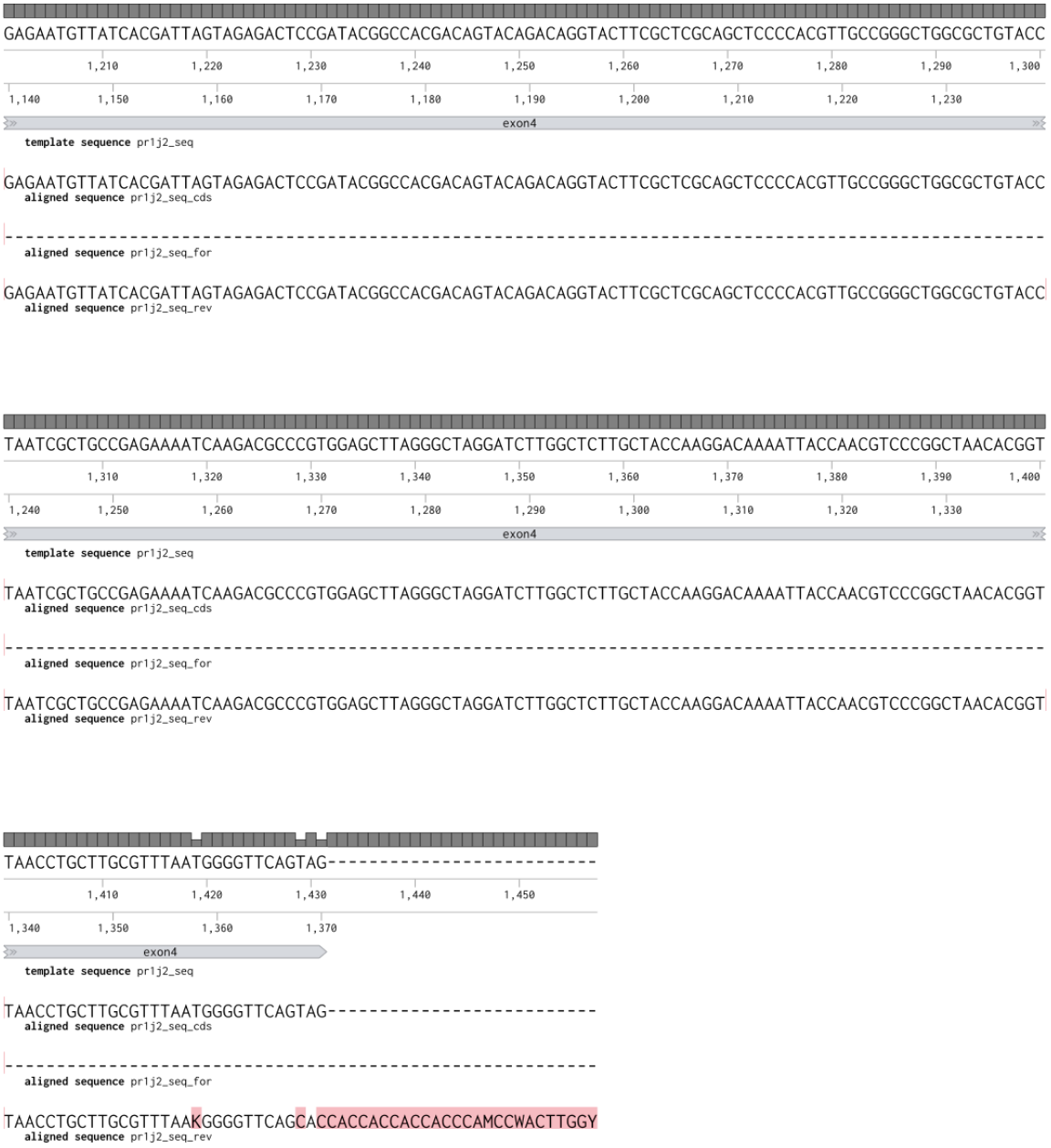

TCGCTCCTGGCAATGAGACTAAGACAGGGAGCGGAACTTCTCAGGCGGCTCCTCATGTTGCTGGGCTGGCCGCCTATCTGGCAGTTGCTAAAAACATCAA
1,210 1,220 1,230 1,240 1,250 1,260 1,270 1,280 1,290 1,300
1,140 1,150 1,160 1,170 1,180 1,190 1,200 1,210 1,220 1,230
exon3
template sequence pr1j1_seq
TCGCTCCTGGCAATGAGACTAAGACAGGGAGCGGAACTTCTCAGGCGGCTCCTCATGTTGCTGGGCTGGCCGCCTATCTGGCAGTTGCTAAAAACATCAA
aligned sequence pr1j1_seq_cds
aligned sequence pr1j1_seq_forward
TCGCTCCTGGCAATGAGACTAAGACAGGGAGCGGAACTTCTCAGGCGGCTCCTCATGTTGCTGGGCTGGCCGCCTATCTGGCAGTTGCTAAAAACATCAA
aligned sequence pr1j1_seq_reverse
CACTGCAAAGGAGTTGAAGGCTAGCATTCTTTCTCTCGGAACCCGTGACAAGGCCACTGCTGTTAAGGACGGCACAGTCAACTTGGTTGCATACAACGGC
1,310 1,320 1,330 1,340 1,350 1,360 1,370 1,380 1,390 1,400
1,240 1,250 1,260 1,270 1,280 1,290 1,300 1,310 1,320 1,330
exon3
template sequence pr1j1_seq
CACTGCAAAGGAGTTGAAGGCTAGCATTCTTTCTCTCGGAACCCGTGACAAGGCCACTGCTGTTAAGGACGGCACAGTCAACTTGGTTGCATACAACGGC
aligned sequence pr1j1_seq_cds
aligned sequence pr1j1_seq_forward
CACTGCAAAGGAGTTGAAGGCTAGCATTCTTTCTCTCGGAACCCGTGACAAGGCCACTGCTGTTAAGGACGGCACAGTCAACTTGGTTGCATACAACGGC
aligned sequence pr1j1_seq_reverse
ATCATATAA
1,410 1,420 1,430 1,440
1,340 1,345
exon3
template sequence pr1j1_seq
ATCATATAA
aligned sequence pr1j1_seq_cds
aligned sequence pr1j1_seq_forward
ATCATACACCACCACCACCACCASKGAGATTCCGGCWKGMWAAAAT
aligned sequence pr1j1_seq_reverse

Anexo 20: Sequenciamento do inserto do plasmídeo pET23d(+)-*pr1J2.*




AAATACTACTAAACGGATCTATTGATAGGTACTAGGCGTTGAGCCTGATGGAATAA-CAATGCCATTAGCGTTGACAACGCAGCAGAACCCTCCGTGGGC
410 420 430 440 450 460 470 480 490 500
350 360 370 380 390 400 410 420 430 440
exon1
template sequence pr1j2_seq
AAATACTACTAAACGGATCTATTGATAGGTACTAGGCGTTGAGCCTGATGGAATAA-CAATGCCATTAGCGTTGACAACGCAGCAGAACCCTCCGTGGGC
aligned sequence pr1j2_seq_cds
----------GTACTAGGCGTTGAGCCTGATGGAATAA-CAATGCCATTAGCGTTGACAACGCAGCAGAACCCTCCGTGGGC
aligned sequence pr1j2_seq_for
--ATTCTCATGGCC----------KTGWWGACCTTGAWTGAATAACCAATGCCWTTAGCG-TGGCAATGCAGCAGATCCYTCCGTGGGC
aligned sequence pr1j2_seq_rev
TCTAAGCGCCATGTCTAGCCGAACACCAGGCCCCCCCAGCCTTATCGGTACGACGATAGCGCTGGCCAAAATACCTTTGCA-TACGTACTAGATACGGGT
510 520 530 540 550 560 570 580 590 600
450 460 470 480 490 500 510 520 530 540
exon1
intron1
template sequence pr1j2_seq
TCTAAGCGCCATGTCTAGCCGAACACCAGGCCCCCCCAGCCT----------
aligned sequence pr1j2_seq_cds
TCTAAGCGCCATGTCTAGCCGAACACCAGGCCCCCCCAGCCTTATCGGTACGACGATAGCGCTGGCCAAAATACCTTTGCA-TACGTACTAGATACGGGT
aligned sequence pr1j2_seq_for
YTTAGGCGCCATGTCTAGCCG-ACTCCAGGCCCCCCCCAGCTTATCGGTACGACGAWAGCGSTGGCCAAAATACCTTTGCAWTACGTACTAGAWWCGGGK
aligned sequence pr1j2_seq_rev
GTCCACGATAAACATGTGGAATTCGGTGGCCGTGTTACCCCGGGCTGGAGTGCTTACGAAGAGGAC-TTCCCAGACAGGCCGCATGGTGATGTTACTGGC
610 620 630 640 650 660 670 680 690 700
550 560 570 580 590 600 610 620 630 640
exon2
template sequence pr1j2_seq
-TCCACGATAAACATGTGGAATTCGGTGGCCGTGTTACCCCGGGCTGGAGTGCTTACGAAGAGGAC-TTCCCAGACAGGCCGCATGGTGATGTTACTGGC
aligned sequence pr1j2_seq_cds
GTCCACGATAAACATGTGGAATTCGGTGGCCGTGTTACCCCGGGCTGGAGTGCTTACGAAGAGGAC-TTCCCAGACAGGCCGCATGGTGATGTTACTGGC
aligned sequence pr1j2_seq_for
GTCCACGAT-AACATGTGGAATTCGGTGGCCGTGTTACCCCGGGCTGGAGTGCTTACGAAGAGGACTTTCCCAGACAGGCCGCATGKTGATGTTWCTGGC
aligned sequence pr1j2_seq_rev
CACGGAACCATGGTTGCCGGTATTATCGCCTCCAATACCTACGGTGTCGCCAAGAAGGCAAATATCATTGCTGTCCAGACGGATCAGACAGTATCAGGAT
710 720 730 740 750 760 770 780 790 800
650 660 670 680 690 700 710 720 730 740
exon2
template sequence pr1j2_seq
CACGGAACCATGGTTGCCGGTATTATCGCCTCCAATACCTACGGTGTCGCCAAGAAGGCAAATATCATTGCTGTCCAGACGGATCAGACAGTATCAGGAT
aligned sequence pr1j2_seq_cds
CACGGAACCATGGTTGCCGGTATTATCGCCTCCAATACCTACGGTGTCGCCAAGAAGGCAAATATCATTGCTGTCCAGACGGATCAGACAGTATCAGGAT
aligned sequence pr1j2_seq_for
CACGGAACCATGGTTGCCGGTATTATCGCCTCCAATACCTACGGTGTCGCCAAGAAGGCAAATATCATTGCTGTCCAGACGGATCAGACAGTATCAGGAT
aligned sequence pr1j2_seq_rev


TGCT-AGGCGGCATAGCGTGGGCCGTGCGGGATATCCAGAGTCAAGGCCGCGTCGGTCAGGCCGTTATTAATTACTCGGGAGGTACGTATGCGAGATGTT
exon2
intron2
template sequence pr1j2_seq
TGCT-AGGCGGCATAGCGTGGGCCGTGCGGGATATCCAGAGTCAAGGCCGCGTCGGTCAGGCCGTTATTAATTACTCGGGAGGTACGTATGC--------
aligned sequence pr1j2_seq_cds
TGCT-AGGCGGCATAGCGTGGGCCGTGCGGGATATCCAGAGTCAAGGCCGCGTCGGTCAGGCCGTTATTAATTACTCGGG----------
aligned sequence pr1j2_seq_for
TGCTWGGGCGGCATAGCGTGGGCCGTGCGGGATATCCAGAGTCAAGGCCGCGTCGGTCAGGCCGTTATTAATTACTCGGG----------
aligned sequence pr1j2_seq_rev
TTGCACCTCTTGGAGTTTTTTGCATCCTAATTAGAAGTAGGGCTCCCCACTATTCCCGATTCTGCCGGATACAAGTATCAACCCGGAATTGCCATGGCCC
intron2
exon3
template sequence pr1j2_seq
----------GTAGGGCTCCCCACTATTCCCGATTCTGCCGGATACAAGTATCAACCCGGAATTGCCATGGCCC
aligned sequence pr1j2_seq_cds
----------AGGGCTCCCCACTATTCCCGATTCTGCCGGATACAAGTATCAACCCGGAATTGCCATGGCCC
aligned sequence pr1j2_seq_for
----------AGGGCTCCCCACTATTCCCGATTCTGCCGGATACAAGTATCAACCCGGAATTGCCATGGCCC
aligned sequence pr1j2_seq_rev
AAAGTATGGATATTGCGTTTAACGAAGGG-ATTCTCTGCGTCATTGCTGCTGGCAACGACGGTAAAGT--AGTTGAACAATCCACAACTCCTTATCAAGG
exon3
intron3
template sequence pr1j2_seq
AAAGTATGGATATTGCGTTTAACGAAGGG-ATTCTCTGCGTCATTGCTGCTGGCA----------
aligned sequence pr1j2_seq_cds
-AAGTATGATTA-TGCGTTTAACGAAGGGAATTCTCTGCGTCATTGCTGCTGGCACGAACGGTWAAGCTAGTTTGAACAATYCCCMMAC----------
aligned sequence pr1j2_seq_for
AAAGTATGGATATTGCGTTTAACGAAGGG-ATTCTCTGCGTCATTGCTGCTGGCAACGACGGTAAAGT--AGTTGAACAATCCACAACTCCTTATCAAGG
aligned sequence pr1j2_seq_rev
CAACTCTACTACTGCTTTAGTGGTCGGCGGAATAAATCAACAATGGGATTTTATGCGCCTCTCCAACCACGGCCCCAGCGTCGATATTCTCGCTCCTGGT
intron3
exon4
template sequence pr1j2_seq
----------GGAATAAATCAACAATGGGATTTTATGCGCCTCTCCAACCACGGCCCCAGCGTCGATATTCTCGCTCCTGGT
aligned sequence pr1j2_seq_cds
----
aligned sequence pr1j2_seq_for
CAACTCTACTACTGCTTTAGTGGTCGGCGGAATAAATCAACAATGGGATTTTATGCGCCTCTCCAACCACGGCCCCAGCGTCGATATTCTCGCTCCTGGT
aligned sequence pr1j2_seq_rev


GAGAATGTTATCACGATTAGTAGAGACTCCGATACGGCCACGACAGTACAGACAGGTACTTCGCTCGCAGCTCCCCACGTTGCCGGGCTGGCGCTGTACC
1,210 1,220 1,230 1,240 1,250 1,260 1,270 1,280 1,290 1,300
1,140 1,150 1,160 1,170 1,180 1,190 1,200 1,210 1,220 1,230
exon4
template sequence pr1j2_seq
GAGAATGTTATCACGATTAGTAGAGACTCCGATACGGCCACGACAGTACAGACAGGTACTTCGCTCGCAGCTCCCCACGTTGCCGGGCTGGCGCTGTACC
aligned sequence pr1j2_seq_cds
aligned sequence pr1j2_seq_for
GAGAATGTTATCACGATTAGTAGAGACTCCGATACGGCCACGACAGTACAGACAGGTACTTCGCTCGCAGCTCCCCACGTTGCCGGGCTGGCGCTGTACC
aligned sequence pr1j2_seq_rev
TAATCGCTGCCGAGAAAATCAAGACGCCCGTGGAGCTTAGGGCTAGGATCTTGGCTCTTGCTACCAAGGACAAAATTACCAACGTCCCGGCTAACACGGT
1,310 1,320 1,330 1,340 1,350 1,360 1,370 1,380 1,390 1,400
1,240 1,250 1,260 1,270 1,280 1,290 1,300 1,310 1,320 1,330
exon4
template sequence pr1j2_seq
TAATCGCTGCCGAGAAAATCAAGACGCCCGTGGAGCTTAGGGCTAGGATCTTGGCTCTTGCTACCAAGGACAAAATTACCAACGTCCCGGCTAACACGGT
aligned sequence pr1j2_seq_cds
aligned sequence pr1j2_seq_for
TAATCGCTGCCGAGAAAATCAAGACGCCCGTGGAGCTTAGGGCTAGGATCTTGGCTCTTGCTACCAAGGACAAAATTACCAACGTCCCGGCTAACACGGT
aligned sequence pr1j2_seq_rev
TAACCTGCTTGCGTTTAATGGGGTTCAGTAG----------
1,410 1,420 1,430 1,440 1,450
1,340 1,350 1,360 1,370
exon4
template sequence pr1j2_seq
TAACCTGCTTGCGTTTAATGGGGTTCAGTAG----------
aligned sequence pr1j2_seq_cds
aligned sequence pr1j2_seq_for
TAACCTGCTTGCGTTTAAKGGGGTTCAGCACCACCACCACCCAMCCWACTTGGY
aligned sequence pr1j2_seq_rev

CURRICULUM VITÆ resumido

ANDREIS, F. C. ; ANDREIS, FABIO CARRER

Fabio Carrer Andreis
Caxias do Sul, Rio Grande do Sul, Brasil, 25/01/1992
fabio.andreis@gmail.com

Formação Acadêmica

2016-2021 Doutorado em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Orientador: Augusto Schrank.
Coorientadora: Claudia Elizabeth Thompson.
Bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.

2014-2016 Mestrado em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Título: Evolução Molecular de Proteases Pr1 (Classe II) de Metarhizium anisopliae, Ano de Obtenção: 2016.
Orientador: Augusto Schrank.
Coorientadora: Claudia Elizabeth Thompson.
Bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brasil.

2019-Atual Graduação em andamento em Matemática Aplicada e Computacional.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

2010-2013 Graduação em Biotecnologia (ênfase Bioinformática).
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Título: Caracterização Filogenética de Ureases.
Orientador: Hugo Verli.
Bolsista da: Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, FAPERGS, Brasil.

Formação Complementar

2019 *Complete Python Bootcamp.* (Carga horária: indefinida).
Udemy, Inc.

2019 Dinâmica Molecular. (Carga horária: 15h).
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

2017 Reconstrução, Modelagem e Análise de Redes Metabólicas. (Carga

horária: 15h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2017 Análise de Transcritomas e microRNAs. (Carga horária: 36h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2016 Cálculos Quânticos de Estrutura Eletrônica Semi-Empíricos. (Carga horária: 6h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2016 Dinâmica Molecular Básica. (Carga horária: 6h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2016 Cálculo de Modos Normais em Proteínas. (Carga horária: 6h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2015 *The Genomics Bootcamp Workshop*. (Carga horária: 40h).
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

2014 Análise de Sequências Biológicas. (Carga horária: 36h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

2013 Computação de alto desempenho em placas GPUs. (Carga horária: 63h).
Laboratório Nacional de Computação Científica, LNCC, Brasil.

Estágios

08/2013 - 12/2013 Estágio Curricular
Unidade de Biologia Teórica e Computacional
Centro de Biotecnologia
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Orientador: Augusto Schrank.

07/2010 - 08/2013 Iniciação Científica
Grupo de Bioinformática Estrutural
Centro de Biotecnologia
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Orientador: Hugo Verli

Prêmios e distinções

2019 Destaque de Sessão - Congresso UFCSPA: conectando saúde e sociedade
Universidade Federal de Ciências da Saúde de Porto Alegre.

2016 Melhor Poster - Categoria Pós-Graduação

VIII Escola de Modelagem Molecular em Sistemas Biológicos.

2012 Destaque
Salão de Iniciação Científica UFRGS, PROPESQ.

2011 Destaque
Salão de Iniciação Científica UFRGS, PROPESQ.

Experiência Profissional

07/2017 - 07/2018 Professor Substituto
Departamento de Informática Teórica
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Disciplinas: Biologia Computacional (Bacharelado em Biotecnologia - Bioinformática); Linguagens Formais e Autômatos (Bacharelado em Ciência da Computação)

Artigos Completos Publicados

ANDREIS, FABIO CARRER; SCHRANK, AUGUSTO; THOMPSON, CLAUDIA ELIZABETH. Molecular evolution of Pr1 proteases depicts ongoing diversification in Metarhizium spp. MOLECULAR GENETICS AND GENOMICS, v. -, p. 1-17, 2019.

DE OLIVEIRA, EDER SILVA; JUNGES, ÂNGELA ; SBARAINI, NICOLAU ; **ANDREIS, FÁBIO CARRER** ; THOMPSON, CLAUDIA ELIZABETH ; STAATS, CHARLEY CHRISTIAN ; SCHRANK, AUGUSTO . Molecular evolution and transcriptional profile of GH3 and GH20 β-N-acetylglucosaminidases in the entomopathogenic fungus Metarhizium anisopliae. GENETICS AND MOLECULAR BIOLOGY (ONLINE VERSION), v. 41, p. 4, 2018.

SBARAINI, NICOLAU; **ANDREIS, FABIO C.**; THOMPSON, CLAUDIA E.; GUEDES, RAFAEL L. M.; JUNGES, ANGELA; CAMPOS, THAIS; STAATS, CHARLEY C.; VAINSTEIN, MARILENE H.; RIBEIRO DE VASCONCELOS, ANA T.; SCHRANK, AUGUSTO. Genome-Wide Analysis of Secondary Metabolite Gene Clusters in Ophiostoma ulmi and Ophiostoma novo-ulmi Reveals a Fujikurin-Like Gene Cluster with a Putative Role in Infection. Frontiers in Microbiology, v. 8, p. 1063, 2017.

SBARAINI, NICOLAU; GUEDES, RAFAEL LUCAS MUNIZ; **ANDREIS, FÁBIO CARRER**; JUNGES, ÂNGELA; DE MORAIS, GUILHERME LOSS; VAINSTEIN, MARILENE HENNING; DE VASCONCELOS, ANA TEREZA RIBEIRO; SCHRANK, AUGUSTO. Secondary metabolite gene clusters in the entomopathogen fungus Metarhizium anisopliae: genome identification and patterns of expression in a cuticle infection model. BMC GENOMICS, v. 17, p. 736, 2016.

STAATS, CHARLEY CHRISTIAN; JUNGES, ÂNGELA; GUEDES, RAFAEL LUCAS; THOMPSON, CLAUDIA ELIZABETH; DE MORAIS, GUILHERME LOSS; BOLDO, JULIANO TOMAZZONI; DE ALMEIDA, LUIZ GONZAGA; **ANDREIS, FÁBIO CARRER**; GERBER, ALEXANDRA LEHMKUHL; SBARAINI, NICOLAU; DA PAIXÃO, RANA LOUISE; BROETTO, LEONARDO; LANDELL, MELISSA; SANTI, LUCÉLIA; DA SILVA, WALTER ORLANDO; SILVEIRA, CAROLINA PEREIRA; SERRANO, THAIANE RISPOLI; DE OLIVEIRA, EDER SILVA; KMETZSCH, LÍVIA; VAINSTEIN, MARILENE HENNING; DE VASCONCELOS, ANA TEREZA; SCHRANK, AUGUSTO. Comparative genome analysis of entomopathogenic fungi reveals a complex set of secreted proteins. BMC Genomics, v. 15, p. 822, 2014.

LIGABUE-BRAUN, RODRIGO; **ANDREIS, FÁBIO CARRER**; VERLI, HUGO; CARLINI, CÉLIA REGINA. 3-to-1: unraveling structural transitions in ureases. Science of Nature, v. 100, p. 459-467, 2013.