UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Pedro Magalhães Martins

PROPEDIA: uma base de dados de estruturas de complexos proteína-peptídeo não redundante baseada em agrupamentos híbridos

Belo Horizonte

2020

Pedro Magalhães Martins

PROPEDIA: uma base de dados de estruturas de complexos proteína-peptídeo não redundante baseada em agrupamentos híbridos

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Doutor em Bioinformática.

Orientadora:
Profa. Dra. Raquel Cardoso de Melo Minardi
Coorientador:
Dr. Diego César Batista Mariano

Belo Horizonte

2020

043 Martins, Pedro Magalhães.
Propedia: uma base de dados de estruturas de complexos proteína-peptídeo não redundante baseada em agrupamentos híbridos [manuscrito] / Pedro Magalhães Martins. – 2020.
93 f. : il. ; 29,5 cm.

Orientadora: Profa. Dra. Raquel Cardoso de Melo Minardi. Coorientador: Dr. Diego César Batista Mariano.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Base de Dados. 3. Elementos Estruturais de Proteínas. 4. Peptídeos. I. Minardi, Raquel Cardoso de Melo. II. Mariano, Diego César Batista. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-graduação em Bioinformática

ATA DA DEFESA DE TESE

PEDRO MAGALHÃES MARTINS

Às quatorze horas do dia **15 de dezembro de 2020**, reuniu-se, através do aplicativo zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Propedia: uma base de dados de estruturas de complexos proteína-peptídeo não redundante baseada em agrupamentos híbridos"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Raquel Cardoso de Melo Minardi**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**Indicação**
Dra. Raquel Cardoso de Melo Minardi	UFMG	Aprovado
Dr. Diego César Batista Mariano	UFMG	Aprovado
Dr. Lucas Bleicher	UFMG	Aprovado
Dr. Tiago Antônio Oliveira Mendes	UFV	Aprovado
Dr. Hugo Verli	UFRGS	Aprovado
Dr. Bruno Silva Andrade	UESB	Aprovado
Dra. Karina dos Santos Machado	FURG	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado

O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 15 de dezembro de 2020.

Dra. Raquel Cardoso de Melo Minardi - Orientadora

Dr. Diego César Batista Mariano- Coorientador

Dr. Lucas Bleicher

Dr. Tiago Antônio Oliveira Mendes

Dr. Hugo Verli

Dr. Bruno Silva Andrade

Dra. Karina dos Santos Machado

Documento assinado eletronicamente por **Diego César Batista Mariano**, **Usuário Externo**, em 15/12/2020, às 18:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tiago Antônio de Oliveira Mendes**, **Usuário Externo**, em 15/12/2020, às 18:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Hugo Verli**, **Usuário Externo**, em 15/12/2020, às 18:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Karina dos Santos Machado**, **Usuário Externo**, em 15/12/2020, às 18:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Silva Andrade**, **Usuário Externo**, em 15/12/2020, às 18:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Lucas Bleicher**, **Professor do Magistério Superior**, em 15/12/2020, às 18:16, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Raquel Cardoso de Melo Minardi**, **Subcoordenador(a)**, em 22/01/2021, às 18:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0431468** e o código CRC **F2CBF330**.

Agradecimentos

Agradeço às agências de fomento à pesquisa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Este projeto foi financiado pela CAPES: edital Biologia Computacional. Número de processo 23038.004007/2014-82.

"Escolhe um trabalho de que gostes e não terás que trabalhar nem um dia na tua vida
(Confúcio)

Resumo

Interações proteína-peptídeo atuam em uma ampla variedade de processo biológicos, como sinalização celular, redes regulatórias, sistema imunológico, e inibição enzimática. Estima-se que 40% das interações proteicas são mediadas por peptídeos, e apesar disso, há poucos dados estruturais disponíveis sobre tais interações. Peptídeos são cadeias polipeptídicas formadas por poucos aminoácidos e por isso são mais flexíveis e versáteis. Os peptídeos possuem uma conformação estrutural de caráter transitório, e por isso podem ser mais facilmente manipulados. Além disso, peptídeos em geral possuem baixa toxicidade e sua área de interface, normalmente pequena, faz deles alvos atrativos para propostas terapêuticas, planejamento racional de fármacos e inibidores proteicos. De fato, estima-se que 10% do mercado farmacêutico seja composto por medicamentos à base de peptídeos e esse percentual continua em ascensão.

O presente trabalho apresenta uma base de dados abrangente e atualizada de estruturas de complexos proteína-peptídeo, denominado Propedia. Propedia compreende mais de 19.000 complexos de alta resolução obtidos do Protein Data Bank (PDB), adicionando ainda novas informações relacionadas a interação proteína-peptídeo. Apesar de existirem outras bases de dados de estruturas proteína-peptídeo, Propedia se destaca por propor um algoritmo de agrupamento híbrido, a fim de unir complexos similares e, consequentemente, reduzir a redundância, conforme a necessidade do pesquisador. O Propedia dispõe de três agrupamento de complexos: por sequências de peptídeos; por estruturas de interface proteína-peptídeo; e por sítio de ligação.

Por fim, foram realizados estudos de caso para validar a utilidade do Propedia em encontrar peptídeos promissores para alvos proteicos. Entre eles, estão uma previsão de estruturas de peptídeos para inibir a principal protease da síndrome respiratória aguda grave do coronavírus 2 (SARS-CoV-2) e também a principal protease da lagarta-da-soja (*Anticarsia gemmatalis* hübner), considerada uma importante praga desfolhadora que causa prejuízos na produção de soja no Brasil.

O Propedia está disponível por meio de uma serviço *web* (bioinfo.dcc.ufmg.br/propediadb), com interface amigável e intuitiva, na qual os usuários podem explorar e analisar os complexos, realizar o *download* de parte ou de toda a base de dados. Propedia permite ainda submeter sequências ou estrutura de interesse para recuperar complexos semelhantes considerando a estrutura primária do peptídeo/proteína alvo ou seu sítio de ligação.

Palavras-chave: base de dados, estrutura de proteínas; complexo proteína-peptídeo, peptídeo, servidor *web*

Abstract

Protein-peptide interactions play in a wide variety of biological processes, such as cell signaling, regulatory networks, immune system, and enzyme inhibition. It is estimated that 40% of protein interactions are mediated by peptides, nevertheless, there is little structural data information available about these interactions. Peptides are polypeptide chains made up of few amino acids and therefore are more flexible and versatile. The peptides have a structural conformation of a transitory character, and therefore can be more easily manipulated. In addition, peptides in general have low toxicity and their normally small interface area make them attractive targets for therapeutic proposals, rational drug design and protein inhibitors. In fact, it is estimated that 10% of the pharmaceutical market be composed of peptide-based, which continues to rise.

The present work presents a comprehensive and up-to-date database of protein-peptide complex structures, named Propedia. Propedia comprises over 19,000 high-resolution complex retrieved from the Protein Data Bank (PDB), adding new features about these protein-peptide interactions. Although there are other databases of protein-peptide structures, Propedia stands out for proposing hybrid clustering algorithm, in order to gather similar complexes and, consequently, remove redundancy, according to the researcher's need. Propedia has three types of clustering: by peptide sequences; by protein-peptide interface structures; and by binding sites.

Finally, case studies were carried out to validate the usefulness of Propedia in finding promising peptides for protein targets. Among them are a prediction of peptide structures to inhibit the main protease of acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and also the main protease of the soybean caterpillar (*Anticarsia gemmatalis* hübner), considered an important defoliating pest that harms the production of soybean in Brazil.

Propedia is available through a web service (http://bioinfo.dcc.ufmg.br/propediadb/), with a user-friendly and intuitive interface, in which users can explore and analyze complexes, download part or all of database. Propedia also allows submitting sequences or structure of interest to recover similar complexes considering the target peptide/protein primary structure or its binding site.

Keywords: database; protein strucuture; protein-peptide complex; peptide; web server

Lista de ilustrações

Figura 1 – Estrutura dos aminoácidos. 17
Figura 2 – Formação de uma ligação peptídica. 17
Figura 3 – Modelo da base de dados do Propedia 25
Figura 4 – Fluxo da coleta e análise de dados para povoamento inicial da base de dados. 26
Figura 5 – Exemplo de um complexo proteína-peptídeo. 28
Figura 6 – Exemplo de um complexo proteína-peptídeo ao qual o peptídeo interage com mais de uma proteína (receptor). Imagem gerada com PyMOL (http://pymol.org). Fonte: próprio autor. 28
Figura 7 – Cálculo da área da interface (AI). 29
Figura 8 – Fluxo do processo de agrupamento de sequências de peptídeos. 32
Figura 9 – Exemplo de logo de sequência do ASP. 33
Figura 10 – Exemplo da tela de execução do MUSTANGmod 34
Figura 11 – Diagrama de caixa dos resultados de RMSD e iRMSD 35
Figura 12 – Fluxo de dados por agrupamento de estrutura de interface. 36
Figura 13 – Sobreposição de duas estruturas como resultado do alinhamento do MUSTANG 37
Figura 14 – Relação da quantidade de grupos formados de acordo com os limites de iRMSD 38
Figura 15 – Exemplo simplificado do método de agrupamento baseado em grafo 38
Figura 16 – Comparação de grupos funcionais de duas proteínas 40
Figura 17 – Fluxo de dados do agrupamento por sítio de ligação (ASL). 41
Figura 18 – Comparação dos tipos de arquivos de saída do ProBiS. 42
Figura 19 – Relação da quantidade de grupos gerados de acordo com os limites de escore Z 43
Figura 20 – Gráfico de distribuição dos tamanhos dos peptídeos. 45
Figura 21 – Diagrama de caixa dos grupos do ASP. 47
Figura 22 – Diagrama de caixa dos grupos do AEI. 48
Figura 23 – Diagrama de caixa dos grupos do ASL. 50
Figura 24 – Página inicial do serviço *web*. 52
Figura 25 – Página para explorar os complexos com vários recursos de filtros e *downloads* 53
Figura 26 – Fluxo de navegação do Propedia utilizando o serviço *web*. 55
Figura 27 – Comparação estrutural e de interação entre os complexos 2jf9-B-Q e 4iv2-A-C. 57
Figura 28 – Poses das conformações estruturais de peptídeos geradas pelo FlexPepDock. 60
Figura 29 – Matriz de especificidade do MEROPS. 61
Figura 30 – Correlação da ΔG_{bind} da MetaD com o RMSD do sítio de ligação e métrica de alinhamento da MPro do SARS-CoV-2 com péptides sugeridos pelo Propedia. 62
Figura 31 – Modelo da serino protease de AG acoplado aos peptídeos 3qgn, cadeia A e 4dii, cadeia L. 64

Figura 32 – Modelo da serino protease de AG juntamente com as 4 primeiras poses dos peptídeos 6rw2, cadeia B, 3kn2, cadeia B e 2obq, cadeia B. 65
Figura 33 – Triplicatas dos panoramas de energia livre 82
Figura 34 – Sobreposição do alinhamento estrutural do modelo de protease de AG (em verde) e os modelos melhores classificados usados no procedimento de modelagem 84
Figura 35 – Mapas de interação para as três melhores resultados de FFC, dos complexos 6rw2_B_A (A), 3kn2_B_C (B) e 2obq_B_C (C) 85

Lista de tabelas

Tabela 1 – Resumo da quantidade de complexos identificados, por quantidade de cadeias de receptores interagindo com peptídeos e peptídeos apenas com aminácidos padrões. 30
Tabela 2 – Comparação das bases de dados de complexos proteína-peptídos 51
Tabela 3 – Características de comparação entre 2jf9-Q-B e 4iv2-C-A 56
Tabela 4 – Lista dos complexos recuperados a partir da busca dos 10 melhores resultados encontrados tendo como alvo a estrutura da Mpro do SARS-CoV-2. 58
Tabela 5 – Resultados do FlexPepDock em contraste com os resultados do Propedia. . 59
Tabela 6 – Lista dos peptídeos recuperados pela busca por sequência (receptor) utilizando sequência serino protease da AG 63
Tabela 7 – Lista dos peptídeos recuperados pela busca por sítio de ligação utilizando o modelo gerado pela sequência serino protease da AG e os resíduos da tríade catalítica (His6, Asp56 e Ser143) 63
Tabela 8 – Métrica e RMSD do HADDOCK para os modelos selecionados para cada cadeia de peptídeo no experimento baseado em sequência 64
Tabela 9 – Métrica do HADDOCK e FCC para os modelos selecionados para cada cadeia de peptídeo no experimento de sítio de ligação 65
Tabela 10 – Correlação entre a energia livre de ligação MetaD e Propedia 78

Lista de abreviaturas e siglas

AG:	*Anticarsia gemmatalis* Hübner
AI:	Área da interface
AEI:	Agrupamento por estrutura de interface
ASA:	*Accessible Surface Area*
ASL:	Agrupamento por sítio de ligação
ASP:	Agrupamento por sequência de peptídeo
CCA:	Conjunto de complexos agrupáveis
ΔG_{bind}:	Energia livre de ligação
FCC:	Fração de Contatos Comuns
iRMSD:	*Interface root mean square deviation*
DM:	Dinâmica molecular
MetaD:	Metadinâmica
Mpro:	Principal protease (*main protease*)
NMR:	*Nuclear magnetic resonance* spectroscopy
HMM:	*Hidden Markov Models*
RMSD:	*Root mean square deviation*
SARS-CoV-2:	Severe Acute Respiratory Syndrome Coronavirus 2
SGDB:	Sistema gerenciador de banco de dados
PDB:	*Protein Data Bank*

Sumário

1 INTRODUÇÃO 16
1.1 Proteínas e aminoácidos 16
1.2 Estrutura de proteínas 16
1.3 Peptídeos e complexos proteína-peptídeo 18
1.4 Bases de dados de proteína-peptídeo 19
1.4.1 PepX 20
1.4.2 PeptiDB 20
1.4.3 PepBind 20
1.4.4 Bases de dados recentes: PixelDB, PepBDB e PepPro 20
1.5 Justificativa 21

2 OBJETIVOS 23
2.1 Objetivo Geral 23
2.2 Objetivo Específicos 23

3 METODOLOGIA 24
3.1 Modelagem da base de dados 24
3.2 Coleta e análise dos dados de entrada 25
3.3 Sistema de agrupamento 30
3.3.1 Sequência de peptídeo 31
3.3.2 Estrutura de interface 33
3.3.2.1 Versões iniciais 34
3.3.2.2 Versão final 35
3.3.3 Sítio de ligação 39
3.4 Criação do serviço *web* 42

4 RESULTADOS E DISCUSSÃO 44
4.1 Tamanho de um peptídeo 44
4.2 Agrupamentos 44
4.2.1 Por sequência de peptídeo (ASP) 46
4.2.2 Por estrutura de interface (AEI) 46
4.2.3 Por sítio de ligação (ASL) 46
4.3 Produtos 49
4.3.1 Base de dados Propedia 49
4.3.2 Serviço *web* 51
4.4 Estudos de caso 56

4.4.1 Receptores de estrogênio com peptídeos diferentes 56
4.4.2 Interações da principal protease do SARS-CoV-2 com peptídeos (6lu7) 57
4.4.3 Metadinâmica aplicada aos os peptídeos de complexos com sítio de ligação similares a principal protease do SARS-CoV-2 60
4.4.4 Protease da lagarta-da-soja (*Anticarsia gemmatalis* Hübner) 61

5 CONCLUSÕES 67

6 PERSPECTIVAS 68

REFERÊNCIAS 69

APÊNDICES 77

APÊNDICE A – PROTOCOLOS APLICADOS A METADINÂMICA . 78
A.1 Seleção de peptídeos para validação de metadinâmica e configuração do sistema 78
A.2 Procedimentos de simulação 79
A.3 Mapas de energia livre e projeções das energias de metadinamicas ao longo das dimensões da CV$_{dist}$ e estimativa da ΔG$_{bind}$ . . . 80

APÊNDICE B – *ANTICARSIA GEMMATALIS* HÜBNER 83

APÊNDICE C – ARTIGOS PUBLICADOS 88
C.1 Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm 88
C.2 Vermont: a multi-perspective visual interactive platform for mutational analysis 89
C.3 Introducing programming skills for life science students 90
C.4 Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes 91
C.5 How to compute protein residue contacts more accurately? . . . 92

APÊNDICE D – PRÊMIOS 93
D.1 Best Poster Award X-Meeting 2016 na categoria “Software Development and Databases” 93

1 Introdução

1.1 Proteínas e aminoácidos

Proteínas são biomoléculas essenciais para existência da vida, pois desempenham papéis vitais nos organismos vivos, atuando no transporte, regulação de funções corporais, proteção imunológica, catálise em reações bioquímicas, entre outros processos biológicos importantes para manter organismos complexos, como o ser humano, até unicelulares, como bactérias (BERG, 2006). Proteínas são constituídas de unidades menores, chamadas de aminoácidos, e estes são compostos por: cadeia principal e cadeia lateral, conforme ilustrado na Figura 1, na qual temos os átomos de nitrogênio (azul), carbono (cinza escuro), oxigênio (vermelho) e hidrogênio (branco). A cadeia principal (cinza claro) é formada por uma amina e uma carboxila, enquanto a cadeia lateral (R, amarelo claro) varia conforme o tipo do aminoácido.

A cadeia principal é similar entre todos os aminoácidos e é composta, além da amina ($-NH_2$) e carboxila ($-COOH$), por um carbono central (denominado Cα) e um hidrogênio, enquanto a cadeia lateral (grupo R) está ligada ao Cα. Essa, por sua vez, é distinta em quantidade de átomos o que, consequentemente, define suas propriedades físico-químicas. Enquanto a cadeia lateral permite a distinção entre os aminoácidos, a cadeia principal tem a propriedade de uni-los. Essa união se dá por meio da ligação química entre o hidrogênio (H) do grupamento amina e hidróxido (OH) do grupamento carboxila, chamada de ligação peptídica, tendo como produto da reação uma molécula de água (H_2O), conforme representado na Figura 2.

Como há perda de átomos, e a liberação de uma molécula de água, os aminoácidos são chamados de resíduo quando formam as proteínas. Os resíduos se unem através de ligações peptídicas. Cada resíduo possui propriedades físico-químicas particulares definidas pela cadeia lateral e tais características permitem a formação de proteínas com as mais diversas conformações e funções.

1.2 Estrutura de proteínas

As funções biológicas desempenhadas pelas proteínas estão intrinsecamente relacionadas à sua conformação estrutural (BERG, 2006). Para adquiri-la, as cadeias polipeptídicas se enovelam (se dobram e se torcem), orientadas por interações intermoleculares, e assumem suas conformações nativas, podendo assumir sua função. Elucidar e entender como as proteínas se enovelam é conhecido como o "problema de dobramento de proteínas" e tem sido um grande desafio na biologia nos últimos 50 anos (DILL; MACCALLUM, 2012). Trabalhos recentes utilizando inteligência artificial conseguiram definir a estrutura de uma proteína, dada sua sequência,

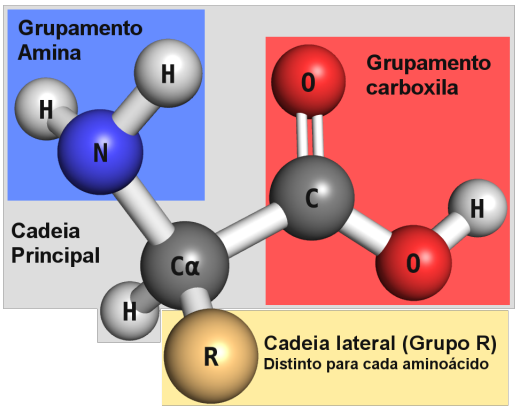


Figura 1 – Estrutura dos aminoácidos.

com aproximadamente 90 pontos de precisão em uma escala de 100 (SENIOR et al., 2020).

A estrutura de uma proteína pode ser analisada em quatro níveis hierárquicos: estrutura primária, secundária, terciária e quaternária. Esta abstração permite que as sequências e estruturas de proteínas possam ser representadas computacionalmente e melhor compreendidas, bem como seus mecanismos de funcionamento.

A estrutura primária consiste na sequência linear dos resíduos da cadeia polipeptídica, composta pela ordem em que os aminoácidos são unidos, começando do N-terminal (ou amino-terminal) em direção ao C-terminal (ou carboxi-terminal). A estrutura secundária apresenta traços do espaço tridimensional e é composta de conformações bem conhecidas que se repetem ao longo da cadeia polipeptídica: α-hélices, folhas-β e laços *loops*. A estrutura terciária é representada pela cadeia polipeptídica, formada pela combinação de estruturas secundárias em toda sua extensão estrutural. Por fim, temos a estrutura quaternária ou os complexos proteicos, compostos por mais de uma cadeia interagindo entre si. Essas interações são principalmente não-covalentes (com exceção de ligações dissulfeto) em suas interfaces. Apesar de serem consideradas

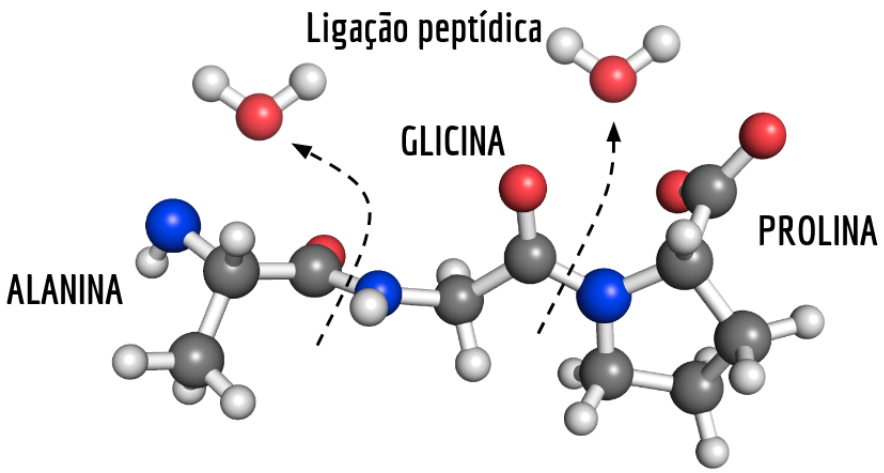


Figura 2 – Formação de uma ligação peptídica.

fracas, comparadas às interações covalentes, estabilizam o complexo, mantendo sua forma, e consequentemente permitindo executar sua função. Essas interações podem ocorrer também com moléculas de DNA (proteína-DNA), RNA (proteína-RNA), ligantes e pequenas moléculas (proteína-ligante) e entre outras proteínas (proteína-proteína e proteína-peptídeos).

1.3 Peptídeos e complexos proteína-peptídeo

A palavra "peptídeo" vem do grego *peptós* ("digerido") mais o sufixo *ídeo*. Assim como as proteínas, os peptídeos são formados por aminoácidos, mas diferem em tamanho. Os peptídeos são menores e podem ter 2 (dipeptídeo), 3 (tripeptídeo), 4 (tetrapeptídeo) e até 50 (oligopeptídeo) resíduos de aminoácidos (CRAIK et al., 2013). A definição do tamanho de um peptídeo é um tanto arbitrária. Observa-se na literatura definições distintas de tamanhos como 5-15 (LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010), 5-25 (JOHANSSON-ÅKHE; MIRABELLO; WALLNER, 2019), 5-30 (XU; ZOU, 2020), 5-35 (VANHEE et al., 2010), 5-50 (FRAPPIER; DURAN; KEATING, 2018), 35 (DAS et al., 2013) e 50 (WEN et al., 2019) aminoácidos.

Estima-se que entre 15 e 40% de todas interações em células são mediadas por peptídeos (NEDUVA et al., 2005; PETSALAKI; RUSSELL, 2008). Os peptídeos são importantes nos processos de sinalização celular e em redes regulatórias, agindo também como hormônios e antimicrobianos.

Conhecer e entender a estrutura de um complexo peptídeo-proteína é a chave para caracterizar e manipular peptídeos, uma vez que podem fornecer pistas fundamentais para o projeto de inibidores direcionados, por exemplo. Existem bases de dados que possuem sequências de peptídeos que interagem com proteínas (provenientes de genomas), como ELM (PUNTERVOLL et al., 2003), PROSITE (FALQUET et al., 2002) e SCANSITE (OBENAUER; CANTLEY; YAFFE, 2003). Porém, há pouca informação estrutural disponível sobre tais interações. O Protein Data Bank (PDB) (BERMAN et al., 2006) é o maior e mais conhecido repositório de estruturas tridimensionais de proteínas e possui atualmente mais de 169.000[1] macromoléculas biológicas, das quais a maior parte é de proteínas.

Como alvos terapêuticos, os peptídeos não possuem biodisponibilidade por via oral em vista da sua massa molecular (>500 Da), além de serem caros para produção em larga escala (OTVOS, 2008). Por isso, no final da década de 1980, as empresas farmacêuticas reduziram suas pesquisas em peptídeos e focaram em moléculas pequenas, que eram mais vantajosas em questões farmacocinéticas, além de serem mais estáveis *in vivo* (OTVOS, 2008). Porém, nos anos 2000, os peptídeos voltaram a serem alvos para fármacos (ANGELOVA et al., 2019). Isso ocorreu devido à inesperada toxicidade e reatividades cruzadas de moléculas pequenas, uma

[1] Disponível em: <www.rcsb.org>. Acesso em 26 Set/2020

vez que essas podem sofrer uma seletividade reduzida por seu tamanho, podendo causar efeitos colaterais (CRAIK et al., 2013).

Em geral, peptídios possuem baixa toxicidade e tendem a possuir alta especificidade com os seus alvos devido a maior superfície de contato, comparada as pequenas moléculas e, consequentemente, maior interação com o alvo. Com o recente surgimento de novas abordagens sintéticas que permitem alterações nas propriedades biofísicas e bioquímicas dos peptídeos, tais moléculas estão sendo novamente consideradas candidatas a medicamentos (ANGELOVA et al., 2019; LEE et al., 2019; VINOGRADOV; YIN; SUGA, 2019). De fato, mais de 60 medicamentos peptídicos foram aprovados nos principais mercados farmacêuticos e centenas de outros estão em desenvolvimento clínico ativo no momento (LAU; DUNN, 2018). Peptídeos como inibidores também são usados para tratar câncer, diabetes e doenças autoimunes e com altas taxas de sucesso para o desenvolvimento comercial (PANT et al., 2020). Vários candidatos a medicamentos de nova geração foram propostos como agentes terapêuticos para o diabetes mellitus tipo 2 (LAU; DUNN, 2018), como por exemplo, derivados da exenatida, uma forma sintética de um peptídeo natural de 39 aminoácidos secretado pelo *Heloderma suspectum*.

Compreender a estrutura e o reconhecimento de complexos proteína-peptídeo pode ajudar no desenho racional de novos peptídeos e compostos à base de peptídeos para o desenvolvimento de fármacos ou outros fins biotecnológicos. Consequentemente, bancos de dados de complexos proteína-peptídeo podem pavimentar o caminho para a análise e compreensão dos mecanismos de reconhecimento proteína-peptídeo. Existem bases de dados de peptídeos, com finalidades distintas, como peptídeos bioativos (WANG et al., 2018), antimicrobianos (WANG; LI; WANG, 2016), peptídeos de penetração celular (GAUTAM et al., 2012; AGRAWAL et al., 2016), peptídeos hemolíticos (GAUTAM et al., 2014). Porém, mesmo tendo disponíveis as estruturas dos peptídeos, em geral, se concentram em descrições e não o consideram como parte de um complexo. Tal propriedade é fundamental, pois o complexo proteína-peptídeo pode servir de modelo para uma proteína alvo para a qual se deseja encontrar um peptídeo com afinidade de ligação.

1.4 Bases de dados de proteína-peptídeo

Existem bases de dados específicas para estruturas de complexos proteína-peptídeo como PepX (VANHEE et al., 2010), PeptiDB (LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010), PepBind (DAS et al., 2013), PixelDB (FRAPPIER; DURAN; KEATING, 2018), PepBDB (WEN et al., 2019) e PepPro (XU; ZOU, 2020). Elas possuem diferentes peculiaridades, como por exemplo, definem peptídeos com tamanhos diferentes, filtram complexos a partir de propriedades específicas e dispõem os dados em plataformas diversas. A seguir, serão descritos as características principais das bases de dados mencionadas.

1.4.1 PepX

O PepX possui 1.431 complexos e agrupa um conjunto final de 505 complexos não redundantes. Para isso, os autores usam similaridade estrutural das interfaces, alinhando os complexos com o programa MUSTANG (KONAGURTHU et al., 2006) e agrupando-os utilizando um algoritmo de aglomeração hierárquica. Apesar do manuscrito do PepX descrever uma página *web* para acesso ao repositório, sua última atualização foi realizada em 2014 e não está mais disponível.

1.4.2 PeptiDB

PeptiDB se apresenta como um conjunto de dados com 103 complexos de alta-resolução. PeptiDB classifica os peptídeos com tamanhos de 5 a 15 aminoácidos. Além disso, somente estruturas com menos de 70% de identidade de sequência foram consideradas, como técnica para prevenir redundância. O PeptiDB não possui um serviço *web* para explorar as estruturas, mas os autores disponibilizam a lista com os códigos das estruturas no formato *protein data bank* (pdb), informando as cadeias da proteína e do peptídeo. Tanto PepX e PeptiDB filtram as estruturas coletados do PDB por uma resolução máxima, sendo 2,5 Å para o PepX e 2,0 Å para o PeptiDB.

1.4.3 PepBind

PepBind é um repositório de estruturas, sequências e observações experimentais com 3.100 complexos proteína-peptídeo. Nessa base de dados, peptídeos são definidos como estruturas de até 35 aminoácidos. Entretanto, a resolução dos complexos não é especificada. Os autores afirmam que o PepBind complementa o PepX, pois provém de maiores detalhes quanto a interface entre o peptídeo e o proteína, além de informações de suas funções celulares (coletados manualmente). PepBind possui ainda uma opção de busca por similaridade estrutural e de sequência. Porém, o serviço *web* é instável e até a última data de visita[2] encontra-se em manutenção.

1.4.4 Bases de dados recentes: PixelDB, PepBDB e PepPro

As mais recentes base de dados de complexos proteína-peptídeo como PixelDB (FRAPPIER; DURAN; KEATING, 2018), PepBDB (WEN et al., 2019) e PepPro (XU; ZOU, 2020), concentram-se em agregar e fornecer elementos estruturais com peptídeos de até 50 aminoácidos. O PixelDB possui 1.966 complexos não redundantes, organizados em grupos para fornecer dados estruturais e conservação dos modos de ligação com os peptídeos. Enquanto isso, o PepBDB agrega 12.241 complexos de peptídeos-proteínas com informações de interação úteis para analisar e comparar os programas de ancoragem molecular (*docking*). Por fim, PepPro é um conjunto de dados de referência para *benchmarks* de algoritmos de ancoragem molecular entre proteínas e

[2] Disponível em: <http://pepbind.bicpu.edu.in/>. Acesso em 26 Setembro de 2020

peptídeos. Ele contém 89 complexos não redundantes, recuperadas de 1.198 arquivos PDB de alta resolução com peptídeos que variam de 5 a 30 aminoácidos.

1.5 Justificativa

Como mencionado previamente, existem várias bases de dados de estruturas de complexos proteína-peptídeo, com diversas características incluindo remoção de redundância, alta resolução, conservação de sítio de ligação, dentre outras. Porém, algumas dessas bases de dados estão desatualizados ou estão indisponíveis. Algumas das bases de dados mais recentes estão acessíveis e fornecem recursos para a análise do modo de ligação e para facilitar estudos baseados em estruturas. No entanto, cada banco de dados tem seu próprio objetivo e usa um conjunto específico de complexos peptídeo-proteína. Portanto, importantes anotações de peptídeos estão espalhadas por vários bancos de dados. Além disso, o agrupamento de todas as informações com propósito mais geral em um único repositório pode ser importante para estudos de biologia estrutural e computacional que ajudem no entendimento, que modelem e prevejam novos peptídeos ou façam a evolução dirigida dos mesmos Wu et al. (2020b). Por fim, a criação de repositório de complexos proteína-peptídeo, aliada a um serviço *web* flexível e de fácil uso para análise, exploração e visualização de uma base de dados abrangente e de propósito geral sobre estrutura e interação de peptídeos com proteínas tem grande relevância e utilidade.

O presente trabalho foi inspirado em projetos anteriores sobre a defesa de plantas contra insetos e patógenos. A soja (*Glycine max*) é um dos principais produtos do agronegócio brasileiro, sendo o Brasil o segundo maior produtor de soja no cenário mundial. Uma das principais pragas que assolam a produção de soja no Brasil é a lagarta-da-soja (*Anticarsia gemmatalis* Hübner). Mesmo sendo uma praga desfolhadora, a lagarta-da-soja pode causar a destruição completa das plantas, trazendo grandes prejuízos às lavouras. A soja, quando devorada pela lagarta-da-soja, produz os inibidores de protease Kunitz e Bowman-Birk, responsáveis por prejudicar a degradação da protease no intestino do inseto (PILON et al., 2017; PATARROYO-VARGAS et al., 2017). Consequentemente, a ingestão de tais inibidores leva à uma deficiência proteica que prejudica o crescimento, desenvolvimento e a reprodução da lagarta-da-soja, em vista da indisponibilidade de aminoácidos para a síntese de proteínas fundamentais nesses processos ((BROADWAY, 1995; AO et al., 2013)). O uso de peptídeos como controle alternativo de pragas reduz o uso de defensivos agrícolas de alta toxicidade, garantindo mais segurança alimentar e preservação ambiental. Com isso, há um interesse em propor moléculas peptídicas para inibir as proteases do intestino da lagarta, com potencial para serem utilizadas no controle ecológico desse inseto-praga. Um repositório de complexos proteína-peptídeo pode elucidar e propor moléculas peptídicas promissoras para inibição das proteases da lagarta-da-soja através da busca por proteínas similares à protease alvo.

Além disto, no último ano do desenvolvimento deste trabalho, uma desconhecida doença

acometeu moradores da província de Wuhan, na China. Ela espalhou-se pelo mundo rapidamente, e pesquisadores vêm gerando dados de diversas naturezas para elucidar as origens do patógeno, forma de contágio, tratamentos e vacinas. Em janeiro de 2020, descobriram tratar-se de um novo coronavírus (SARS-CoV-2), tipo de vírus cujo genoma é formado por RNA de fita simples, e infecta mamíferos, aves e seres humanos. Não se sabe ao certo as origens deste vírus. É possível que outros vírus transponham a barreira de espécies e infectem humanos. Sete outros coronavírus foram descritos em humanos, entre eles o SARS-CoV (causador de síndrome respiratória aguda severa - SARS) e MERS-COV (síndrome respiratória do Oriente Médio). A nova doença foi denominada COVID-19, e seus principais sintomas são febre, tosse e falta de ar, podendo evoluir para pneumonia, SARS e insuficiência renal. Não existem medicamentos que comprovadamente tratem a COVID-19; apenas tratamentos de suporte aos sintomas são utilizados. Na perspectiva molecular, a IA, integrada a métodos de bioinformática e simulações moleculares, bem como validações experimentais in vitro, pode auxiliar na compreensão da função de proteínas do patógeno, identificar medicamentos para reposicionamento, propor novos compostos para o desenvolvimento de fármacos, identificar alvos para vacina, melhorar o diagnóstico e aumentar o entendimento do vírus, sua infecciosidade e gravidade. Neste contexto, o desenho de peptídeos que possam interagir com proteínas alvo do SARS-CoV-2 também é uma área de pesquisa ativa. Nesse trabalho, usamos o Propedia para prospectar peptídeos que possam ser promissores para inibição de sítios de alvos do vírus (trabalhos em andamento) ou possam ser base para evolução dirigida.

2 Objetivos

2.1 Objetivo Geral

Construir uma base de dados de estruturas de complexos proteína-peptídeo abrangente, atualizada, alta disponibilidade, resposta ágil e com interface amigável capaz de agrupar os complexos homólogos considerando diferentes aspectos. Essa base deve prover mecanismos de redução de redundância permitir a realização de buscas por sequência ou estrutura, a partir alvos de interesse.

2.2 Objetivo Específicos

- Modelar um banco de dados relacional para armazenar as informações extraídas de estruturas de proteínas, tornando ágil a recuperação do dados, em vista do grande volume dados.
- Desenvolver códigos para coleta de dados, processamento e povoamento da base dados.
- Criar modelos e algoritmos capazes de agrupar complexos proteína-peptídeo homólogos, com intuito de remover redundância, considerando três aspectos distintos: sequência do peptídeo; estrutura de interface do complexo; e sítio de ligação.
- Produzir um serviço *web* com interface amigável e intuitiva, com funcionalidades para exploração e visualização dos complexos.
- Gerar mecanismos de busca por complexos homólogos, baseado sequência (tanto da proteína quanto do peptídeo) e sítio de ligação.

3 Metodologia

Neste capítulo serão abordadas as etapas, ferramentas e métodos utilizados para modelagem e povoamento da base de dados de complexos proteína-peptídeo. Serão descritos também os detalhes para o desenvolvimento do serviço *web*. Através dele, o usuário poderá realizar buscas por similaridade de sequência ou sítio de ligação, uma vez fornecidas a sequência ou estrutura alvo, respectivamente. A primeira parte da metodologia consiste na modelagem da base de dados e seu povoamento inicial. Para isso, foi realizada uma coleta de todas as estruturas de proteínas do PDB. Em seguida, *scripts* em Python foram desenvolvidos para extrair informações relevantes de cada arquivo, identificando complexos proteína-peptídeo e coletando suas características. Um sistema de agrupamento de complexos foi criado, levando em conta três aspectos distintos de similaridade: sequência do peptídeo, estrutura de interface do complexo e sítio de ligação. Vale ressaltar que estas duas últimas etapas foram as mais importantes e complexas tendo em vista que eles realizam cálculos que exigem alto poder de processamento, como alinhamento estrutural de proteínas em larga escala. As informações foram armazenadas na base de dados, e por fim, foi criado um serviço (web) para exploração interativa dos complexos proteína-peptídeo. As seções a seguir descrevem essas etapas em detalhes.

3.1 Modelagem da base de dados

Antes de iniciar a etapa de coleta e análise de cada arquivo PDB, em busca de complexos proteína-peptídeo, modelamos a base de dados. Utilizamos o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL (http://www.mysql.com). O MySQL é um SGBD gratuito, robusto e amplamente utilizado pela comunidade científica.

A Figura 3 apresenta o modelo final. A tabela de complexos (*complex*), em branco, é a principal entidade da base de dados e interliga todas as outras tabela. Nela foram armazenados os dados pertinentes ao complexo, como o peptídeo e proteína que o compõem, área e quantidade de resíduos da interface, entre outros. Entidades relacionadas aos metadados de cada PDB estão representadas em azul (tabelas *pdb*, *pdb_groups* e *groups*), sendo a tabela *groups* responsável por rotular cada pdb com uma ou mais classificações. Tomando como exemplo a estrutura da protease do coronavírus 3CLpro (PDB id: 1p9u), a mesma é classificada no PDB como uma *"Hydrolase/Hydrolase Inhibitor"* e com isso, na base de dados esta estrutura seria reclassificada com dois grupos *"Hydrolase"* e *"Inhibitor"*. Em amarelo estão as tabelas de peptídeos (*peptide*), proteínas (*receptor*) e organismos (*organism*). Vale esclarecer que o termo "receptor" foi aqui utilizado para definir as cadeias de proteínas que interagem com um peptídeo, e que doravante serão chamadas de receptores. As tabelas em verde definem os grupos de complexos, definidos para cada tipo de agrupamento: por sequência de peptídeo (*cluster_sequence*);

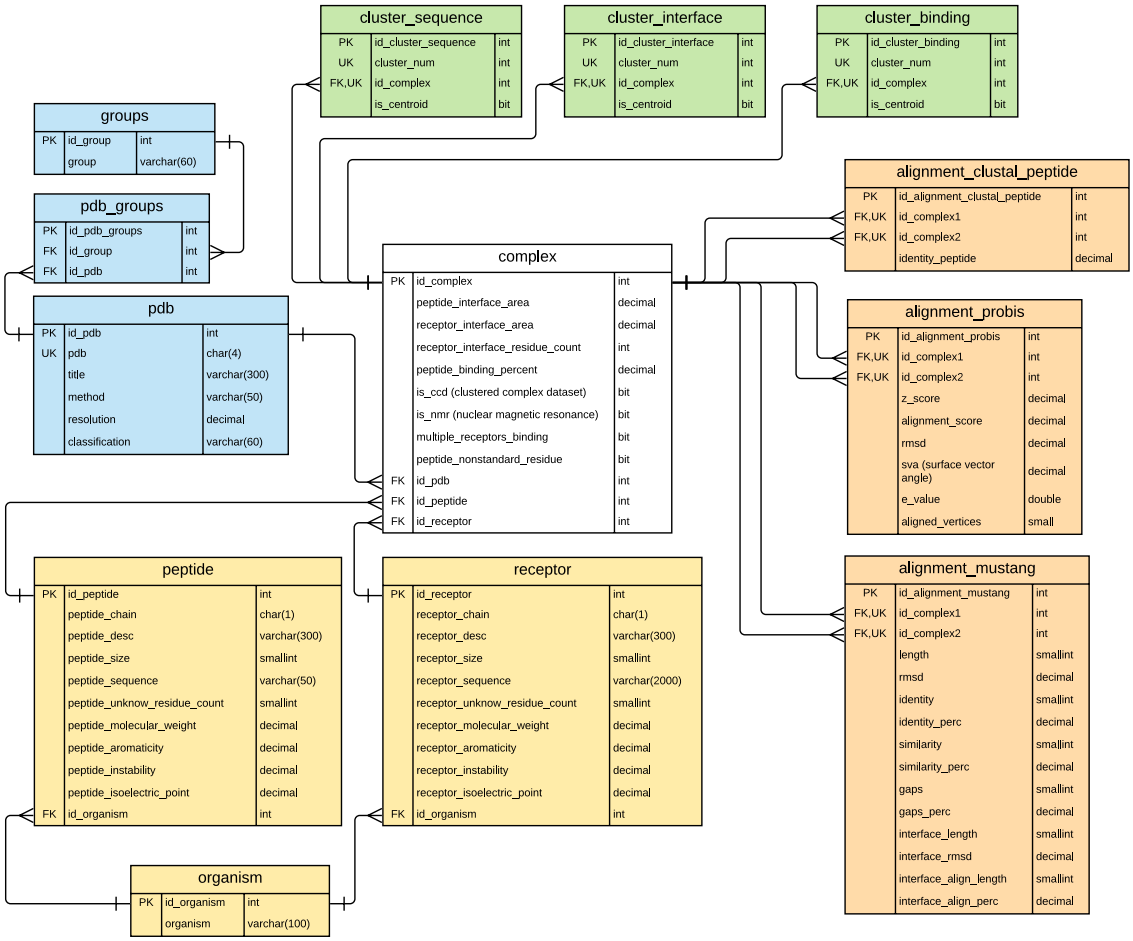


Figura 3 – Modelo da base de dados do Propedia

por estrutura de interface (*cluster_interface*); e por de sítio de ligação (*cluster_binding*). Por fim, em laranja, estão as tabelas que armazenarão os resultados dos alinhamentos de sequência (*alignment_clustal_peptide*), de estrutura de interface (*alignment_mustang*) e de sítio de ligação (*alignment_probis*), realizados entre os complexos e utilizados para definir os agrupamentos.

3.2 Coleta e análise dos dados de entrada

Para encontrar possíveis complexos proteína-peptídeo, todo os arquivos de estrutura do PDB foram recuperados e armazenados localmente, conforme apresentado na Figura 4 A.

O *script rsync_PDB.sh* (arquivo de execução de comandos de sistema operacional) utiliza o *rsync* (TRIDGELL; MACKERRAS et al., 1996), um programa utilitário nativo de sistemas operacionais baseados em Unix, como Linux e Mac OS, que sincroniza dois diretórios distintos, sejam eles pertencentes ao mesmo computador ou não. O PDB possui um serviço de protocolo de transferência de arquivos (do inglês: *File Transfer Protocol*, FTP), que utilizando o *rsync*, permite sincronizar o diretório de arquivos de estrutura do PDB com um diretório

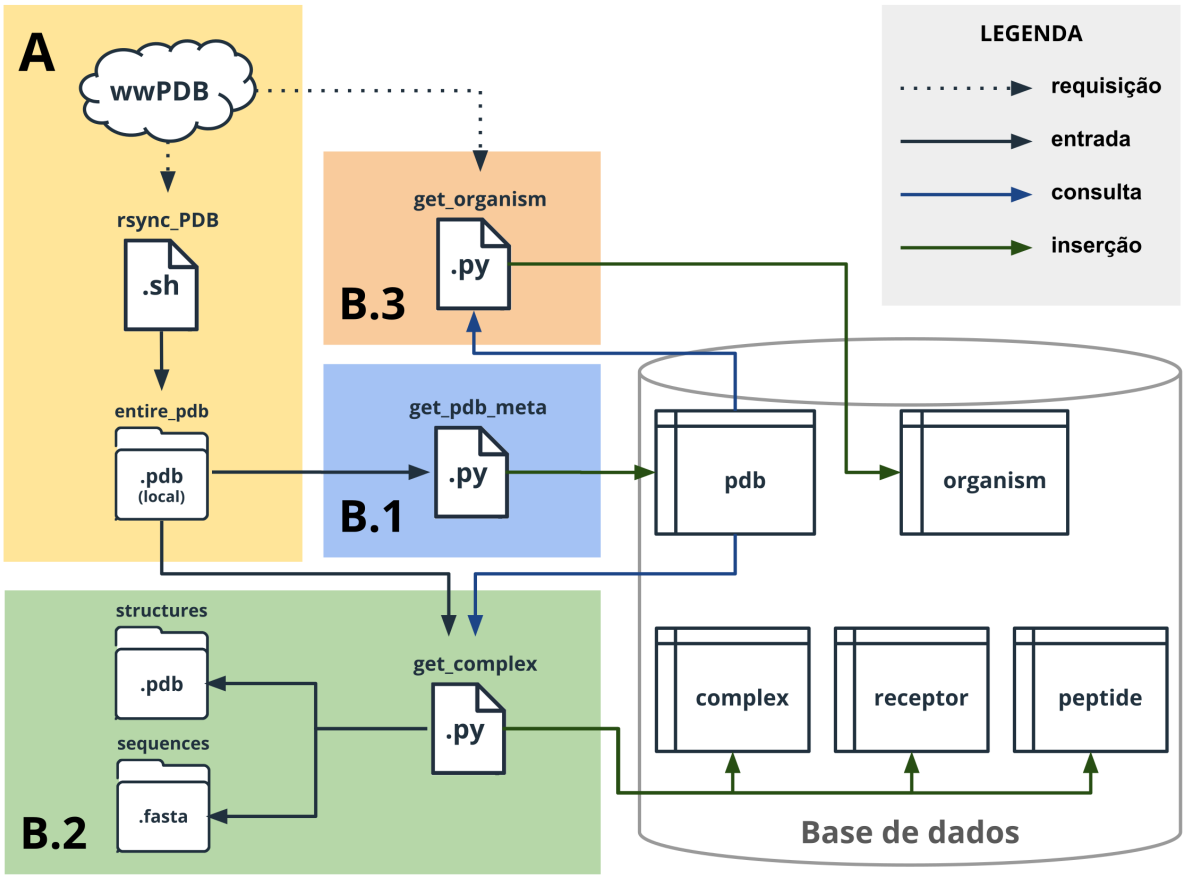


Figura 4 – Fluxo da coleta e análise de dados para povoamento inicial da base de dados.

local. Com isso, 164.281[1] arquivos PDB foram recuperados, totalizando cerca de 33.7 GB de espaço de armazenamento. Apesar de ser um tarefa que exige banda de dados e espaço em disco, este procedimento garante que todo o escopo atual de estruturas de proteínas seja analisado. Além disso, a utilização do *rsync* será útil para futuras atualizações, uma vez que não será mais necessário realizar o *download* de todos os arquivos novamente, pois o mesmo garante que o diretório local de estruturas permaneça idêntico ao existente no PDB, adicionando somente as novas aquisições e removendo estruturas consideradas obsoletas pelo PDB. De posse dos arquivos de estrutura (dados de entrada) e da base de dados pronta para ser povoada, iniciou-se a etapa de extração de complexos e suas características, seguida pela etapa de inserções de registros na base de dados. Para essa fase, *scripts* em Python foram desenvolvidos para analisar todos os arquivos de estrutura recuperados do PDB (Figura 4 B). A critério de definição, um "peptídeo" será uma cadeia de proteína que possua tamanho entre 2 a 50 aminoácidos. O argumento utilizado para esta escolha está relacionado aos limites que as bases de dados de complexos proteína-peptídeo usam (Tabela 2).

O *script get_pdb_meta.py* (Figura 4 B.1) busca pelos metadados de cada arquivo PDB

[1] Disponível em: <www.rcsb.org>. Acesso em 02 Mai/2020

contidos no diretório local que tenham potencial para conter complexos de proteína-peptídeo. Para isso, os seguintes requisitos são avaliados: conter duas ou mais cadeias polipeptídicas (para encontrar complexos); ao menos uma cadeia de 2 até 50 aminoácidos (para identificar possíveis peptídeos); estruturas resolvidas por difração de raio X, com resoluções de 0 a 2.5 Å; estruturas resolvidas por ressonância magnética nuclear (*nuclear magnetic resonance spectroscopy* - NMR). Ao fim da busca, ao todo 8.133 arquivos PDB foram encontrados com tais características, e seus metadados (título, método experimental, resolução, classificação, quantidade de cadeias de peptídeos e receptores) foram inseridos na base de dados (tabela pdb). Para encontrar os complexos, o *script get_complex.py* (Figura 4 B.2) foi desenvolvido para filtrar elementos estruturais, identificar os complexos proteína-peptídeo e por fim armazenar as informações obtidas na base de dados. Ele utiliza como escopo somente os arquivos de estrutura que foram registrados na tabela pdb. Para analisar arquivos pdb foi utilizado a biblioteca Biopython (COCK et al., 2009). Com ela é possível transformar o conteúdo textual do arquivo em um objeto contendo as cadeias polipeptídicas, resíduos de aminoácidos e átomos, em uma estrutura hierárquica. Para cada arquivo PDB, foi aplicado um filtro, removendo moléculas de água, átomos de hidrogênio, posições atômicas alternativas (HAMELRYCK; MANDERICK, 2003) e artefatos cristalográficos. Artefatos cristalográficos são elementos estruturais utilizados no processo de resolução das estruturas que podem estar presentes nos arquivos PDB. Para definir tais elementos, foi utilizado a lista desenvolvida por Fassio et al. (2019).

Após a filtragem, iniciou-se o processo para identificar complexos proteína-peptídeo. Para cada arquivo PDB filtrado, uma busca é realizada por todas as suas cadeias polipeptídicas. Se a cadeia possuir entre 2 e 50 aminoácidos, a mesma é classificada como um "peptídeo". Caso a cadeia possua 60 ou mais aminoácidos será classificada como "receptor". Este intervalo de 10 aminoácidos entre peptídeos e receptores foi criado arbitrariamente para evitar complexos aos quais o peptídeo e o receptor possuam tamanhos próximos. Para identificação de complexo, verificou-se se havia uma interface de interação entre os peptídeos e os receptores encontrados. Se existe ao menos um resíduo do peptídeo a uma distância de 6 Å de ao menos um resíduo de uma proteína, então existe um interação entre eles (BICKERTON; HIGUERUELO; BLUNDELL, 2011; PLAXCO; SIMONS; BAKER, 1998). Logo, todos os resíduos até essa distância são identificados como parte da interface. A Figura 5 ilustra um exemplo dessas ocorrência. Os resíduos da Cadeia A (azul escuro) que estão a 6 Å de distância do peptídeo (vermelho), compõem, juntamente com o peptídeo, a interface proteína-peptídeo, formando assim um complexo. Observa-se que, para este caso, somente a Cadeia A (azul claro) possui interação com peptídeo, enquanto a Cadeia B (amarelo) interage apenas com a Cadeia A.

Algumas vezes foram encontrados complexos aos quais um peptídeo interagia com mais de uma proteína, conforme ilustrado na Figura 6. Neste exemplo, é apresentado um peptídeo (vermelho) interagindo com três cadeias de proteínas distintas (azul, laranja e amarelo), onde os bastões (*sticks*) são os resíduos dos receptores da interface proteína-peptídeo. Para diferenciar os complexos ao qual um peptídeo interage com múltiplos receptores e dos complexos onde o

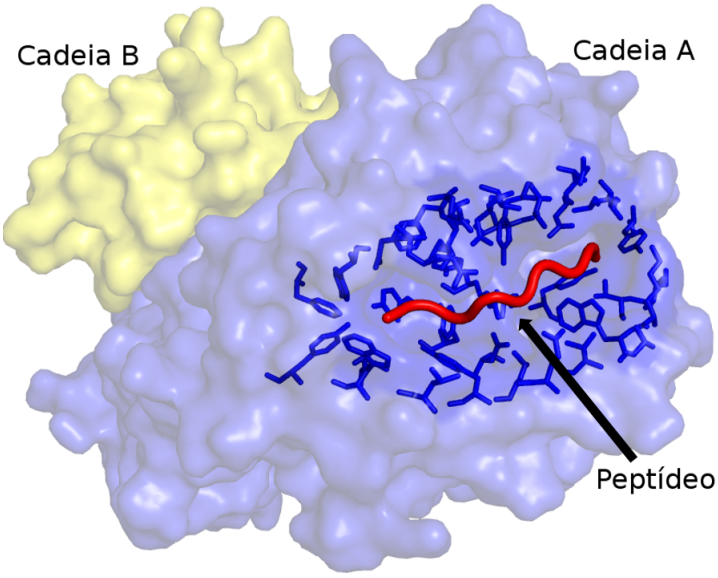


Figura 5 – Exemplo de um complexo proteína-peptídeo. Imagem geradas com PyMOL (http://pymol.org). Fonte: próprio autor.

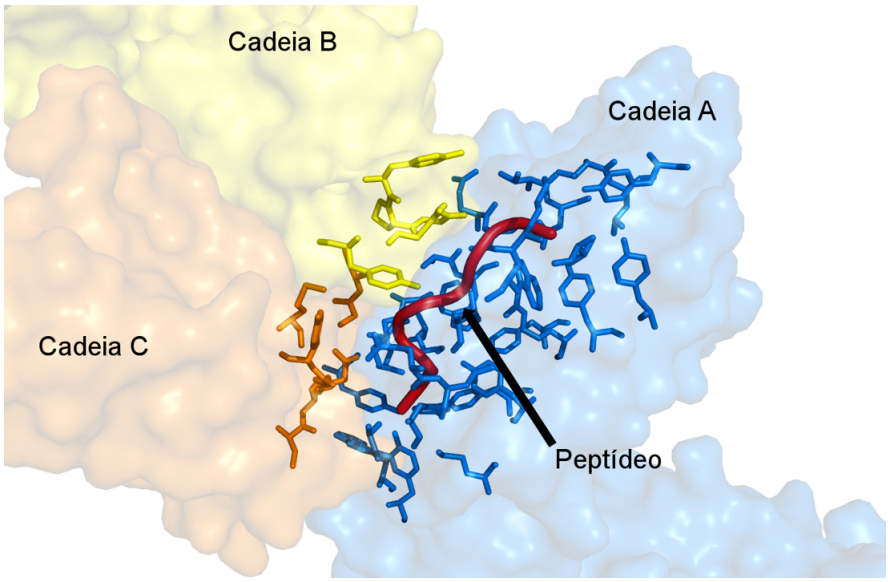


Figura 6 – Exemplo de um complexo proteína-peptídeo ao qual o peptídeo interage com mais de uma proteína (receptor). Imagem gerada com PyMOL (http://pymol.org). Fonte: próprio autor.

peptídeo interage apenas com um, foi feito o cálculo do percentual de resíduos das interfaces dos receptores que interagiam com o peptídeo. Assim sendo, para a Figura 5 (PDB id: 1a1m), o receptor da cadeia A interage com uma interface de 41 resíduos (100%) com o peptídeo, enquanto que para na Figura 6 (PDB id: 1bd2) observa-se três proteínas, A, B e C, interagindo com interfaces de 43 (71,67%), 10 (16,67%) e 7 (11,67%) resíduos, respectivamente.

Para cada complexo, foi coletado o tamanho, sequência, quantidade de resíduos não padrão, área da interface (AI). Essas características foram extraídas tanto do receptor como do

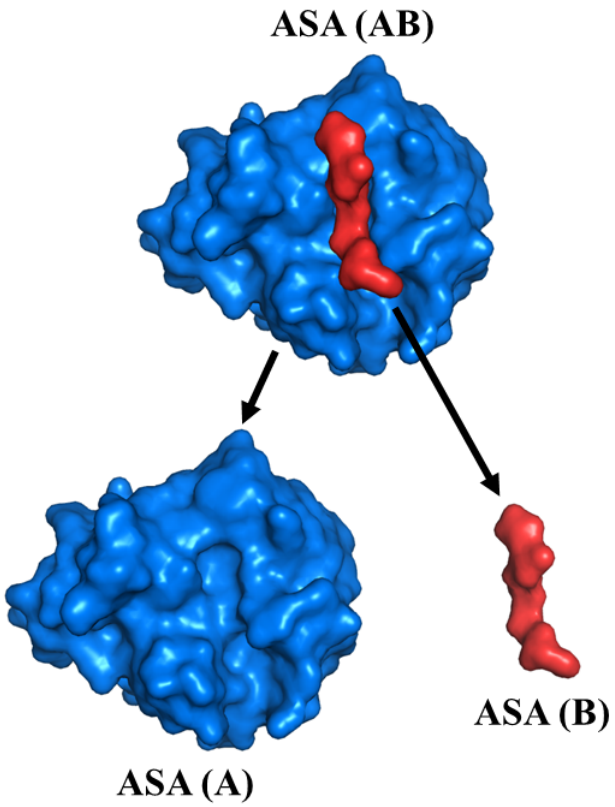


Figura 7 – Cálculo da área da interface (AI). Imagem gerada com PyMOL (http://pymol.org). Fonte: próprio autor.

peptídeo. Além do critério de distância de 6 Å entre um peptídeo e um receptor, como critério de definição de complexo, a AI também foi considerada. Logo, se a AI for 0.0, então o complexo é desconsiderado.

Para computar a AI do receptor e do peptídeo, foi utilizado o método proposto por Lee e Richards (1971), que calcula a área de acessibilidade ao solvente (*Accessible Surface Area, ASA*) de uma proteína em $Å^2$. O programa NACCESS (HUBBARD; THORNTON, 1993) realiza este cálculo a partir de um arquivo PDB e retorna o ASA de cada átomo encontrado. Conforme pode ser visto na Figura 7, para definir o AI da proteína e do peptídeo, computa-se o ASA do complexo (AB) e depois o ASA da proteína (A) e peptídeo (B) separadamente e, por fim, identifica-se o conjunto de átomos que ganharam acessibilidade ao solvente quando verificada a diferença de ASA, entre cada átomo, conforme equação:

$$AI = (ASA(A) + ASA(B)) - ASA(AB)$$

O último *script* a ser executado, *get_organism.py* (Figura 4 B.3), tem como objetivo realizar requisições no PDB, de acordo com o escopo da tabela pdb, para buscar pelos organismos das cadeias de peptídeos e receptores, permitindo enriquecer o filtro de busca que será utilizado no serviço *web*.

Todas as informações geradas foram inseridas em arquivos texto, no formato sql, que

Qtde. de proteínas interagindo com o peptídeo	Número de complexos	Números de complexos com peptídeos apenas com aminoácidos padrões
1	8.990	5.971
2	7.040	4.232
3	2.205	1.449
4	1.204	656
5	290	50
6	84	84
Total	**19.813**	**12.442**

Tabela 1 – Resumo da quantidade de complexos identificados, por quantidade de cadeias de receptores interagindo com peptídeos e peptídeos apenas com aminácidos padrões.

foram usados para povoar as respectivas tabelas da base, conforme visto na Figura 4. O principal motivo para criar arquivos como intermediadores para povoar a base, ao invés de realizar operações diretas usando *scripts*, visa não onerar a base de dados com inserções únicas. Além disso, vale ressaltar que a etapa descrita demanda grande poder computacional, uma vez que realiza diversos cálculos. Assim, para reduzir o tempo de processamento foram aplicados técnicas de processamento paralelo no código para que o mesmo use o máximo de núcleos do(s) processador(es) disponíveis no computador.

Ao todo 19.813 complexos proteína-peptídeo foram encontrados e suas estruturas (arquivos pdb) foram armazenadas no diretório *structures*. Tais estruturas contemplam os complexos, receptores, peptídeos e interface (resíduos da interface proteína-peptídeo). Além disso, arquivos de sequências (complexos, receptores e peptídeos) no formato fasta, também foram criados e inseridos no diretório *sequences*. Esses arquivos foram utilizados nas próximas etapas da metodologia e também serão disponibilizados para *download* no serviço *web*.

Por fim, um resumo quantitativo dos complexos obtidos pode ser observado na Tabela 1. Nela, estão disponíveis a quantidade de complexos com a qual um peptídeo interage. Além disso, foi discriminada a quantidade de complexos com a qual o peptídeo interagem e que são constituídos apenas de aminoácidos padrão, ou seja, apenas os 20 aminoácidos comumente encontrados nos seres vivos.

3.3 Sistema de agrupamento

Nesta seção serão apresentados os detalhes do desenvolvimento do sistema de agrupamento. O sistema considera três aspectos distintos de similaridade: identidade de sequência dos peptídeos, estrutura de interface do complexo e sítio de ligação do complexo. Um agrupamento visa unir elementos que possam ser semelhantes entre si, dado um critério. O método de agrupamento pode ser útil tanto para remoção de redundância, uma vez que pode reduzir um grupo a um único elemento representativo como também permite detectar um grande conjunto de dados de elementos que sejam similares. Quando se leva em consideração as interações proteína-peptídeo,

um agrupamento pode ser capaz de criar um conjunto de de dados de complexos que contenham sequências ou estruturas distintas, ou seja, um conjunto com alta diversidade. Também é possível criar um conjunto que contenham complexos semelhantes, verificando assim as muitas conformidades que o mesmo possa assumir.

Para criar o sistema foram desenvolvidos três módulos em Python para controlar o fluxo dados de forma sistêmica. Os *scripts* foram executados manualmente e utilizaram programas capazes de executar procedimentos de alinhamento de sequência (Clustal Omega (SIEVERS et al., 2011)) e de estrutura (MUSTANG (KONAGURTHU et al., 2006)), agrupamento de peptídeo por sequência (Hammock (KREJCI et al., 2016)) e detecção de sítio de ligação similares (ProBiS (KONC; JANEŽIČ, 2010)). Essa fase exigiu um grande poder de processamento, uma vez que realiza vários cálculos em um grande número de iterações, e por isso, esses programas foram escolhidos para permitir que a execução fosse realizada em larga escala e em um intervalo de tempo viável.

Nem todos os 19.813 complexos puderam fazer parte do sistema de agrupamento. Por exemplo, as ferramentas Hammock, MUSTANG e ProBiS não respondem bem a estruturas com aminoácidos diferentes dos 20 padrão. Além disso, um peptídeo que interage com múltiplos receptores pode causar alterações na conformação estrutural da interface proteína-peptídeo e não seria apropriado compará-lo com complexos com apenas um receptor. Por isso, optamos por criar um recorte para o Propedia: o peptídeo é constituído apenas por aminoácidos padrão e o mesmo interage apenas com um único receptor. Esse escopo de 5.971 complexos será chamado doravante de conjunto do complexos agrupáveis (CCA). Eles serão utilizados no sistemas de agrupamentos a seguir.

3.3.1 Sequência de peptídeo

O agrupamento de sequências de peptídeos (ASP) visa criar grupos de complexos que possuam peptídeos que sejam similares, levando em consideração sua sequência de aminoácidos. Para criar o ASP, utilizamos o programa Hammock (versão 1.2.0) (KREJCI et al., 2016). O Hammock usa cadeias ocultas de Markov (*Hidden Markov Models - HMM*) (NOGUCHI et al., 2002) para realizar esse agrupamento. Além disso, três programas externos são usados pelo Hammock para realizar alinhamento múltiplo, busca por similaridade e comparação HMM-HMM, sendo eles Clustal Omega (SIEVERS et al., 2011), HMMER 3.0 (FINN; CLEMENTS; EDDY, 2011) e HHSuite (SÖDING, 2005), respectivamente.

A Figura 8 ilustra o fluxo de dados feito para criar o agrupamento. O processo é divido em duas fases: criação do grupos de sequências de peptídeos utilizando o Hammock e refinamento para definição da identidade de sequência entre os peptídeos do mesmo grupo utilizando Clustal Omega (SIEVERS et al., 2011).

Na primeira fase (Figura 8 A), todas as sequências dos peptídeos que compõem o CCA

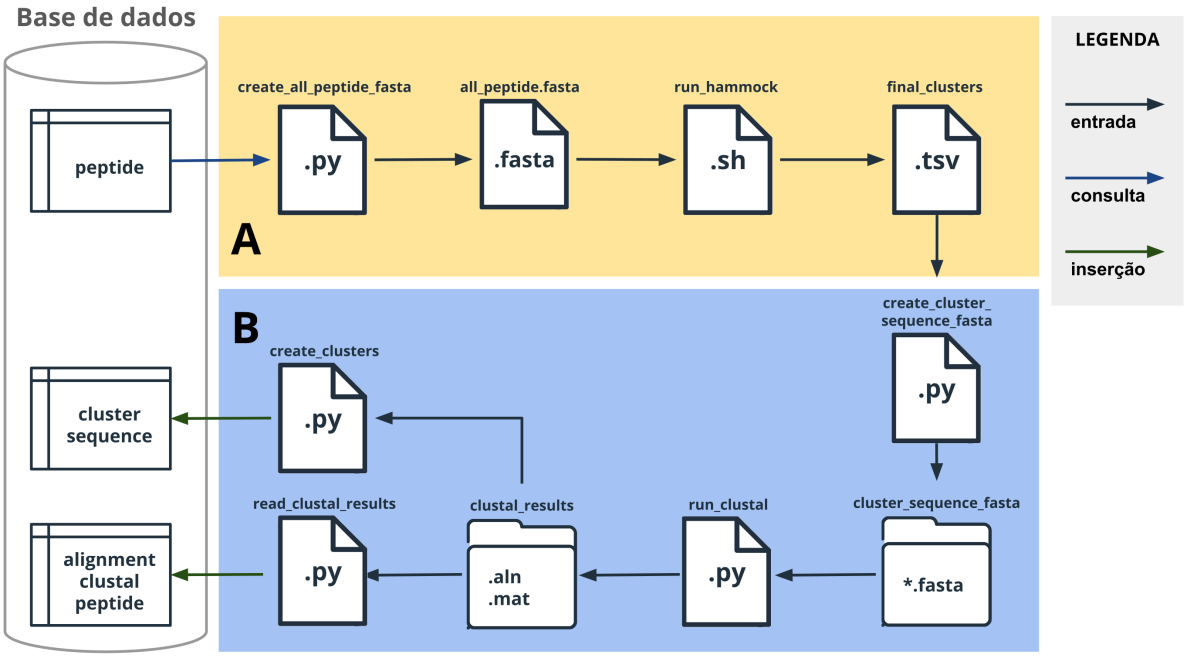


Figura 8 – Fluxo do processo de agrupamento de sequências de peptídeos.

foram recuperadas da base de dados pelo *script create_all_peptide_fasta.py*. Um arquivo único no formato fasta é então criado com todas as sequências (*all_peptide.fasta*), para ser usado como argumento de entrada para Hammock. O *script* também cria um arquivo de execução do Hammock (*run_hammock.sh*), contendo todos os parâmetros necessários para sua execução. Um desses parâmetros é a lista de rótulos de cada sequência de peptídeo. Cada rótulo possui o seguinte formato: identificador do complexo na base de dados; identificador PDB, letra da cadeia da proteína; e letra da cadeia do peptídeo, todos separados por "_", exemplo: 23_1a1m_A_C. Esse formato foi concebido para conter todas as informações que facilitam a identificação do complexo nas próximas etapas do ASP. Além disso, o parâmetro com a lista de rótulos é opcional no Hammock, mas que é fundamental quando se trata de um conjunto grande de sequências. Por isso, é justificável a criação desse arquivo de execução de forma automática, uma vez que seria inviável inserir manualmente essa lista. Por fim, após a execução do Hammock, um diretório de saída é criado, contendo informações de todo o processo de agrupamento, mas o resultado final do agrupamento se concentra no arquivo *final_clusters.tsv*. Nele, pode-se extrair os grupos finais, contendo em cada um a sequência principal (sequência mais frequente) e seus peptídeos.

Apesar de ser possível realizar os agrupamentos utilizando o Hammock, essa ferramenta não retorna um valor que possa ser usado para indicar o quanto as sequências de um grupo são similares. Uma métrica útil seria identidade de sequência entre a sequência principal e as sequências dos peptídeos de cada grupo. A identidade seria então uma unidade de distância e, apesar do Clustal Omega ser executado pelo Hammock, ele não é parametrizado para retornar a matriz de distância entre as sequências de um grupo. Sendo assim, foram desenvolvidos

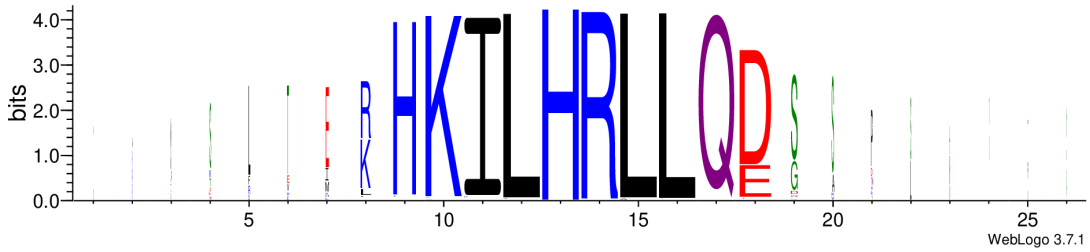


Figura 9 – Exemplo de logo de sequência do ASP.

scripts para realizar alinhamentos múltiplos de sequências para cada grupo e por fim, inserir os resultados do agrupamento na base.

A segunda fase (Figura 8 B) começa com o *script create_cluster_sequence_fasta.py*, responsável por criar arquivos de sequências de peptídeos (formato fasta) para cada grupo a partir do resultado de agrupamento do Hammock (*final_clusters.tsv*). Para cada grupo de sequências de peptídeos, o Clustal Omega é então executado para realizar o alinhamento múltiplo de sequência (*run_clustal.py*), retornando o resultado do alinhamento (arquivos .aln) e matrizes de distâncias (arquivos .mat) entre todas as sequências. A unidade de distância da matriz é a identidade entre as sequências do grupo. Cada arquivo de alinhamento (.aln) contém os complexos de um grupo, com isso, o *script create_clusters.py* identifica os grupos e centroides (sequências iguais a sequência principal do grupo, definida pelo Hammock) e os inserem na tabela *cluster_sequence*. Por fim, os resultados do Clustal Omega são interpretados (*read_clustal_results.py*) para armazenar as identidades de sequência de cada par de complexos do mesmo grupo na tabela *alignment_clustal_peptide*.

Para enriquecer as informações apresentadas, foram criados logos de sequências (CROOKS et al., 2004) para cada grupo, apresentando a sequência de resíduos conservados. A Figura 9 apresenta um exemplo de logo de sequência do maior grupo do ASP, contendo 412 sequências ao qual é possível observar a sequência maior consenso: HKILHRLLQD.

Essa fase resultou em 1.845 grupos de sequências de peptídeos, ao qual 1.074 são unitários e 771 grupos possuem mais de um elemento (não unitários).

3.3.2 Estrutura de interface

O agrupamento de estrutura de interface (AEI) foi desenvolvido para criar grupos de complexos que possuem similaridade quanto a estrutura da interface proteína-peptídeo. Para desenvolver o AEI, foi utilizado o programa MUSTANG (KONAGURTHU et al., 2006), para realizar o alinhamento estrutural par-a-par entre os complexos. O MUSTANG é um alinhador estrutural múltiplo de código-aberto (*open-source*) desenvolvido em C++ que sobrepõe estruturas a partir das posições dos Cα dos aminoácidos.

3.3.2.1 Versões iniciais

Nas primeiras versões do AEI. o alinhamento estrutural par-a-par foi feito de forma integral, considerando todas as possíveis combinações entre os pares de complexos. Anteriormente o número de complexos era inferior comparado ao CCA, totalizando 4.141. O custo computacional para realizar essa operação é muito elevado, e estratégias de processamento paralelo e de retomada (em caso de queda de energia) foram necessárias. Ao todo, 8.571.870 ($n * (n-1)/2$, onde n é o número de complexos) alinhamentos foram realizados utilizando o MUSTANG. Esse número é desafiador uma vez que se trata de um processo computacional custoso. Esse procedimento foi realizado em servidor Linux, distribuição Ubuntu Server 14.04 com 32GB de RAM e processador Intel 2.1GHz com 32 núcleos. Ao todo o processamento levou cerca de 3 meses para ser concluído.

O MUSTANG retorna seus resultados em arquivos de texto. Informações quanto ao alinhamento de sequência (a partir do alinhamento de estrutura), identidade, similaridade e *gaps*, por exemplo, são escritos no arquivo de saída *results.html* (nome padrão). Já o resultado do RMSD e matrizes de rotação são armazenados em *results.rms_rot*. Para que fosse possível realizar mais de 8 milhões de alinhamentos em um tempo hábil, foram feitas alterações no código-fonte do MUSTANG. Essas alterações visaram mudar o formato de saída dos resultados, para que valores de RSMD, identidade, similaridade, gaps, etc, fossem retornados em uma única saída.

Após a modificação e execução do MUSTANG (daqui pra frente denominado MUSTANGmod), o resultado é impresso na tela de forma sucinta e direta, conforme Figura 10. Linhas numeradas foram adicionadas a figura para facilitar sua explicação. Na linha 1 temos o comando de execução. Linhas 2-5 os valores numéricos do alinhamento. Linhas 6-9 e 10-13 temos dados detalhados quanto aos alinhamentos dos resíduos de cada complexo, e suas distâncias Cα-Cα em Å (linhas 9 e 13), após a sobreposição das estruturas. A distância Cα-Cα, em especial, é um dado que não é apresentado nos resultados do MUSTANG original e foram extraídos diretamente do código-fonte e emitido na saída pelo MUSTANGmod. Esse dado é crucial para calcular o RMSD da interface (iRMSD), que será utilizado como critério para definir os grupos.

```
1  pmartins@pop-os:~$ ./mustang_mod -i 1a1m_C_A.pdb 1a1n_C_A.pdb -r ON -s OFF
2  RMSD:   0.693156
3  Length: 288
4  Identity: 272
5  Similarity: 272
6  1a1m_C_A.pdb
7             G,S,H,S,M,R,Y,F,Y,T,A,M,S,R,P,G,R,G,E,P,R,F,I,A,V,G,Y,V,D,D,T,Q,F,V,R,  ...
8             0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26, ...
9             6.10525,0.76278,0.328537,0.251085,0.269233,0.165644,0.174202,0.290518,  ...
10 1a1n_C_A.pdb
11            G,S,H,S,M,R,Y,F,Y,T,A,M,S,R,P,G,R,G,E,P,R,F,I,A,V,G,Y,V,D,D,T,Q,F,V,R,  ...
12            0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26, ...
13            6.10525,0.76278,0.328537,0.251085,0.269233,0.165644,0.174202,0.290518,  ...
```

Figura 10 – Exemplo da tela de execução do MUSTANGmod

O alinhamento par-a-par com o MUSTANGmod teve que ser reprocessado mais de uma vez, seja por problemas encontrados nas estruturas de entrada ou por atualizações e correções necessárias nas fases anteriores. Com isso, foi inviável repeti-lo dessa forma, pois seria sempre necessário uma infraestrutura de computadores poderosos para realizar sua replicação em um tempo hábil, além de dificultar reprocessamentos com adição de novos complexos. Dessa forma, se fez necessário aplicar um filtro avaliando a identidade de sequência. Logo, o alinhamento estrutural só ocorreria entre dois pares de complexos se a identidade do alinhamento de sequência estiver acima de um limite. Apesar do esforço computacional gasto nas versões anteriores, os dados coletados foram úteis para definir esse limite. Assim, considerando apenas os pares de complexos ao qual a identidade de sequência entre os mesmos seja superior a 50% (inspirado no PeptiDB que utiliza 70%), verificou-se 396.050 alinhamentos elegíveis, apenas 21.64% do total (8.571.870), reduzindo a quantidade de alinhamentos necessários em cerca de 80%.

Para observar como os resultados dos alinhamentos se comportam depois dessa redução, a Figura 11 apresenta a distribuição dos resultados de RMSD e iRMSD sobre os 396.050 alinhamentos. Vale ressaltar que o iRMSD é o cálculo do RMSD considerando somente os resíduos das interfaces dos complexos. Os *outliers* foram removidos para reduzir o ruído da figura. O RMSD dos pares de complexos com mais de 50% de identidade de sequência concentram os seus valores entre 0 (mínimo) e 2,69 Å (máximo), tendo sua média de 1,26 (sinal de + na figura) e mediana igual a 0,6. Para o valor de iRMSD, o intervalo mínimo e máximo estão entre 0 a 2,34 Å respectivamente, sendo a média igual a 1,61 e mediana 0,77. Com isso, observa-se que a redução concentra os valores de alinhamentos (RMSD e iRMSD) dos pares de complexos entre 0 e aproximadamente 2,5 Å. Assim sendo, esses valores são razoáveis tomando como referência o limite máximo (3,0 Å) utilizado pelo PepX para agrupar complexos (VANHEE et al., 2010).

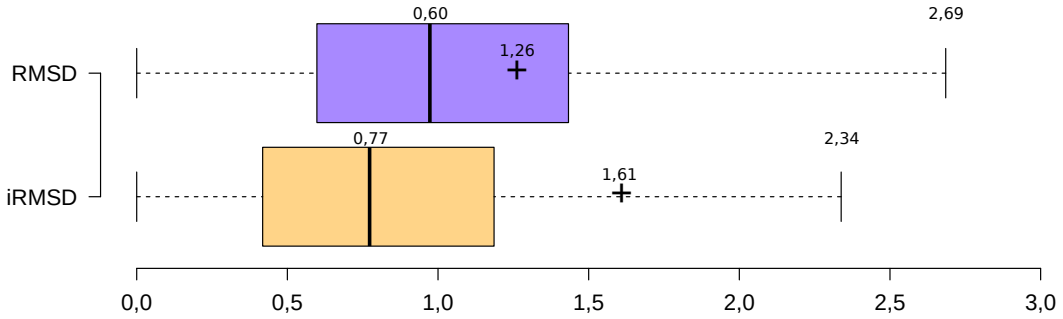


Figura 11 – Diagrama de caixa dos resultados de RMSD e iRMSD de 396.050 alinhamentos estruturais de complexos, considerando apenas complexos com identidade de sequência igual ou superior a 50%.

3.3.2.2 Versão final

Para desenvolver a versão final do AEI foi então aplicada a redução da quantidade de alinhamentos estruturais, baseado na identidades de sequências dos receptores. A Figura 12 apresenta o fluxo de dados para criação do AEI. Em A, é apresentada a etapa de identificação e

execução do alinhamento estrutural e em B o armazenamento dos resultados de alinhamento e criação dos grupos na base de dados.

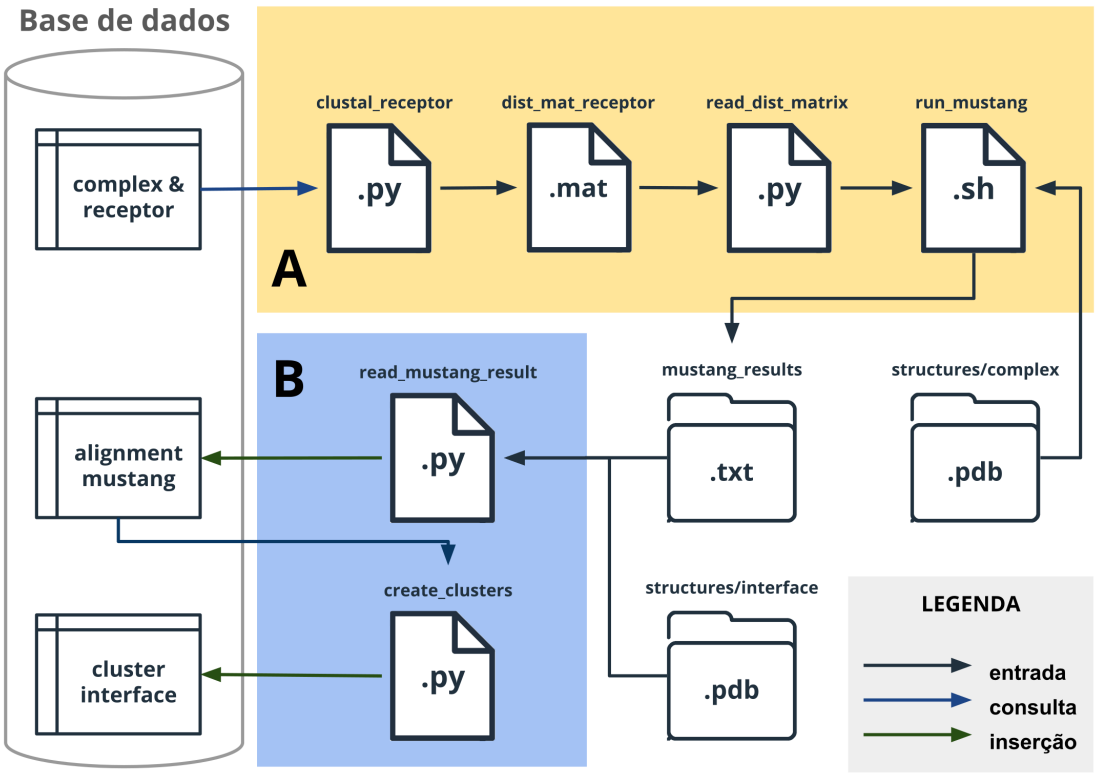


Figura 12 – Fluxo de dados por agrupamento de estrutura de interface.

O processo se inicia com o *script clustal_receptor.py*, que realiza a busca das sequências dos receptores que compõem o CCA. O *script* então executa o Clustal Omega para alinhar todas as sequências dos receptores (par-a-par), armazenando o resultado no arquivo de matriz de distância (*dist_mat_receptor*). Em seguida, o *script read_dist_matrix.py* lê o arquivo de matriz de distância para selecionar os pares que possuem pelo menos 50% identidade de sequência. Ao todo 17.823.435 $((5.971 * 5.970)/2)$ alinhamentos de sequências foram realizados pelo Clustal Omega, mas apenas 455.333 (2,55%) foram selecionados para serem executados pelo MUSTANGmod, reduzindo drasticamente a quantidade de alinhamentos estruturais necessários. Posteriormente, o *script run_mustang.sh* realiza os alinhamentos estruturais dos complexos, considerando tanto a estrutura do receptor, como a do peptídeo. Esta etapa, apesar de ter sito reduzida ainda é custosa, e para otimizar o seu tempo técnicas de processamento paralelo foram aplicadas. Para isso, sua execução foi divida vários processos, utilizando 30 núcleos (*cores*), permitindo a conclusão de todos os alinhamentos em aproximadamente 4 horas de execução apenas. A Figura 13 apresenta um exemplo do resultado de um alinhamento estrutural realizado pelo MUSTANG, onde duas interfaces estão sobrepostas. As esferas azuis e vermelhas-escuras representam os Cα da interface das estruturas comparadas. O número que aparece em branco são as distâncias entre os Cα. Esferas brancas são os Cα que não foram alinhados.

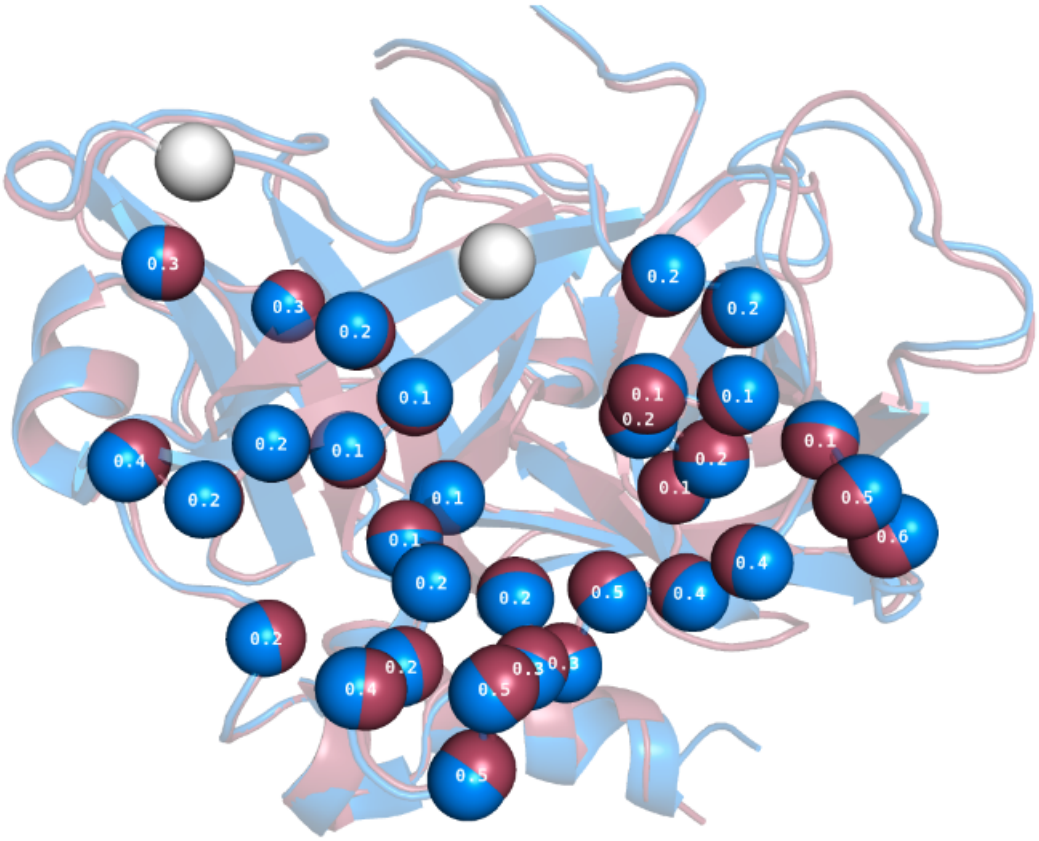


Figura 13 – Sobreposição de duas estruturas como resultado do alinhamento do MUSTANG

Após a execução do *script run_mustang.sh*, os resultados dos alinhamentos foram salvos no diretório *mustang_results*. O *script read_mustang_result.py* então interpreta os arquivos de saída gerados pelo MUSTANGmod (conforme apresentado na Figura 10) e também calcula o iRMSD, identificando o resíduos da interface através dos arquivos pdb do diretório *structures/interface*. O dados gerados são então armazenados na tabela *alignment_mustang*.

Por fim, o *script create_clusters.py* foi criado para criar os grupos do AEI e inserir os dados na tabela *cluster_interface*. Para definir os grupos, considerou-se o iRMSD como unidade de similaridade. Como abstração computacional, foi utilizado o conceito de grafos, com o auxílio da biblioteca NetworkX (HAGBERG; SWART; CHULT, 2008). Um grafo é representado como um conjunto de vértices e arestas. As arestas são constituídas por pares de vértices. Grafos são utilizados para representar modelos em que existem relações entre os objetos. Assim, um grafo foi criado para representar as relações entre os complexos, onde os vértices são os complexos e as arestas são os alinhamentos entre eles. Os valores iRMSD e o percentual de resíduos foram usados como peso em cada aresta. Os vértices que não possuem nenhuma aresta foram considerados automaticamente como grupos unitários. Para definir os grupos, foi aplicado o método de agrupamento baseado em grafo (*graph-based clustering*). Nele, os complexos foram mapeados como vértices de um grafo e seus alinhamentos como arestas, considerando valor do iRMSD como peso. O limite de corte definido para remover arestas foi de 2,0 Å. Assim, as

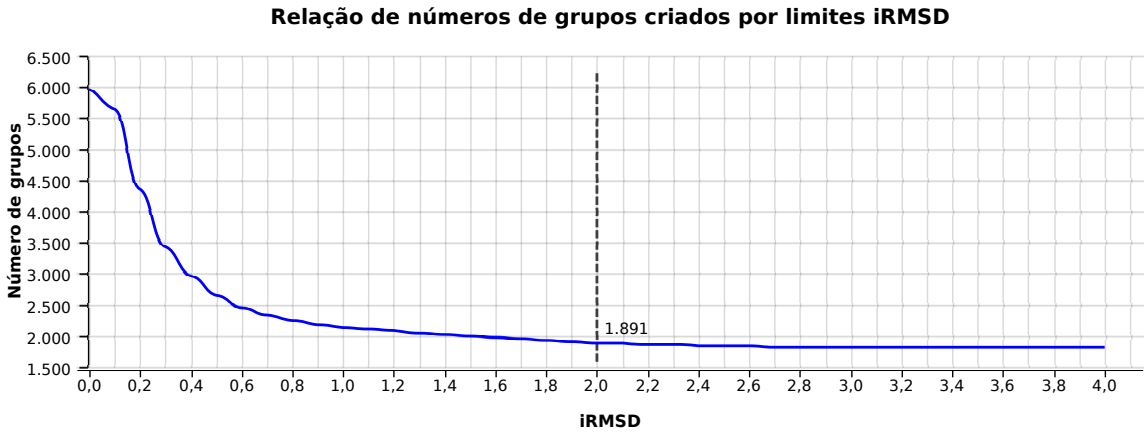


Figura 14 – Relação da quantidade de grupos formados de acordo com os limites de iRMSD, variando em passos de 0,1 Å.

arestas que tiverem o valor de iRMSD menor ou igual a este limite foram removidas. Esse limite foi o mesmo escolhido como um dos critérios de agrupamento utilizados pelo PepX (VANHEE et al., 2010). Além desse argumento, conforme pode ser observado na Figura 14, que apresenta uma relação entre os números de grupos criados o valor limite de iRMSD, a quantidade de grupos formados deixa de variar próximo de 2,0 Å.

A Figura 15 apresenta um exemplo simplificado para ilustrar a aplicação do método de agrupamento baseado em grafo.

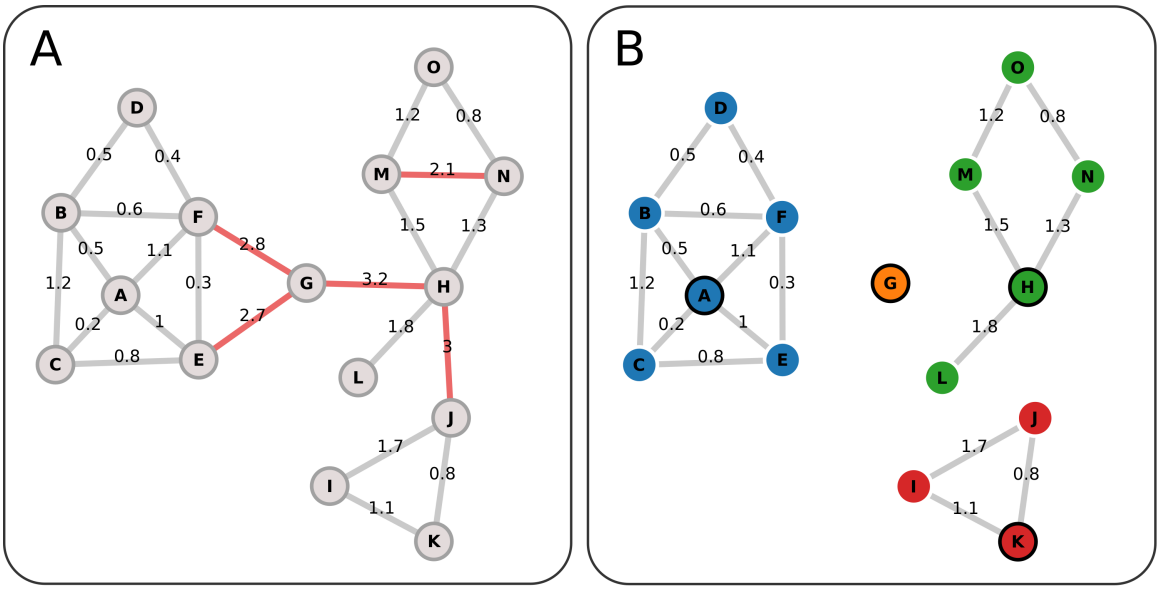


Figura 15 – Exemplo simplificado do método de agrupamento baseado em grafo, aplicado para criar o AEI. Nos grafos, os vértices são complexos (representados por letras) e as arestas são os alinhamentos, aos quais os peso são representados pelo iRMSD. Em A, são destacadas as arestas que serão removidas (em vermelho) e em B, os grupos formados após a remoção dos alinhamentos (arestas) acima de 2,0 Å, formando 4 grupos (subgrafos).

Em A, apresenta-se um grafo, onde os vértices são os complexos (círculos cinzas com letras) e as arestas os alinhamentos. Os números que aparecem nas arestas são os pesos, representados pelo valor de iRMSD do alinhamento entre dois complexos. Arestas vermelhas serão removidas por estarem acima do limite definido (2,0 Å). Em B, estão ilustrados os grupos formados pelos subgrafos resultantes (azul, laranja, vermelho e verde) após a remoção das arestas. Os complexos destacados (círculos pretos) são os centróides de cada grupo. O centróide de cada grupo foi definido como o complexo com maior número de alinhamentos entre os outros complexos do mesmo grupo. Em caso de empate, definiu-se o centroide como sendo aquele ao qual a soma entre todos os valores de iRMSD de seus alinhamentos dentro do grupo for a menor.

Ao final deste processo, 1.891 grupos do AEI foram criados, dos quais 1.356 são unitários e 535 grupos possuem mais de um complexo.

3.3.3 Sítio de ligação

Para criar o agrupamento por sítio de ligação (ASL) os grupos de complexos foram formados considerando as propriedades físico-químicas que ocorrem nas interfaces proteína-peptídeo. Para isso, foi utilizado o ProBiS (KONC; JANEŽIČ, 2010), um algoritmo capaz de detectar sítios de ligação similares, realizando uma comparação estrutural local, orientada à superfície das estruturas. Para realizar a comparação, a superfície é transformada em um grafo, no qual cada vértice é um ponto no espaço tridimensional que representa um grupo funcional de um resíduo. Grupos funcionais são grupos específicos de átomos em resíduos com propriedades físico-químicas, que incluem aceptores de ligação de hidrogênio, doadores de ligação de hidrogênio, aceptores/doadores misto, aromáticos e alifáticos.

Conforme ilustrado na Figura 16, cada grupo funcional é representado por uma esfera colorida. Subgrafos (grafo resultante) são gerados a partir do alinhamento da proteína consulta com a proteína alvo, considerando o clique maximal no grafo. Um subgrafo é representado como parte de um grafo, onde alguns (ou todos) vértices e arestas são selecionados. Se um subgrafo possui todos os seus vértices conectados por arestas entre si, o mesmo é considerado um clique. Com isso, o clique maximal é representado como o clique de um grafo que possui a maior quantidade de vértices dentre os outros cliques possíveis.

O ProBiS permite que apenas uma região da superfície seja selecionada (conjunto de resíduos), e no caso do ASL, os residuos que compõem a interface proteína-peptídeo foram escolhidos. Em contraste com o MUSTANG, utilizado no AEI, o ProBiS é capaz de detectar similaridade de sítio de ligação mesmo quando os complexos comparados possuem interfaces diferentes, a nível de conformação estrutural. Ao final do alinhamento, o ProBiS retorna uma métrica de alinhamento, obtido com base em outros valores do alinhamento, partir da equação:

$$al_{metrica} = log\left(\frac{n_{vert} * log(1 + 1/e_{value})}{rmsd}\right)$$

onde n_{vert} representa a quantidade de vértices alinhados, $rmsd$ o RMSD entre os pares de vértices alinhados sobrepostos e e_{value} é o valor esperado do alinhamento, calculado a partir da equação de Karlin–Altschul (KARLIN; ALTSCHUL, 1990). Assim, quanto maior a métrica de alinhamento entre um par de complexos, mais semelhantes são seus sítios de ligação. A Figura 17 exibe o fluxo de dados feito para criar o ASL. Em A, ocorrem as extração das superfícies e alinhamento das estruturas pelo ProBiS. Em B, os *script* interpretam os resultados dos alinhamentos realizados pelo ProBiS, definem os grupos do agrupamento e por fim inserem os dados em suas respectivas tabelas na base de dados.

O processo inicia-se com extração das superfícies da interface de cada complexos do CCA, pelo *script extract_surface_binding.py*. Para acelerar a comparação entre duas estruturas, o ProBiS é capaz de gerar arquivos de superfície (formato .srf) a partir do arquivo pdb dos complexos. Além disso, é possível indicar os resíduos de interesse que serão comparados. Assim, o *script* cria os arquivos de superfície, considerando os resíduos que compõem a interface dos complexos, utilizando como referência os arquivos de estrutura da interface (diretório *structures/interface*), conforme apresentado na Seção 3.2. Os arquivos de superfície gerados são então armazenados no diretório *probis_surfaces*, para serem utilizados em seguida pelo *script mult_alignment_probis.py*. Então, o mesmo realiza o alinhamento par-a-par do escopo CCA, executando o ProBiS e considerando os arquivos de superfície como parâmetros de entrada. Essa fase em especial requer um alto poder computacionalmente, por realizar alinhamentos par-a-par de 5.971 complexos. Diferente da redução de pares de complexos pela sequência primária aplicada no AEI, no ASL todo os pares foram comparados. Para realizar tal façanha, foi necessário que sua execução ocorresse em um servidor de alto poder de processamento. Além disso, o próprio ProBiS permite o uso de computação paralela para realizar os alinhamentos, o que auxiliou na conclusão desta etapa.

Após o término da execução do *mult_alignment_probis.py*, os arquivos de saída, contendo os valores e dados dos alinhamentos (formato NoSQL) foram então armazenados no diretório *probis_results*. Apesar desses arquivos estarem no formato de texto, não é possível entender os

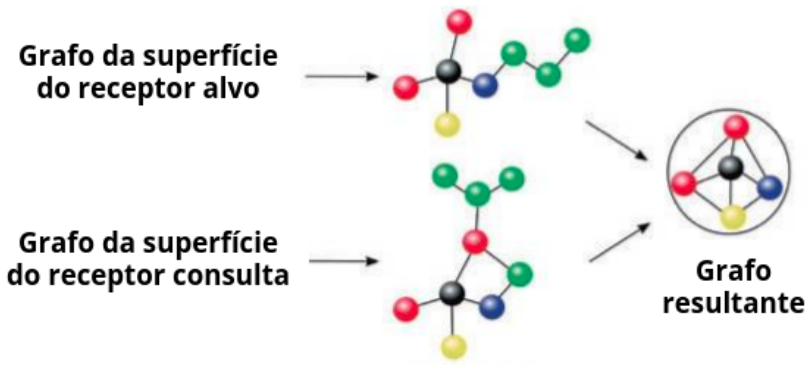


Figura 16 – Comparação de grupos funcionais de duas proteínas, onde suas superfícies representadas como grafos. Adaptado de Konc e Janežič (2010).

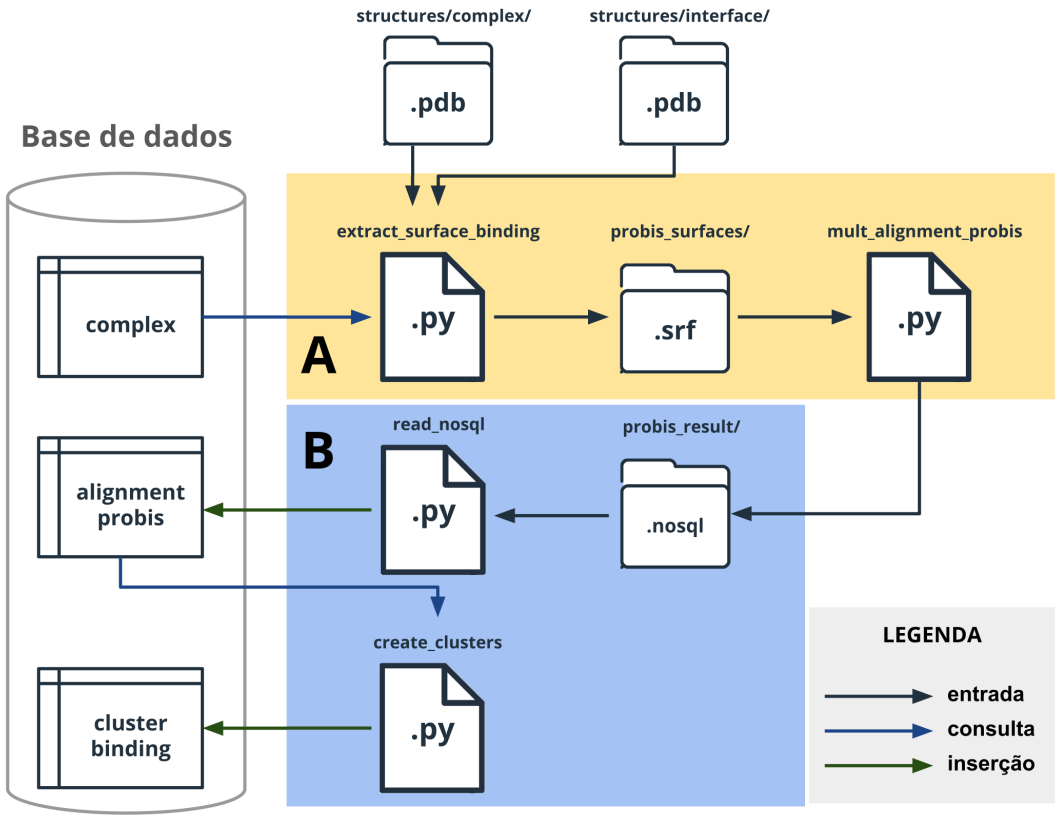


Figura 17 – Fluxo de dados do agrupamento por sítio de ligação (ASL).

valores apresentados sem conhecer a sua estrutura contextual e, por isso, o ProBiS possui opções capazes de converter o arquivo NoSQL em um JSON (JavaScript Object Notation). Entretanto, para este último, a conversão pelo ProBiS acaba reduzindo as casas decimais, e com isso há uma perda considerável do valor real dos resultados. A Figura 18 ilustra um exemplo comparando os dois formatos, demonstrando a perda do valor real. A esquerda é apresentado o formato JSON, com os resultados rotulados. À direita, o formato NoSQL, onde é possível ver que e necessário conhecer a estrutura do arquivo previamente para interpretá-lo corretamente. É possível observar (destacado em vermelho) o arredondamento feito nos valores de RMSD de 0,38917 para 0,4 e na métrica de alinhamento de 9,8156 para 9,82. Apesar desse arredondamento parecer mínimo, ele implica em resultados de métrica de alinhamento errado, uma vez que o RMSD é o usado para seu cálculo. Por esse motivo, foi utilizado apenas o formato NoSQL. A sua análise correta, considerando as posições de cada resultado, foi possível através da leitura manual do código-fonte do ProBiS. A implementação foi realizada no *script read_nosql.py*, que interpreta e insere os resultados na tabela de alinhamentos por sítio de ligação (*alignment_probis*).

Para finalizar a criação do ASL, o *script create_clusters.py* foi utilizado para consultar os resultados dos alinhamentos na tabela *alignment_probis* e definir os grupos de forma similar ao processo utilizado para criar o AEI. Um grafo foi criado, utilizando os complexos como vértices e os seus alinhamentos como arestas, faltando apenas a definição do limite para criar os grupos. O ProBiS, além do algoritmo para definir sítio de ligação, possui também uma base de dados

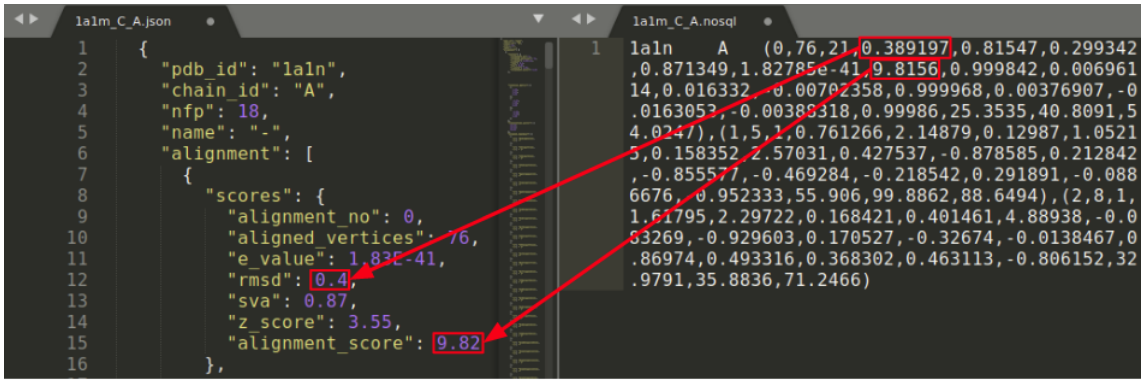


Figura 18 – Comparação dos tipos de arquivos de saída do ProBiS. À esquerda é apresentado a saída no formato JSON. À direita, no formato NoSQL. A transformação de NoSQL para JSON pelo ProBiS, apesar de tornar o resultado legível, implica em arredondamento considerável e consequentemente na perda do valor real do resultado.

(KONC et al., 2012). Nela, foi adicionado escore Z (*Z score*) para representar a significância do alinhamento baseado na população. Seguindo a mesma ideia, o escore Z foi definido para cada alinhamento utilizando como população as métricas de alinhamento contidos na tabela *alignment_probis*, conforme a seguinte equação:

$$Z_{escore} = \frac{al_{metrica} - \mu}{\sigma}$$

onde μ e σ são a média da população e o seu desvio padrão, respectivamente. Considerando o escopo CCA, a média apresentou o valor de 1,55, com desvio padrão de 4,91, sendo assim possível definir o escore Z para cada alinhamento. O ProBiS sugere que escores Z acima de 2.0 possui uma alta significância, porém, a média da população e o seu desvio padrão utilizadas são diferentes. Por isso, foi realizado uma avaliação, relacionando a quantidade grupos criados. A Figura 19 apresenta a quantidade de grupos formados conforme variação no limite do escore Z. O limite foi definido empiricamente com base na observação do crescimento da curva, que apresenta a quantidade de grupos formados o limite de escore Z aumenta. Assim, o valor de 1,6 do escore Z foi escolhido como critério de corte, considerando o conceito de grafo similar ao apresentando na Figura 15, porém, onde as arestas são representados pelo alinhamento do ProBiS e seus pesos são o escore Z. Assim, ao remover os alinhamentos (arestas) com valores de escore Z inferiores a 1,6 os subgrafos resultantes definiram os grupos de complexos. Ao todo, 1.812 grupos de complexos foram criados, sendo 501 grupos com mais de um complexo e 1.311 grupos unitários.

3.4 Criação do serviço *web*

O serviço *web* criado visa ser uma interface interativa capaz de recuperar, explorar e visualizar as informações contidas na base de dados. Além disso, o serviço permite que o usuário

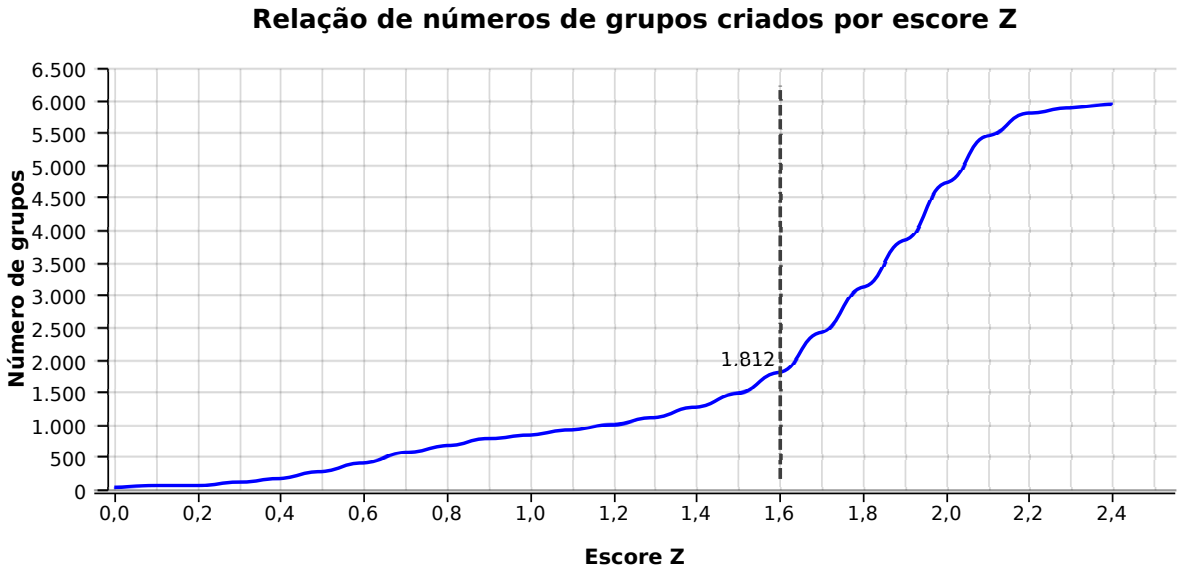


Figura 19 – Relação da quantidade de grupos gerados de acordo com os limites de Z, variando em passos de 0,1.

insira um sequência para buscar por peptídeos ou receptores com sequências similares e também por sítios de ligação, caso o usuário forneça a estrutura da proteína alvo e os resíduos de interesse.

Para criar o serviço *web*, foi utilizado servidor HTTP Apache (*Apache HTTP Server*) versão 2.0 e a linguagem interpretada PHP. Como *framework* de desenvolvimento, o CodeIgniter (UPTON, 2007) foi escolhido por ser seguro, ágil e de fácil manejo. Para tornar as páginas mais interativas, metodologias de requisições assíncronas utilizando Ajax (*Asynchronous Javascript and XML*) foram amplamente utilizadas, além de bibliotecas *JavaScript* incluindo: DataTables (<datatables.net>), 3Dmol.js (REGO; KOES, 2014) e D3 (<d3js.org>). Tais bibliotecas foram cruciais para manter a interatividade do serviço *web*, permitindo que os complexos pudessem ser listados em tabelas dinâmicas e paginadas. Permitiu ainda que suas estruturas tridimensionais fossem visualizadas e comparadas utilizando os navegadores de internet atuais.Por fim, as bibliotecas foram úteis na criação de gráficos dinâmicos dirigidos a dados. Tanto a base de dados como o serviço *web* estão hospedados em um servidor Ubuntu 14.04 com processador Intel Xeon com 12 núcleos (cores) de processamento e 64GB de memória principal (RAM).

Recursos de busca foram desenvolvidas para que o usuário possa inserir uma proteína alvo para encontrar complexos similares, baseado em sua sequência (receptor ou peptídeo) e também sitio de ligação. O mecanismo de busca de sequência (receptor/peptídeo) é baseado na ferramenta BLASTp, da suíte NCBI-BLAST+ (ALTSCHUL et al., 1990; CAMACHO et al., 2009) enquanto para encontrar receptores com sítios de ligação similares, foi utilizado o algoritmo ProBiS (KONC; JANEŽIČ, 2010).

4 Resultados e discussão

Neste capítulo serão descritos os resultados obtidos através das observações e análises dos dados contidos na base de dados, além de discussões quanto aos produtos gerados por este trabalho, além dos estudos de casos desenvolvidos.

4.1 Tamanho de um peptídeo

Conforme explicado anteriormente, há outros repositórios de dados sobre peptídeos disponíveis. Neles, o tamanho de "peptídeo" pode variar consideravelmente (ver Seção 1.3). É evidente que não há um consenso sobre esta definição e para o presente trabalho, o termo "peptídeo" foi definido como sendo uma cadeia polipeptídica de 2 a 50 aminoácidos. Optamos por usar os limiares mínimo e máximo encontrados na literatura (VANHEE et al., 2010; LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010; DAS et al., 2013; FRAPPIER; DURAN; KEATING, 2018; WEN et al., 2019; XU; ZOU, 2020) para manter a maior abrangência possível. Seus autores o definem de forma empírica ou baseado em outros trabalhos. Ao todo, a base de dados desenvolvida nesta tese possui 19.813 complexos, dos quais pôde-se extrair 13.626 peptídeos distintos. A Figura 20 apresenta os gráficos de distribuição e diagrama de caixa dos tamanhos dos peptídeos. Em vermelho, estão representados os tamanhos de todos peptídeos distintos da base de dados (13.626). Neste escopo, verifica-se no gráfico de linha (Figura 20 A) uma alta concentração de peptídeos com tamanhos entre 3 a 14 aminoácidos, e em conformidade com o gráfico de dispersão (Figura 20 B), observa-se o intervalo de 4 a 44 para o mínimo e máximo (considerando percentis 5º a 95º, respectivamente) e 8 a 20 para o primeiro e terceiro quartil, respectivamente. Em azul, são apresentados apenas os peptídeos constituídos dos aminoácidos padrão (8.796). Em contraste com todos os peptídeos, observa-se no gráfico de linha que a concentração dos tamanhos estão entre 4 e 11, não passando da quantidade de 300 peptídeos neste intervalo. Nota-se que os diagramas de caixa possuem perfil similar, com diferença apenas no primeiro e terceiro quartil (9 a 21, respectivamente), mas mantendo a mediana no tamanho em 11.

Uma boa escolha seria utilizar os dados entre quartis (comprimentos entre 8 a 21) ou entre os valores entre o 5º e o 95º percentil (comprimentos entre 4 e 44).

4.2 Agrupamentos

Os agrupamentos de complexos visam remover redundância e agregar complexos que sejam similares, dependendo do aspecto utilizado para definir a similaridade. Ao todo, foram criados três tipos agrupamentos para os complexos da base de dados, considerando sequência do

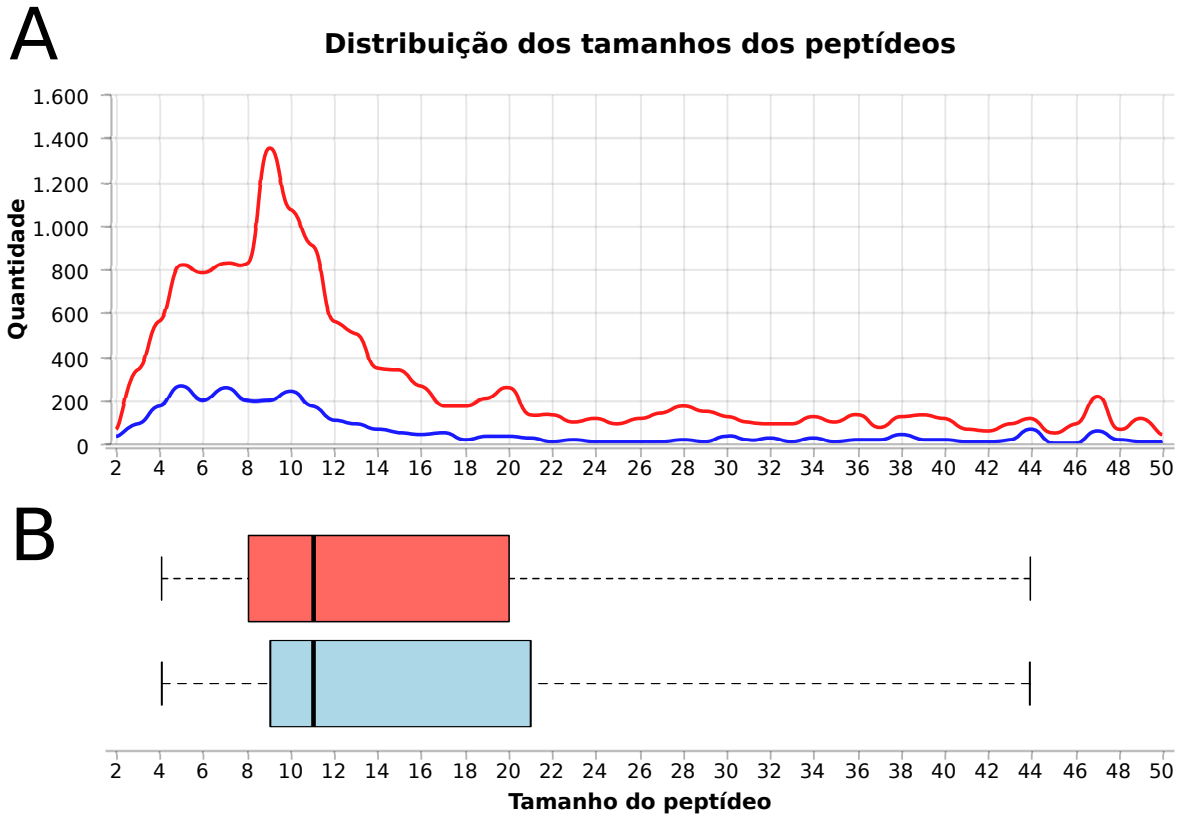


Figura 20 – Gráfico de distribuição dos tamanhos dos peptídeos, representado como gráfico de linhas (A) e diagrama de caixa (B). Em vermelho, estão representadas as distribuições de todos os peptídeos da base de dados, enquanto em azul, estão os que possuem apenas os aminoácidos principais.

peptídeo (ASP), a estrutura de interface proteína-peptide (AEI) e o sítio de ligação (ASL). O ASP tem como objetivo distinguir as distintas sequências dos peptídeos nos complexos, e para isso, leva em consideração a identidade de sequência entre os mesmos. O AEI leva em conta o iRMSD de cada par de complexo para definir seus grupos. E finalmente, o ASL utiliza o escore Z, que está relacionada com a métrica de alinhamento como aspecto de similaridade.

Avaliar o comportamento da cada grupo, verificando cada par de complexos e seus critérios de similaridade, é uma tarefa complexa uma vez que o número de pares dentro de único grupo pode ser enorme. Tomando como exemplo o grupo I0 (primeiro e maior grupo do AEI), se fosse representado como um grafo, o mesmo teria 659 vértices (complexos) e 210.715 arestas (alinhamentos). Por conta desse colume, é inviável a visualização desses dados. Fizemos então diagramas de caixas para apresentar os agrupamentos. As Figuras 21, 22 e 23 apresentam diagramas de caixas dos agrupamentos ASP, AEI e ASL, respectivamente, considerando os valores das distribuição seus devidos aspectos entre os pares de complexos do grupo: identidade de sequência, iRMSD e Z, respectivamente. Removemos os *outliers* para reduzir a complexidade visual. Entraremos mais em detalhes a seguir.

4.2.1 Por sequência de peptídeo (ASP)

O ASP possui 1.845 sequências de peptídeos distintos, em 1.074 grupos de unitários (sequência única) e 771 agrupamentos de fato. Deste último (771 grupos), a Figura 21 exibe apenas 203, sendo eles os grupos que possuem 3 ou mais sequências com identidade de sequências diferentes de 100%. Evitamos assim apresentar grupos "perfeitos", ao qual seus elementos sejam idênticos entre si. Os grupos são ordenados pela mediana, assim os grupos da parte superior na Figura 21 tem mais alta similaridade entre si (homogeneidade) quando comparados aos da parte de inferior.

Considerando a mediana como referência, observa-se que apenas 5 grupos (S153, S186, S206, S35 e S210) possuem identidade de sequência abaixo de 30%. Para o limite inferior a 40%, 50%, 60%, e 70% de identidade de sequência, essa quantidade aumentam para 20, 39, 55 e 60 grupos, respectivamente. É evidente a existência de grupos com dissimilares de sequências significativa, porém, perante ao agrupamento como um todo, são poucos os grupos com tais dissimilaridades. Por exemplo, considerando a mediana como critério, apenas 91 grupos possuem identidade de sequência abaixo de 80%, o que representa cerca de 5% de todo ASP. Em outra palavras, mais de 95% de todo agrupamento possui a mediana de identidade de sequência superior a 80%.

4.2.2 Por estrutura de interface (AEI)

A Figura 22 apresenta os grupos do AEI que possuem 10 ou mais alinhamentos entre seus complexos, totalizando 172 grupos exibidos, de um total de 1.891. Esse redução foi aplicada para dar preferência a grupos maiores, o que consequentemente aumenta seu grau de dissimilaridade. Os grupos apresentados foram ordenados por suas medianas. Na parte superior do gráfico estão os grupos com maior similaridade em comparação aos da parte inferior (mais distante do iRMSD de 0,0). O AEI foi agrupado considerando o critério de corte de 2,0 Å para formar os grupos, logo, considerando esse mesmo limite, apenas 2 grupos (I15 e I162) possuem uma mediana superior ao iRMSD de 2,0. Considerando o 3º quartil (75% da população) como referência, 10 grupos ultrapassam a marca de 2,0 Å. Quando muda-se a referência para o valor máximo de cada distribuição, 48 grupos são encontrados. Novamente, assim como colocado para o ASP, esses 48 grupos possuem um impacto muito pequeno quando comparados ao todo. Quando aplicado a todos os grupos do AEI (total de 1.891) esta quantidade representa apenas 2,53%.

4.2.3 Por sítio de ligação (ASL)

As distribuição dos valores de alinhamento dos grupos do ASL estão representados na Figura 23. Assim como definido na discussão do AEI, grupos com 10 ou mais alinhamentos são exibidos para permitir uma análise sobre os menos similares, onde encontra-se 210 grupos. A figura apresenta o valores de escore Z dos alinhamento, ao qual foram derivados das métricas

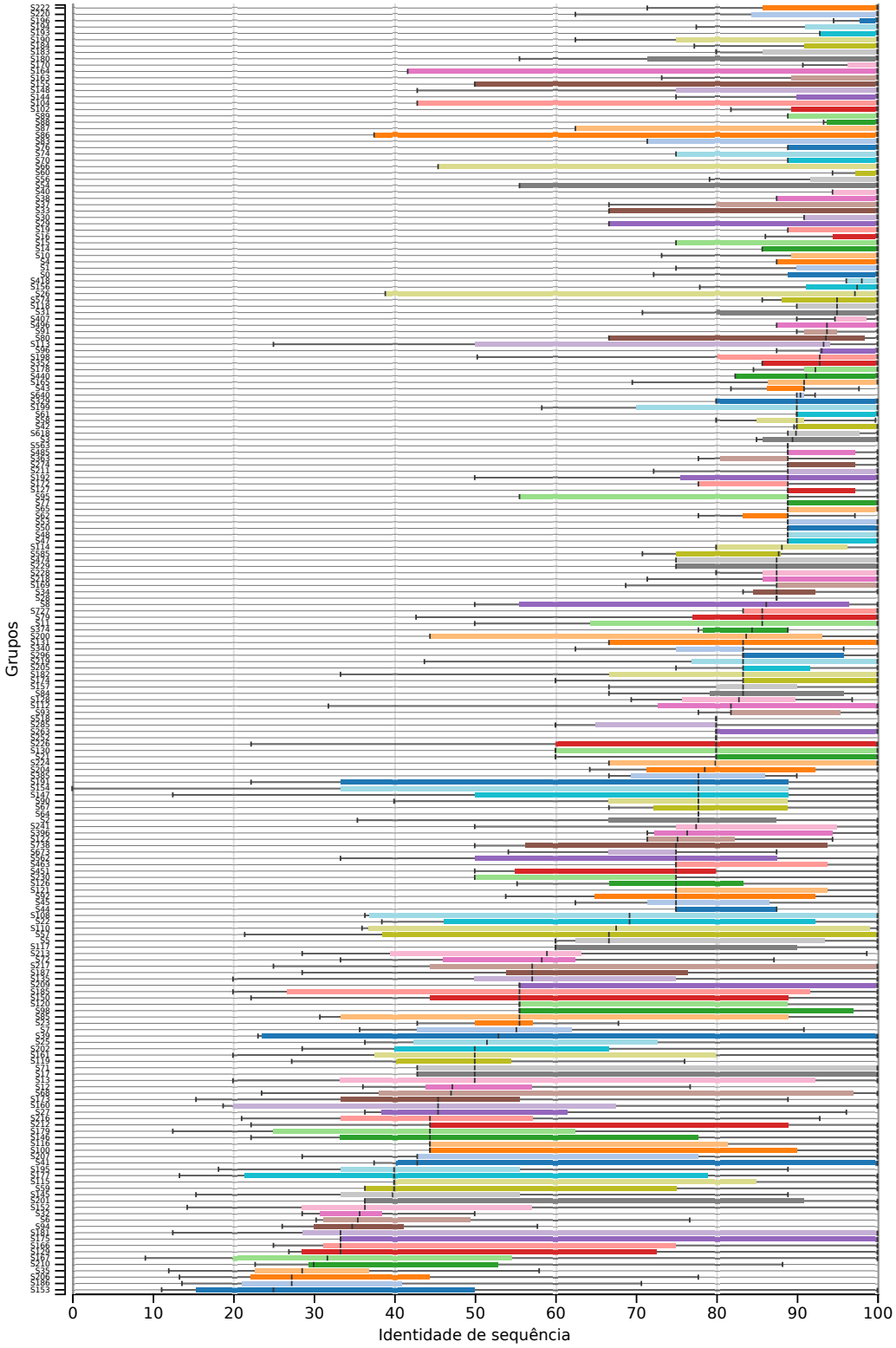


Figura 21 – Diagrama de caixa dos grupos do ASP. Somente os grupos com mais de 3 sequências e nos quais ao menos um dos valores de identidade de sequência seja inferior 100%, totalizando 203 grupos.

de alinhamentos, juntamente com a média e desvio padrão da população. Vale ressaltar que, diferente do ASP e AEI, onde os valores possuem um valor "ótimo" de similaridade (100%

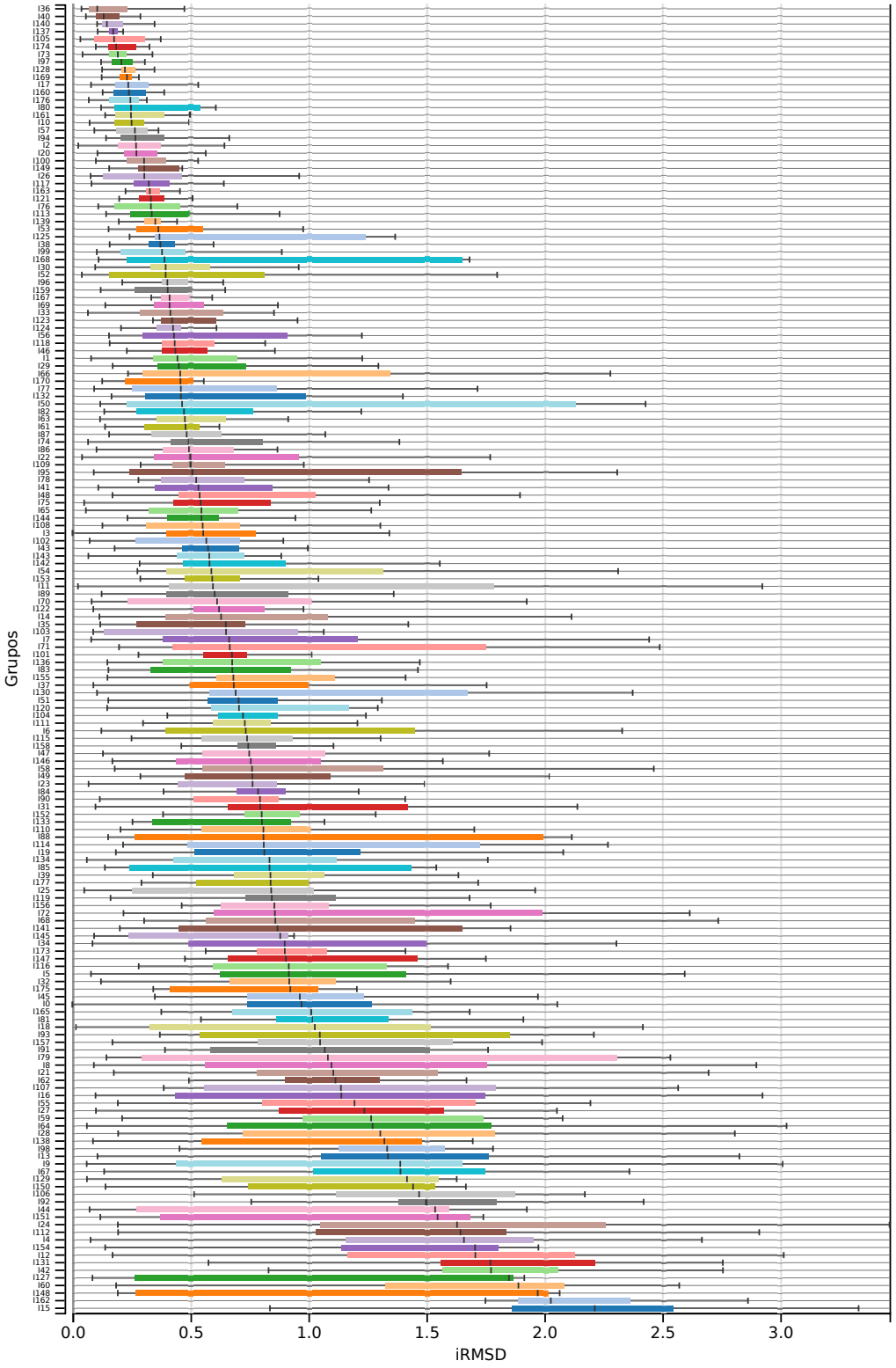


Figura 22 – Diagrama de caixa dos grupos do AEI. Apenas os grupos com quantidade de alinhamentos de 10 ou mais são exibidos, totalizando 172 grupos.

identidade de sequência e 0.0 Å de iRMSD, respectivamente), o ASL, utiliza o escore Z de 1.6 para definição dos grupos, ao qual quanto maior o seu valor, mais similares são os sítio de ligação. Assim, não há um valor limite para definição do “ótimo”, porém, quanto maior o escore

Z, mais similares são os sítios de ligação entre complexos.

Ainda observando a Figura 23, os grupos estão ordenados pela mediana, sendo perceptivo a sua sutil entre os escores Z próximos de 1,3 a 2,1, com exceção dos grupos B185 e B117 (topo) que obtiveram medianas acima de 5,5. 26 grupos são encontrados quando considera-se a mediana como referência com escore Z abaixo de 1,6. Passando a referência para o 1º quartil, o número de grupos aumenta para 81 e quando usa-se o valor mínimo, encontra-se 151 grupos o que representa 8,33% de todo ASL.

A discussão aqui realizada visa uma avaliação qualitativa e quantitativa sobre os agrupamentos obtidos. Além disso, por se tratar de agrupamentos baseados em grafos (com exceção do ASP), muitos grupos possuem uma alta variação entre os valores de alinhamento entre seu complexos, indicando que talvez seja necessário avalia-los manualmente para verificar se existem possíveis subgrupos. Apesar disso, definir grupos em um agrupamento é uma tarefa muito subjetiva, e muitas vezes não é possível encontrar um agrupamento "perfeito". Sendo assim, o objetivo desta análise foi justamente verificar, a partir de aspectos numéricos, como os complexos se agruparam a partir de três aspectos distintos. Uma vantagem desse agrupamento híbrido é a possibilidade de poder compará-las. Assim, é possível por exemplo, verificar os distintos peptídeos (com ASP) dentro de um grupo do AEI ou ASL, ou qualquer uma dessas três combinações possíveis.

4.3 Produtos

4.3.1 Base de dados Propedia

A base de dados foi o principal produto gerado por este trabalho. Todos os complexos recuperados foram anotados junto com suas características como código PDB, estrutura e sequência do receptor/peptídeo, classificação, organismo, área de interface, peso molecular, aromaticidade, instabilidade, ponto isoelétrico, hidrofobicidade e assim por diante. Ao todo, a base de dados dispõe de 19.813 complexos, dos quais 5.971 (complexos com peptídeos formado de aminoácidos naturais com interação com apenas um receptor) formaram o escopo CCA, que permitiu o agrupamento de complexos (ASP, AEI e ASL) a partir de três aspectos distinto. Tais agrupamentos não apenas ajudam a remover a redundância, e definir complexos únicos na base de dados, como também permitem comparações entre os agrupamentos considerando: a sequência do peptídeo; a estrutura da interface proteína-peptídeo e o sítio de ligação. Dessa forma, a base de dados pretende ser um recurso disponível para as comunidades de biologia estrutural e computacional. Além disso, a base de dados pode ser útil para o estudo do reconhecimento proteína-peptídeo ou para a construção conjuntos de treinamento e testes para abordagens de ancoragem molecular, a fim de auxiliar também em projetos de desenhos de peptídeos como fármacos ou soluções biomoleculares.

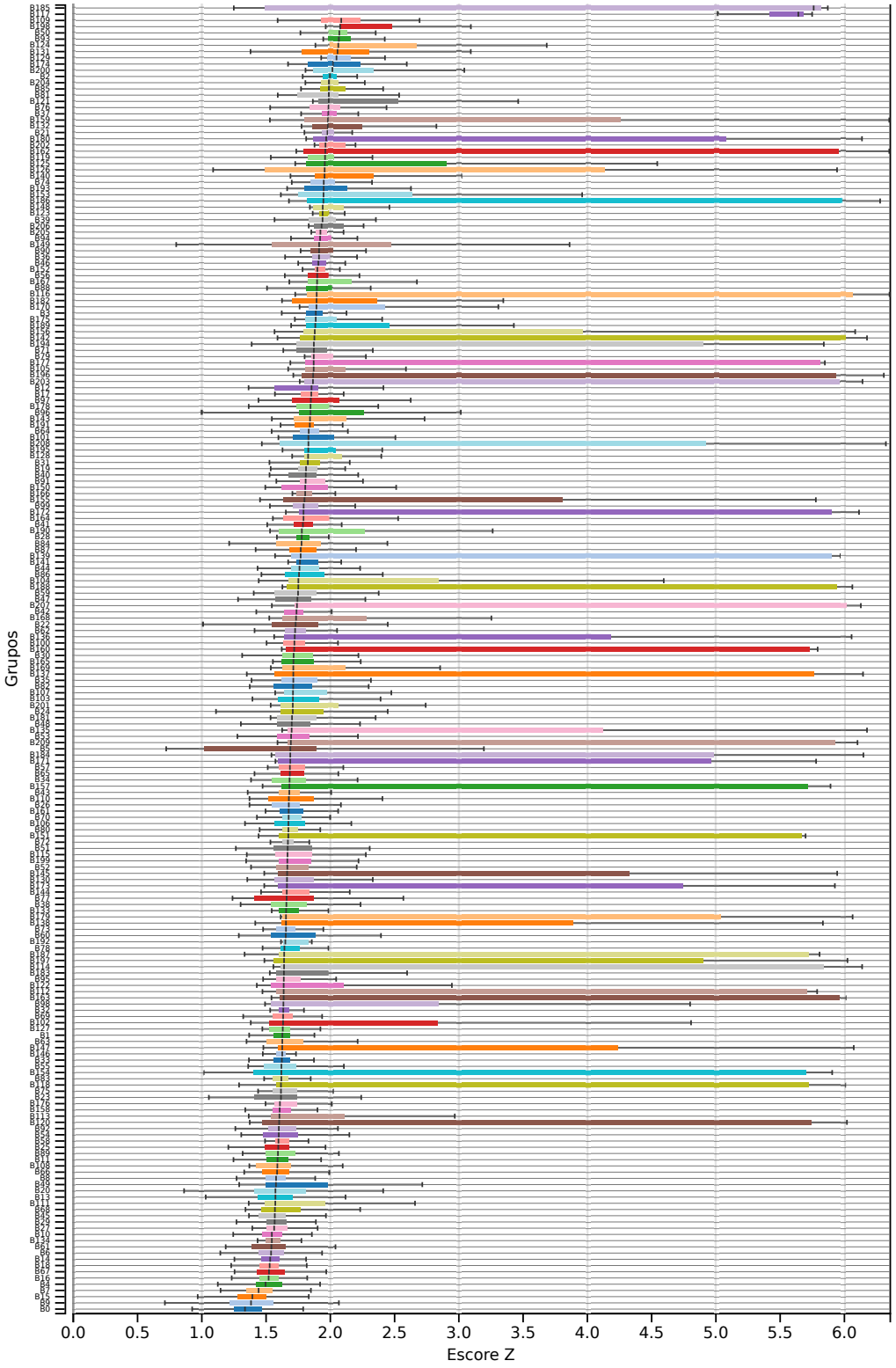


Figura 23 – Diagrama de caixa dos grupos do ASL. Somente os grupos com 10 ou mais alinhamentos são apresentados, representando um total de 210

Comparada a outras bases de dados similares (ver Seção 1.4), o Propedia possui um vasta quantidade de complexos, além de um serviço *web* que permite explorar seu conteúdo de forma fácil e ágil. A Tabela 2 apresenta um comparativo de características e propriedades entre a base

de dados Propedia e as bases existentes e mencionadas na Seção 1.4. Em comparação com as outras bases de dados, o Propedia é a base de dados mais abrangente e atualizada. Assim como o PepBDB, o Propedia apresenta seus complexos através de um serviço *web*, porém, dispõe de uma interface mais amigável e rica em recursos de recuperação da informação, além de ser capaz de realizar buscas por sítio de ligação, a partir de uma estrutura alvo.

Nome	**Quantidade de Complexos**	**Tamanho do peptídeo (aa)**	**Resolução (Å)**	**Tipo**	**Disponível**	**Busca por sequência**	**Busca por estrutura**
Propedia	19.813	2-50	< 2,5	Serviço *web*	✓	✓	✓
PepX	1.431	5-35	< 2,5	Serviço *web*	N.D.	×	×
PeptiDB	103	5-15	< 2,0	Lista PDB	✓	×	×
PepBind	5.314	<= 35	-	Serviço *web*	N.D.	×	×
PixelDB	1.966	5-50	< 2,5	GitHub	✓	×	×
PepBDB	12.241	< 50	-	Serviço *web*	✓	×	✓
PepPro	1.198	5-30	< 2,5	Lista PDB	✓	×	×

N.D.: Não disponível. Acesso em 26 de Setembro de 2020.

Tabela 2 – Comparação das bases de dados de complexos proteína-peptídos

4.3.2 Serviço *web*

O serviço *web* está disponível através do endereço eletrônico: <http://bioinfo.dcc.ufmg.br/propediadb> e oferece recursos visuais interativos que permitem explorar os complexos com muita flexibilidade. Cada entrada é composta por um complexo peptídeo-proteína identificados no seguinte formato: <*pdb*>_<*cadeia_peptídeo*>_<*cadeia_receptor*>. A Figura 24 apresenta a página inicial (*home*), que exibe a quantidade total atual de complexos, e o número de complexos não redundantes dos agrupamentos ASP, AEI e ASL, respectivamente.

A interface do serviço *web* permite explorar os complexos considerando filtros, conforme Figura 25 A. Por eles é possível realizar buscar por: pdb/complexo; organismo (peptídeo ou receptor); classificação; resolução; (para estruturas resolvidas por difração de raio X); tamanho do peptídeo ou receptor. Além disso, o usuário pode optar realizar o *download* de todos os complexos filtrados, ou apenas os centroides de cada agrupamento, através dos botões apresentados na Figura 25 B. Os complexos são exibidos em tabelas dinâmicas, ilustradas pela Figura 25 C, ao qual apresentam algumas informações dos complexos, como resolução da estrutura, tamanho do peptídeo, e os grupos ao qual eles pertençam, caso façam parte do escopo CCA.

A partir da página *explore*, o usuário pode navegar em vários escopos, podendo acessar um complexo em particular, um grupo específico ou realizar o *download* completo da base de dados. Além disso, o usuário pode executar buscas a partir de sequências de interesse (menu *Search -> by Sequence*), ou estruturas de receptores (menu *Search -> by Binding Site*), para encontrar sítio de ligações similares, e consequentemente peptídeos que possam interagir nessas regiões.

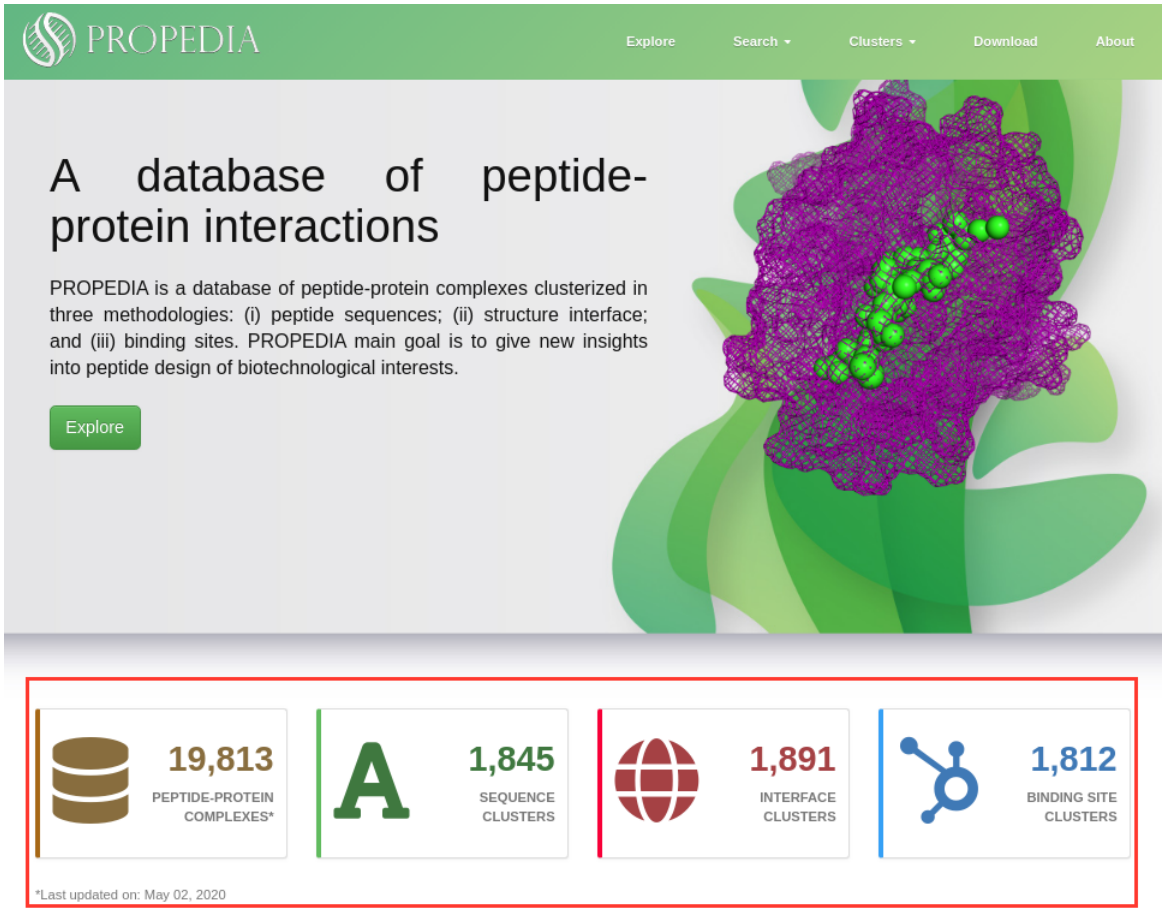


Figura 24 – Página inicial do serviço *web*. Em destaque a quantidade de complexos total da base de dados, e também a quantidade de complexos não redundantes considerando os escopos ASP, AEI e ASL. Última extração do PDB foi realizada em 02/Mai/2020

A Figura 26 apresenta o fluxo de navegação ao qual o usuário pode acessar as páginas e usufruir das funcionalidades que o serviço *web* fornece. Em A, é a apresentado um esboço da tela de resultados de busca por sequência utilizando o BLASTp. Os complexos retornados são ordenados pela pontuação dada pelo BLAST. Identidade de sequência, cobertura, dentre outros, são colunas que aparecem no resultado. A consulta pode ser feita tanto por sequência de receptores como peptídeos. A página em B exibe o resultado de busca por sítio de ligação. O usuário precisa fornecer a estrutura alvo (formato PDB), os resíduos que compõem o sítio de ligação e escolher o escopo de busca (CCA ou todos complexos). Os resultados dos melhores alinhamentos são ordenados pela métrica de alinhamento (dado pelo ProBiS) e o usuário pode visualizar e comparar a estrutura alvo com os complexos resultantes da busca, a qual contém um peptídeo (sempre apresentado na cor amarela) que possa ser promissor para o sítio de ligação alvo. Um diagrama de caixa é apresentado para exibir a distribuição dos resultados da métrica de alinhamento e o usuário para usar o controle deslizante (*slider*) para filtrar os resultados, podendo ser assim filtrar apenas os melhores resultados por exemplo.

A Figura 26 C ilustra a página contendo informações de cada complexo da base

Figura 25 – Página para explorar os complexos com vários recursos de filtros e *downloads*

de dados. Nela, encontra-se informações relacionados ao PDB, receptor, peptídeo, e também dos agrupamentos e complexos similares, caso o mesmo pertença ao CCA. Além disso, *links* para outras bases de dados, como PDB, UniProt (CONSORTIUM, 2015) e PubMed (http://www.ncbi.nlm.nih.gov/pubmed) estão disponíveis na página. A estrutura tridimensional do complexo pode ser visualizada, onde o peptídeos são apresentado como bastões (*sticks*) amarelos, enquanto o receptor está no formato de superfície (*surface*), com conformação estrutural da interação do peptídeo com o receptor.

Para cada agrupamento, os complexos são apresentados conforme ilustrado na Figura 26 D, E e F, sendo elas os agrupamentos ASP, AEI e ASL, respectivamente. Nelas, os valores de alinhamentos entre outras informações, são apresentadas de acordo com um complexo de referência (centroide selecionado como padrão) que pertence ao grupo e ao qual pode ser alterado. Diagramas de caixa são apresentados com os valores de alinhamento do complexos referencia com todos os outros do grupo. Assim como na página de resultados de busca por sítio de ligação, um controle deslizante permite filtrar os complexos apresentados.

Em todas as páginas estão disponíveis botões para *download* dos complexos apresentados

nas tabelas, sendo possível escolher o formato pdb do complexo (receptor e peptídeo), receptor, peptídeo ou apenas os resíduos da interface. Além do formato de estrutura, o *download* pode ser feita no formato de sequência (fasta), tanto do receptor como do peptídeo. Por fim, toda a base de dados pode ser recuperada no formato csv, pdb ou fasta pela página da aba "Download". Por fim, nesta mesma página encontra o formato de base de dados (sql), para o SGBD MySQL.

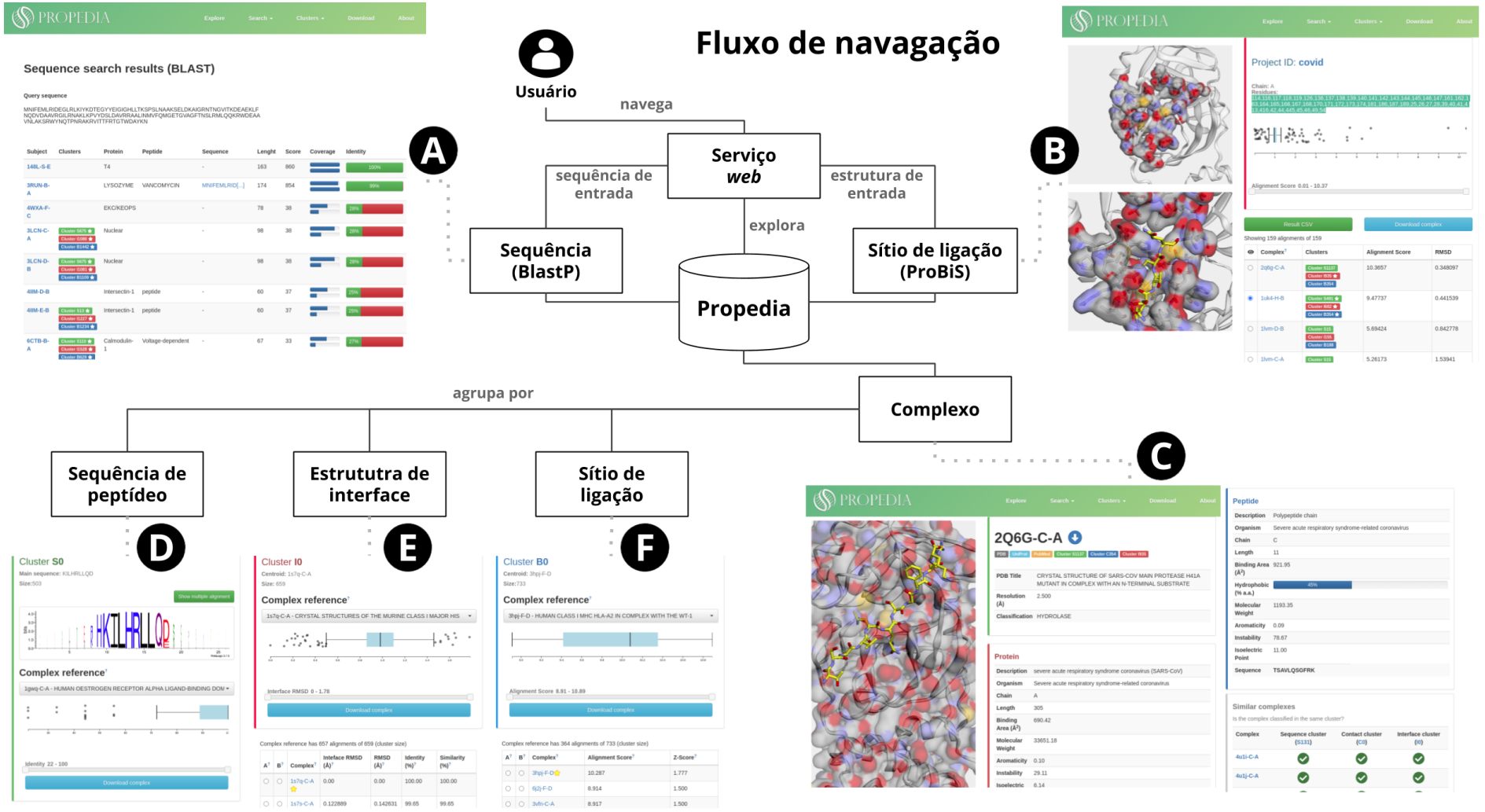


Figura 26 – Fluxo de navegação do Propedia utilizando o serviço *web*. A: Página de resultado de busca por sequência (BLASTp). B: Página de resultado da busca por sítio de ligação (ProBiS). C: Página com informações dos complexos, como características do receptor/peptídeo, grupos e complexos similares. D, E e F: Página dos agrupamentos ASP, AEI e ASL, respectivamente.

4.4 Estudos de caso

4.4.1 Receptores de estrogênio com peptídeos diferentes

Foi realizado um estudo de caso com dois receptores de estrogênio alfa. O primeiro complexo é identificado pelo código PDB 2jf9 (*estrogen receptor alpha ligant-binding domain in complex with a tamoxifen-specific peptide antagonist*). Ele possui em sua cadeia B o receptor de estrogênio alfa interagindo com um peptídeo de tamoxifeno, representado pela cadeia Q. O segundo complexo, de código PDB 4iv2 (*crystal structure of the estrogen receptor alpha ligand-binding domain in complex with dynamic way-derivative*), possui o receptor de estrogênio alfa na cadeia A, e o seu peptídeo coativador 2 de receptor nuclear (*nuclear receptor coactivator 2*) se encontra na cadeia C.

Embora os complexos possuam estrutura de interface similares (iRMSD de 0,74 Å) e estão no grupo I1 (do AEI), seus peptídeos e sítio de ligação são diferentes. O complexo 4iv2-C-A faz parte dos grupos S0 e B1 enquanto o complexo 2jf9-Q-B compõe os grupos S1024 (único) e B838 (também único). A tabela 3 apresenta uma comparação das característica de cada complexo.

		4iv2-C-A	2jf9-Q-B
Proteína	Cadeia	A	B
	Tamanho (aa)	232	210
	Área da Interface ($Å^2$)	484.85	519.3
Peptídeo	Cadeia	C	Q
	Tamanho (aa)	10	13
	Área da Interface ($Å^2$)	559,07	547,74
	Hidrofobicidade (%)	40%	38%
	Peso molecular	1272,5	1539,71
	Aromaticidade	0	0,15
	Instabilidade	95,31	34,72
	Ponto isoelétrico	8,76	5,79
	Sequência	HKILHRLLQD	SPGSREWFKDMLS
Grupos	ASP	S 0	S 1024 (unitário)
	AEI	I 1	I 1
	ASL	B 1	B 838 (unitário)

Tabela 3 – Características de comparação entre 2jf9-Q-B e 4iv2-C-A

Com isso, o alinhamento estrutural de ambos os complexos foi feito manualmente utilizando o PyMol (DELANO et al., 2002) e a comparação dos resultados são vistos na Figura 27. Na Figura 27(A) temos a sobreposição do alinhamento estrutural entre 2jf9-Q-B (branco) e 4iv2-C-A (azul). Os resíduos de aminoácidos da região do receptor foram conservados (*sticks* maiores), mas as sequências de aminoácidos dos peptídeos não (*sticks* menores e *cartoon*). Na Figura 27(B) e (C), temos o receptores de estrogênio alfa separados (não sobrepostos) com seus respectivos peptídeos. É possível observar que, embora as sequências dos peptídeos sejam

diferentes, a conformação estrutural em alfa-hélice dos mesmos permanece similar. Isso não foi o suficiente para que os sítios de ligação fossem semelhantes (mesmo grupo), embora o iRMSD entre os complexos seja de 0,75 Å.

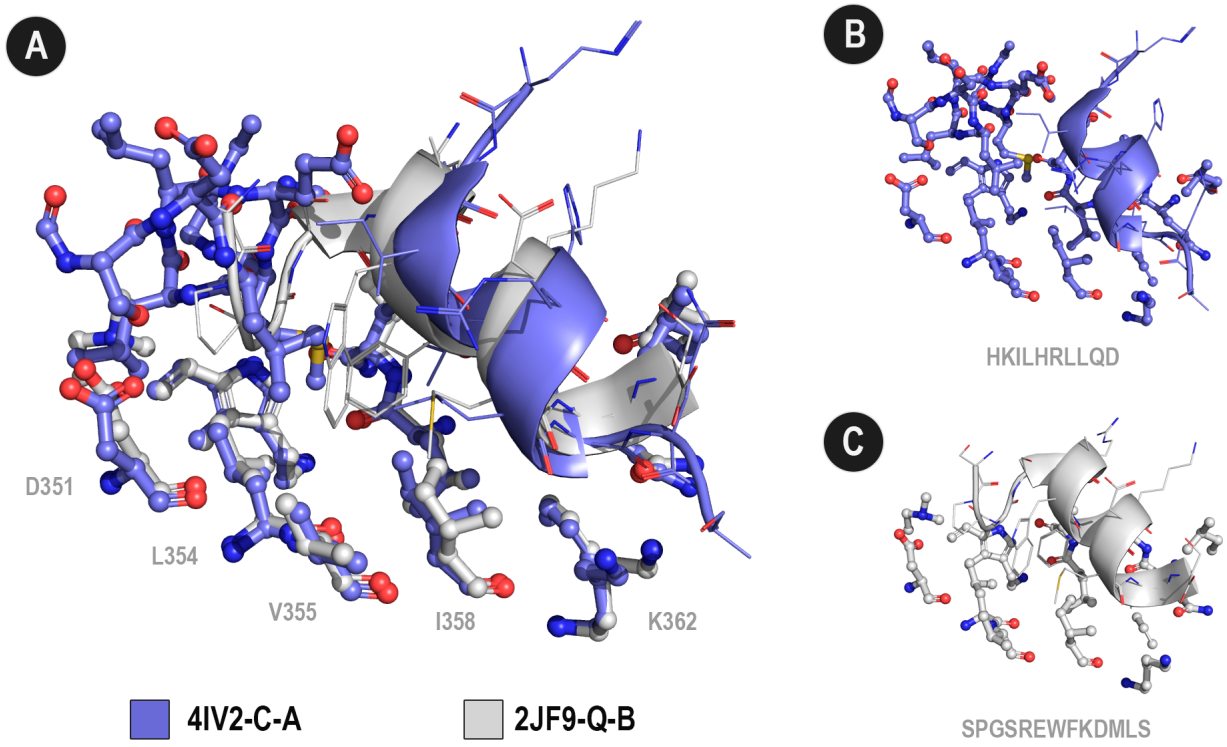


Figura 27 – Comparação estrutural e de interação entre os complexos 2jf9-B-Q e 4iv2-A-C.

4.4.2 Interações da principal protease do SARS-CoV-2 com peptídeos (6lu7)

O novo coronavírus SARS-CoV-2 é o agente etiológico responsável pelo surto de pneumonia viral da doença de coronavírus 2019 (COVID-19) que se espalhou por todo o mundo em 2019-2020 (ZHU et al., 2020; LI et al., 2020; WU et al., 2020a; ZHOU et al., 2020). Atualmente, não existem agentes terapêuticos direcionados para o tratamento desta doença, e as opções de tratamento eficazes permanecem muito limitadas. Por esse motivo, este estudo de caso visa utilizar o recursos de busca desenvolvido e disponibilizado pelo serviço *web* para encontrar estruturas de peptídeos que possam ser promissoras para interagir e consequentemente inibir a protease principal do SARS-CoV-2.

A partir da estrutura da principal protease (Mpro) do SARS-CoV-2 (PDB id: 6lu7), foi realizada uma busca por sítios de ligação similares, utilizando o escopo de pesquisa CCA e definindo os resíduos da cadeia A (protease) que estão a 6 Å do inibidor N3 (cadeia C). Os detalhes dos dados de entrada, como estrutura da protease (cadeia A) e resíduos selecionados,

bem como os resultados da busca estão disponíveis para consulta na página do Propedia (bioinfo.dcc.ufmg.br/propedia/search/binding/covid). Ao todo, 159 complexos similares foram encontrados. Considerando as 10 melhores métricas de alinhamento (limite inferior de 3,18), a Tabela 4 apresenta dados relativos a estes complexos e seus respectivos peptídeos.

PDB id	Descrição	Cadeia receptor	Cadeia peptídeo	Sequência do peptídeo
2q6g	SARS-CoV main protease H41A mutant	A	C	-TSAVLQSGFRK
1uk4	SARS-CoV main proteinase	B	H	--NSTLQ-----
1lvm	Thermotoga maritima methyltransferase	B	D	-ENLYFQ-----
1lvm	Thermotoga maritima methyltransferase	A	C	-ENLYFQ-----
3mmg	Tobacco vein mottling virus protease	A	C	-ETVRFQS----
1lvb	Tobacco etch virus protease	B	D	TENLYFQSGT--
1lvb	Tobacco etch virus protease	A	C	TENLYFQSGT--
6hgj	SARS-CoV main protease variant NewBG-III	A	B	*
3caa	Cleaved antichymotrypsin A347R	A	B	*
5om5	Human alpha1-antichymotrypsin	A	B	-TSAVLQSGFR-

Tabela 4 – Lista dos complexos recuperados a partir da busca dos 10 melhores resultados encontrados tendo como alvo a estrutura da Mpro do SARS-CoV-2 (PDB id: 6lu7). *Sequência omitida devida a seu tamanho.

A busca por sítio de ligação foi capaz de recuperar outras 3 proteases de SARS-CoV (PDB id: 2q6g, 1uk4 e 6hgj) e outras proteases virais (PDB id: 3mmg e 1lvb) junto com peptídeos que possam ter potencial para inibir a protease do SARS-CoV-2. A partir desses resultados, foi realizado experimentos de ancoragem (*docking*) molecular com os peptídeos retornados pela busca utilizando o protocolo do Rosetta FlexPepDock (RAVEH; LONDON; SCHUELER-FURMAN, 2010). O Rosetta FlexPepDock é capaz de definir estruturas de alta resolução a partir do modelo aproximado do peptídeo em interação com o sitio de ligação do receptor, permitindo a flexibilização das cadeias principais e laterais dos resíduos dos peptídeos. Assim, para preparar os dados de entrada para execução do FlexPepDock, foi realizada a sobreposição estrutural de cada complexo (10 recuperados) com a protease SARS-CoV-2 (PDB id: 6lu7). Esse procedimento foi bem-sucedido em 8 complexos, nos quais a protease SARS-CoV-2 foi adequadamente sobreposta aos receptores considerando o RMSD com limite igual ou inferior a 3 Å. As cadeias dos receptores dos complexos foram então removidas, deixando apenas os peptídeos, com suas posições espaciais agora relativas ao sítio da protease SARS-CoV-2. Para os dois complexos cujos os receptores não se alinharam bem devido a alta dissimilaridade estrutural com a protease SARS-CoV-2, foi utilizado o *software* HADDOCOK (ZUNDERT et al., 2016) para execução de uma ancoragem cega (*blind docking*) considerando os peptídeos destes complexos sobre a superfície da protease SARS-CoV-2. Assim, foram selecionadas as melhores posições (considerando os valores mais negativos da métrica do HADDOCK) dos peptídeos para serem submetidos como estrutura inicial no processo de ancoragem molecular, juntamente com os 8 peptídeos anteriores.

De posse dos peptídeos e suas posições espaciais relativas a protease SARS-CoV-2, deu-se início a etapa de execução do FlexPepDock. Devido a limitação quanto ao tamanho dos peptídeos (máximo 30 resíduos) considerada pelo FlexPepDock, foi necessário descartar os peptídeos 6hgj:B e 3caa:B do processo. Consequentemente, 8 modelos foram obtidos pela ancoragem molecular e todos resultaram em uma considerável afinidade com a protease SARS-CoV-2 (Tabela 5, coluna "Métrica Rosetta"). Além de métricas aceitáveis, verificamos posições aparentemente adequadas (Figura 28) de cada peptídeo, considerando a proximidade ao átomo de enxofre do resíduo Cis145, que junto com a His41, que compõem a díade catalítica da protease SARS-CoV-2 (QAMAR et al., 2020). Os resultados são apresentados na Tabela 5. A coluna "Distância P1-Cis145" apresenta a distância entre o átomo de enxofre do resíduo Cis145 da protease SARS-CoV-2 e o Cα do resíduo P1 da matriz de especificidade gerada pela base de dados de enzimas proteolíticas MEROPS. (RAWLINGS; BARRETT; BATEMAN, 2010).

		Propedia		FlexPepDock		
PDB id	Cadeia receptor	Métrica alinhamento	RMSD (sítio)	Métrica Rosetta	RMSD receptor	Distância P1-Cis-145 (Å)
2q6g	A	10,36	0,34	-542,507	0,997	3,6
1uk4	B	9,47	0,44	-525,999	0,659	3,6
1lvm	B	5,69	0,84	-525,907	2,175	4,0
1lvm	A	5,26	1,53	-528,398	2,085	5,5
3mmg	A	4,67	0,47	-530,833	1,785	3,7
1lvb	B	4,54	1,21	-530,517	2,521	3,6
1lvb	A	4,53	1,22	-538,306	2,546	3,5
6hgj	A	3,38	2,20	-	11,156	-
3caa	A	3,25	2,16	-	9,943	-
5om5	A	3,18	1,63	-538,985	6,665	3,7

Tabela 5 – Resultados do FlexPepDock em contraste com os resultados do Propedia, considerando a Mpro do SARS-CoV-2 como alvo, comparada com os receptores dos complexos com sítio de ligação similares.

De fato, de acordo com o MEROPS (RAWLINGS; BARRETT; BATEMAN, 2010) e Goetz et al. (2007), as principais proteases do coronavírus mostram preferência por substratos da forma geral: P4=V/T/A/S P3=V/W/K P2=L P1=H/Q. Essas posições são representadas em tons de azul na Figura 29. De acordo com esses trabalhos anteriores, o sítio P1 está bem conservado, mas os outros são muito mutáveis. Os peptídeos identificados pelo Propedia possuem resíduos destacados em caixas amarelas. Os resultados deste estudo de caso está acessível no endereço eletrônico: <http://bioinfo.dcc.ufmg.br/propediadb/search/binding/covid>. Observa-se que este conjunto de peptídeos é geralmente consistente com peptídeos conhecidos por inibir as proteases do coronavírus.

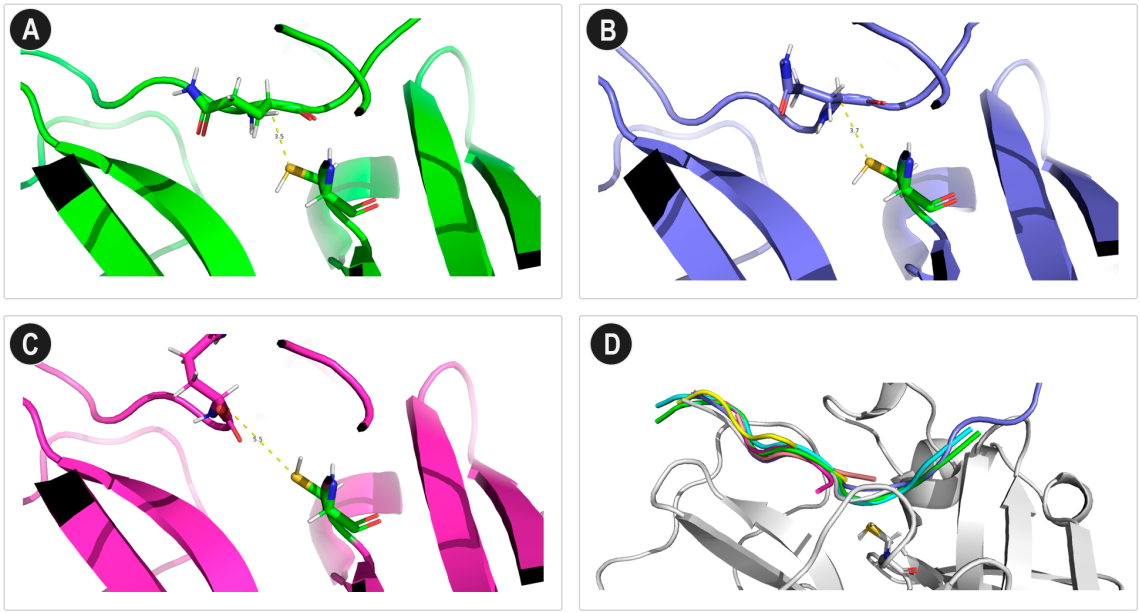


Figura 28 – Poses das conformações estruturais de peptídeos geradas pelo FlexPepDock. (A) PDB id: 1lvb; cadeia do peptídeo: C; cadeia do receptor: A; Métrica Rosetta: -538.306; distância: 3,5 (B) PDB id: 5om5; cadeia do peptídeo: B; cadeia do receptor: A; Métrica do Rosetta: -538.985; distância: 3,7 (C) PDB id: 1lvm; cadeia do peptídeo: C; cadeia do receptor: A; Métrica Rosetta: -528.398; distância: 5.5 (D) Todo o conjunto de peptídeos avaliados.

4.4.3 Metadinâmica aplicada aos os peptídeos de complexos com sítio de ligação similares a principal protease do SARS-CoV-2

A aplicação de processos de metadinâmica (MetaD), realizados em triplicatas, visa estimar os valores de energia livre de ligação (ΔG_{bind}) necessário para desassociar (*unbinding*) os peptídeos do sítio de ligação da Mpro do SARS-CoV-2. Para isso, foram selecionados os peptídeos dos quatro melhores receptores (Tabela 5), sendo eles: 2q6g, cadeia C; 1uk4, cadeia H; 1lvm, cadeia D; e 1lvb, cadeia D. Os panoramas de energia livre para os respectivos processos estão representados na Figura 33 (Apêndice A), mapeados em suas triplicatas de 1-3, 4-6, 7-9 e 10-12, respectivamente. Em cada sistema, os mínimos dentro da proteína (A) e no meio aquoso (B) puderam ser caracterizados com precisão suficiente para estimar o ΔG_{bind} de acordo com as Equações A.1, A.2 e A.3. Os detalhes sobre o protocolo utilizado para a realização da estão descritos no Apêndice A.

Uma convergência significativa foi obtida para o valores de ΔG_{bind} retornados da MetaD para cada sistema no protocolo aplicado, com desvio padrão máximo de 1,57 kcal·mol^{-1} para os sistemas 1lvm (cadeia D), 1lvb (cadeia D) e 2q6g (cadeia C) e um desvio relativamente maior de 4,32 kcal·mol^{-1} apenas para 1uk4 (cadeia H). Na verdade, tal convergência não é surpreendente, uma vez que a situação já consolidada da técnica de metadinâmica como uma ferramenta computacional precisa para estimar ΔG_{bind}, tanto para ligantes e peptídeos. Além

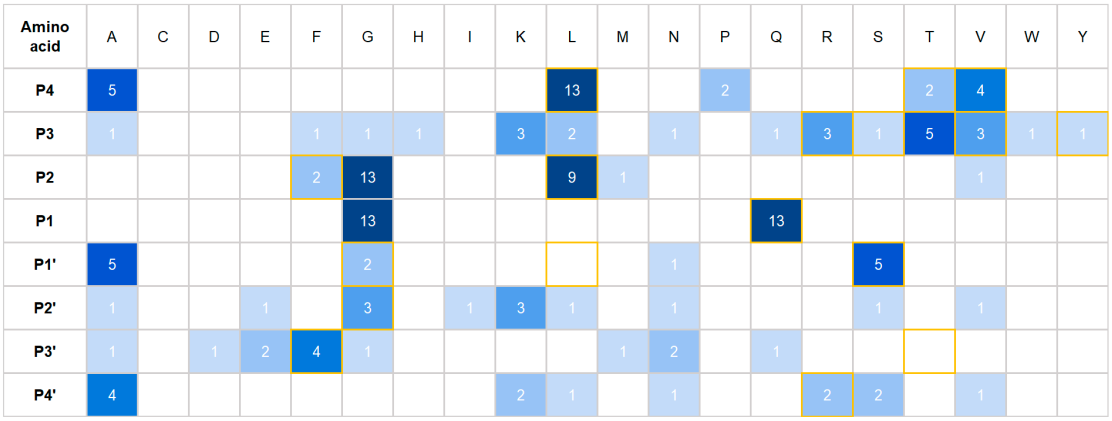


Figura 29 – Matriz de especificidade do MEROPS. As especificidades dos resíduos estão destacados em tons de azul. Resíduos dos peptídeos sugeridos pelo Propedia estão destacados em caixas amarelas.

disso, este método é amplamente utilizado em procedimentos de triagem de drogas (CAVALLI et al., 2015; SÖLDNER; HORN; STICHT, 2019; BRANDT et al., 2016). A precisão desta técnica, desta forma, torna-se um instrumento providencial para validar o Propedia na triagem de peptídeos com afinidades diferenciais para a Mpro do SARS-CoV-2, dada a ainda escassa disponibilidade de dados experimentais de afinidade peptídica neste alvo novo e importante. Desta forma, foi realizada a correlação entre da ΔG_{bind} retornado pelo método de alto desempenho da metadinâmica e as métricas recuperadas do Propedia, visando a validação desta ferramenta computacional.

É notório pela Figura 30 (detalhes na Tabela 10), a correlação negativa significativa da ΔG_{bind} da MetaD com a métrica de alinhamento retornado pelo Propedia (R^2 of 0,98) e a correlação positiva do RMSD Å de sítio de ligação (R^2 de 0,96). Ainda, é digno de nota que tanto as pontuações do resultado do Propedia quanto a ΔG_{bind} da MetaD colocam o substrato específico conhecido da MPro (PDB: 2q6g) e seu inibidor análogo ao substrato (1uk4) no topo da classificação de afinidade com esta proteína. Desta forma, tanto a correlação com os resultados do método de metadinâmica de alto desempenho, quanto a consistência com dados funcionais conhecidos podem ser tomados em conjunto como um indicativo de validação utilizando o Propedia, bem como sua aplicabilidade na triagem para peptídeos funcionais para este e outros alvos importantes.

4.4.4 Protease da lagarta-da-soja (*Anticarsia gemmatalis* Hübner)

A lagarta-da-soja, *Anticarsia gemmatalis* Hübner (AG) Hübner, é uma das principais pragas desfolhadoras nas Américas, afetando principalmente as lavouras de soja e uma das principais causas de perdas econômicas na agricultura (VIANNA et al., 2011; MOSCARDI et al., 2012). Nos últimos anos, alternativas para o controle de pragas, como o desenvolvimento de

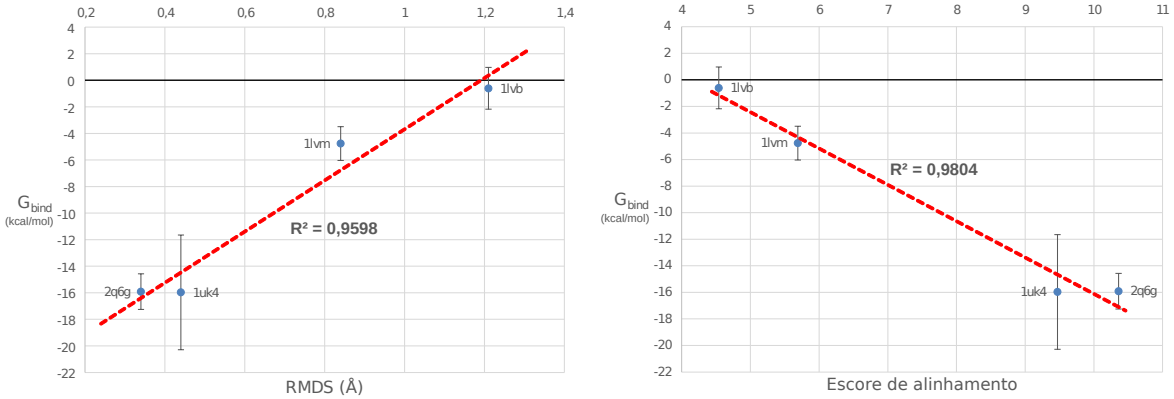


Figura 30 – Correlação da ΔG_{bind} da MetaD com o RMSD do sítio de ligação (a esquerda) e métrica de alinhamento (direita) da MPro do SARS-CoV-2 com peptideos sugeridos pelo Propedia: 2q6g (cadeia C), 1uk4 (cadeia H), 1lvm (cadeia D), e 1lvb (cadeia D)

biopesticidas, têm sido exploradas. Por exemplo, o uso de inibidores de protease é altamente considerado no manejo de pragas de insetos, pois afeta a biodisponibilidade de aminoácidos essenciais, o que impede o crescimento de larvas e o desenvolvimento de insetos para várias espécies, como apresentam os trabalhos de Moreira et al. (2011) e Pilon et al. (2018). Devido a este problema, este estudo de caso visa procurar um peptídeo na base de dados que possa ser promissor para inibir a protease da AG. Foi utilizado a sequência da serino-protease (*trypsin-like serine protease*) extraída do intestino da AG, depositada no GenBank ((BENSON et al., 2005)) com identificador: JX898746.1 ((PILON; OLIVEIRA, 2013)). Além disso, utilizando o I-TASSER (YANG et al., 2015), modelos da estrutura da serino protease da AG foram criados a partir de sua sequência. Alinhamentos estruturais foram realizados sobre os modelos gerados afim de identificar os resíduos altamente conservados da tríade catalítica na protease (PERONA; CRAIK, 1995). Esses resíduos foram identificados como His6, Asp56 e Ser143 no modelo. A Figura 34 (Apêndice B) mostra a superposição das estruturas onde os resíduos da tríade são destacados.

Consultas foram realizadas através do serviço *web* do Propedia, utilizando a sequência da serino protease da AG como parâmetro de busca por sequência, e a estrutura do modelo, selecionando os resíduos da tríade catalítica, para busca por sítio de ligação em dois experimentos separados. Os 10 principais resultados em cada um deles foram selecionados e os resultados são apresentados nas Tabelas 6 e 7, respectivamente. Em seguida, experimentos de ancoragem molecular utilizando apenas o HADDOCK, uma vez que foi encontrado um número considerável de peptídeos que continham resíduos não naturais.. Além disso, para o experimento de consulta por sítio de ligação, duas entradas de peptídeos dos resultados do Propedia foram descartados (que não são listadas na Tabela 7): 1p11, cadeia P, devido ao seu formato não ser suportado por HADDOCK; e 3kf2, cadeia D, devido à alta similaridade de sequência com o peptídeo 3kf2, cadeia D.

Para o conjunto de dados baseado em sequência, definimos os resíduos da tríade catalítica

PDB id	Descrição	Cadeia receptor	Cadeia peptídeo
1ekb	Enteroquinase bovina	B	A
1ekb	Enteroquinase bovina	B	C
2stb	Tripsina do salmão-do-atlântico	I	E
2sta	Tripsina do salmão-do-atlântico	I	E
3qgn	Trombina humana	A	B
2zdv	Trombina humana	L	H
1ca8	Trombina humana	A	B
1ca8	Trombina humana	A	B
4dii	Trombina humana	L	H
4dih	Trombina humana	L	H
4lz1	Trombina humana	B	A

Tabela 6 – Lista dos peptídeos recuperados pela busca por sequência (receptor) utilizando sequência serino protease da AG

PDB id	Descrição	Cadeia peptídeo	Cadeia receptor
3qgj	Lysobacter enzymogenes protease	D	C
1p11	Lysobacter enzymogenes protease	I	E
2obq	Hepacivirus NS3-4A protease	B	C
2oin	Hepacivirus NS3-4A protease R155K	C	A
2o8m	Hepacivirus NS3-4A protease S139A	D	B
3kn2	Hepacivirus NS3-4A protease	B	C
3kf2	Hepacivirus NS3-4A protease	C	A
3sga	Streptomyces griseus protease	P	E
6rw2	Human Ephrin type-A receptor 2	B	A
4a1t	Hepacivirus NS3-4A protease	D	B

Tabela 7 – Lista dos peptídeos recuperados pela busca por sítio de ligação utilizando o modelo gerado pela sequência serino protease da AG e os resíduos da tríade catalítica (His6, Asp56 e Ser143)

como resíduos ativos para o procedimento da ancoragem molecular, bem como a cadeia completa de cada peptídeo. Foram selecionados as melhores estruturas resultantes de acordo com a métrica HADDOCK (mais negativa) e, em seguida, de acordo o seu RMSD ($\leq$ 3 Å) de cada estrutura em relação ao modelo geral de menor energia. A Tabela 8 resume os resultados. Os cinco melhores poses dos peptídeo no modelo da serino protease da AG foram selecionadas de acordo com a métrica HADDOCK, para o qual foram identificados os resíduos mais próximos do resíduo Ser143 no sítio S1 (região ao qual P1 interage), considerando a distância entre os Cα. Os resíduos mais próximos encontrados foram os resíduos de cisteína localizados nos modelos 3qgn, cadeia A (3,9 Å), 4dii, cadeia L (4,4 Å) e 1ca8, cadeia A (5,1 Å). A presença de resíduos de cisteína próximos à serina na tríade catalítica indica um uso potencial do peptídeo como um inibidor, uma vez que substratos com esses resíduos na posição P1 não são geralmente clivados por serina proteases semelhantes à tripsina (PAGE; CERA, 2008). A Figura 31 mostra os modelos 3qgn, cadeia A e 4dii, cadeia L, onde a distância entre os resíduos é destacada.

PDB id	Cadeia peptídeo	RMSD HADDOCK	Métrica HADDOCK	Resíduos próximos de S1
1ekb	C	2.564	-49.202	-
1ekb	A	2.499	-63.477	-
2stb	I	0.000	-86.803	-
2sta	I	4.275	-81.387	-
3qgn	A	0.000	-97.979	CYS (3.9 Å)
2zdv	L	0.000	-100.560	GLU (7.4 Å)
1ca8	A	1.530	-102.975	CYS (5.1 Å)
1ca8	C	1.158	-75.138	-
4dii	L	2.598	-95.836	CYS (4.4 Å)
4dih	L	1.508	-95.317	ARG (5.4 Å)

Tabela 8 – Métrica e RMSD do HADDOCK para os modelos selecionados para cada cadeia de peptídeo no experimento baseado em sequência

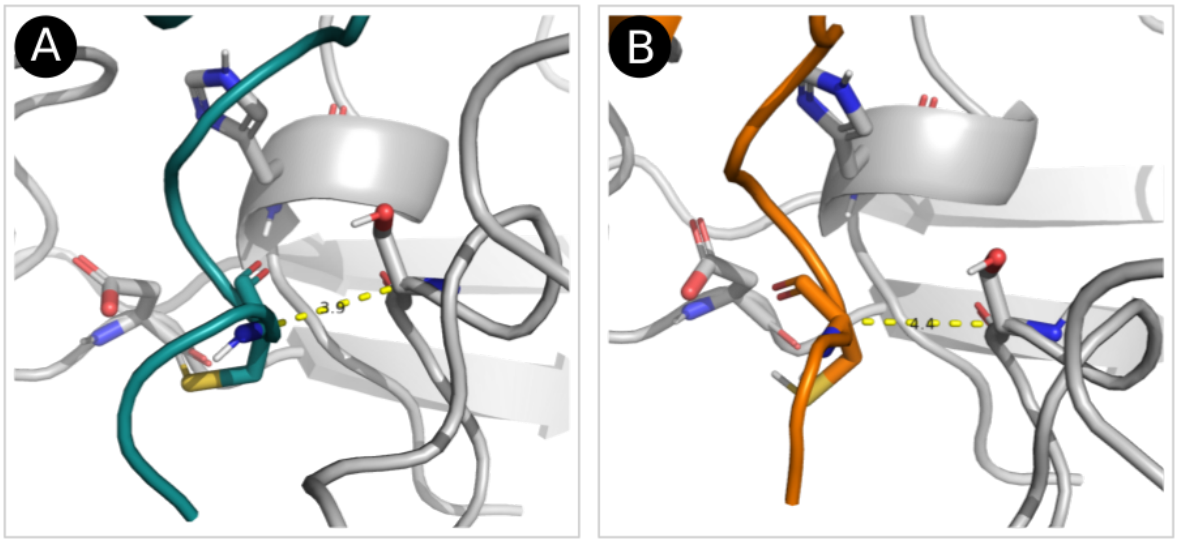


Figura 31 – Modelo da serino protease de AG (em cinza) acoplado aos peptídeos 3qgn, cadeia A (A) e 4dii, cadeia L (B). A distância entre o resíduo Ser143 da região S1 na protease para os resíduos de cisteína nos peptídeos são 3,9 Å e 4,4 Å, respectivamente.

Semelhante ao conjunto de dados baseado em sequência, foi realizada a ancoragem molecular o conjunto de dados do sítio de ligação usando os resíduos da tríade catalítica, bem como cadeias peptídicas completas como resíduos ativos. A assinatura do sítio de ligação representa a forma com que uma proteína interage com seu ligante e quais aminoácidos são essenciais para manter o complexo estável. Uma métrica adequada para verificar a semelhança dos locais de ligação é a fração de contatos comuns (FCC). O FCCAB é a proporção de contatos entre as estruturas A e B para todos os contatos em A, cujo valor varia de 0, quando as cadeias não compartilham contatos, a no máximo 1, quando todos os contatos de A estão em B Rodrigues et al. (2012). Portanto, para os experimentos de ancoragem molecular do sítio de ligaão, um valor médio mais alto de FCC em um grupo indica maior similaridade das interações entre diferentes

poses de peptídeo e o modelo da serino protease da AG, o que também significa que o sítio de ligação é mais conservado.

Para cada peptídeo, foi selecionado o grupo com a pontuação FCC mais alta (em relação aos seus modelos de energia mais baixos produzidos pelo HADDOCK), a partir dos quais escolhemos os melhores modelos de acordo com a métrica HADDOCK. Os valores FCC e as métricas do HADDOCK são mostrados na Tabela 9 para todos os peptídeos. Os quatro melhores modelos para cada um dos 3 principais grupos são mostrados na Figura 32. Em todos os modelos, os contatos estão centrados na tríade catalítica (destacada em vermelho), enquanto as áreas de contato restantes se ligam a diferentes ligantes, onde resíduos vizinhos no sítio de ligação da serino protease têm grande relevância por estabelecer ligações hidrofóbicas e de hidrogênio. O mapa de interação completo de cada modelo está disponível na Figura 35. Isso enfatiza a importância do uso de FCC como uma métrica adequada para análise de sítio de ligação em vez de RMSD, e também demonstra a precisão do Propedia na determinação de padrões de sítio de ligação em relação à especificidade do ligante.

PDB id	Cadeia do peptídeo	FCC	Menor métrica HADDOCK
6rw2	A	0.883	-75.354
3kn2	B	0.840	-78.998
2obq	B	0.833	-80.258
2oin	C	0.696	-85.060
3sga	P	0.650	-61.368
4a1t	D	0.648	-87.944
1p11	I	0.621	-49.383
3qgj	D	0.409	-34.823
3kf2	A	0.236	-55.432
2o8m	D	0.196	-54.616

Tabela 9 – Métrica do HADDOCK e FCC para os modelos selecionados para cada cadeia de peptídeo no experimento de sítio de ligação

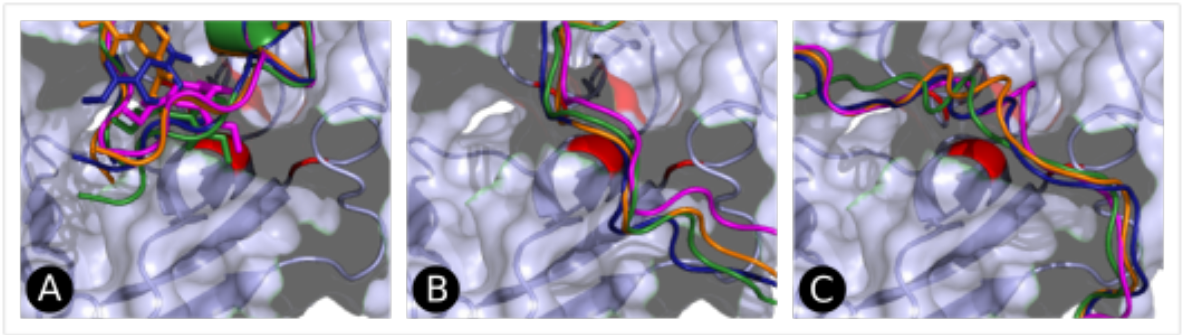


Figura 32 – Modelo da serino protease de AG (em cinza) juntamente com as 4 primeiras posições dos peptídeos 6rw2, cadeia B (A), 3kn2, cadeia B (B) e 2obq, cadeia B (C). Os resíduos em vermelho representam os resíduos catalíticos da tríade catalítica.

Esses estudos de caso mostram a utilidade do Propedia como uma base de dados e ferramenta que permite recuperar peptídeos e seus complexos de interesse. Acreditamos que os complexos existentes e a compreensão dos padrões de interação proteína-peptídeo possam ser muito relevantes na evolução dirigida de peptídeos bem como no desenho de novas moléculas peptidomiméticos.

5 Conclusões

Neste trabalho foi proposta uma base de dados de complexos proteína-peptídeo e uma ferramenta de busca e análise dos dados nela contidos. Ao todo, 19.813 complexos podem ser encontrados, sendo 5.971 classificados como agrupáveis (compostos apenas por resíduos padrão e ligados a apenas um receptor). Deste escopo, podem ser extraídos sem redundância, reduzindo para: 1.845, quando se leva em consideração a sequência do peptídeos; 1.891, para estruturas de interface proteína-peptídeo diferentes; e por fim, 1.812, considerando sítio de ligação como aspecto de similaridade.

O serviço *web* foi desenvolvido para servir de interface, de forma simples e interativa, com funcionalidades para recuperação, exploração e visualização dos complexos. Além disso, o serviço dispõe de funcionalidades que permitem que os usuários possam submeter suas próprias sequências, a fim de encontrar peptídeos/receptores similares e também a submissão de estruturas (formato PDB) quando deseja recuperar sítios de ligações similares, afim de sugerir peptídeos promissores para interação molecular. O Propedia pode ser acessível pelo endereço eletrônico: <http://bioinfo.dcc.ufmg.br/propediadb> e todas suas informações estão disponíveis para *download* (incluindo a base de dados completa).

Espera-se que o base de dados Propedia, junto com o serviço *web* possa servir a comunidade científica como um recurso útil para recuperação de estruturas de complexos proteína-peptídeo e peptídeos promissores para alvos proteicos de interesse.

6 Perspectivas

O Propedia (base de dados e serviço *web*) foi projetado para possuir um aspecto dinâmico, e por isso atualizações serão realizadas periodicamente, permitindo a adição de novos dados e funcionalidades.

A base de dados foi criada a partir da recuperação dos arquivos de estrutura PDB, realizada em 2 de Maio de 2020. Pretende-se realizar uma nova extração, programada para Dezembro de 2020 (9 meses), para avaliação do trabalho, e posteriormente esse intervalo será reduzido, onde a extração será feita trimestralmente.

Quanto ao serviço *web*, novas versões serão lançadas para aperfeiçoar a sua interatividade, considerando o *feedback* dado por seus usuários.

A partir dos dados do Propedia pode-se derivar novos aspectos, como uma biblioteca de fragmentos de proteína-peptídeo, ao qual pretende-se desenvolver futuramente. Tais fragmentos podem ser feitos com um número menor de resíduos, considerando o peptídeo. Para exemplificar, considerado fragmentos da interface do complexos com três resíduos do peptídeos, a partir do escopo CCA, seria possível montar um biblioteca com 74.653 fragmentos. Obviamente, seria aconselhável aplicar métodos de agrupamentos para remover a redundância, afim de reduzir esta quantidade. Além disso, com este montante de dados seria muito bem-vinda a aplicação de técnicas de aprendizado de máquina para o desenvolvimento de um modelo de predição de peptídeos e evolução dirigida. Os fragmentos poderiam funcionar como peças que auxiliariam o desenho de novos peptídeos a partir de um alvo. Utilizando ferramentas de ancoragem molecular (por exemplo DOCK 6 (ALLEN et al., 2015)), os usuários poderão desenhar péptides a partir da interface *web*, sendo orientados por sugestões do modelo de aprendizado, auxiliando ainda na sugestão de resíduo adicionar ao longo da cadeia polipeptídica, tendo como objetivo a redução de energia de livre. Com isso, toda a complexidade por trás desse processo, ficaria a cargo do servidor (*backend*), deixando o usuário livre para se preocupar apenas em encontrar um bom modelo de peptídeo para o seu alvo proteico a partir dos resultados apresentados pelo serviço.

Referências

AGRAWAL, P. et al. Cppsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D1098–D1103, 2016. Citado na página 19.

ALLEN, W. J. et al. Dock 6: Impact of new features and current docking performance. *Journal of computational chemistry*, Wiley Online Library, v. 36, n. 15, p. 1132–1156, 2015. Citado na página 68.

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990. Citado na página 43.

ANGELOVA, A. et al. Pep-lipid cubosomes and vesicles compartmentalized by micelles from self-assembly of multiple neuroprotective building blocks including a large peptide hormone pacap-dha. *ChemNanoMat*, Wiley Online Library, v. 5, n. 11, p. 1381–1389, 2019. Citado 2 vezes nas páginas 18 e 19.

AO, G. P. P. et al. Biochemical responses of anticarsia gemmatalis (lepidoptera: Noctuidae) in soybean cultivars sprayed with the protease inhibitor berenil. *Journal of agricultural and food chemistry*, ACS Publications, v. 61, n. 34, p. 8034–8038, 2013. Citado na página 21.

BARDUCCI, A.; BONOMI, M.; PARRINELLO, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, Wiley Online Library, v. 1, n. 5, p. 826–843, 2011. Citado na página 78.

BENSON, D. A. et al. Genbank. *Nucleic acids research*, Oxford University Press, v. 33, n. suppl_1, p. D34–D38, 2005. Citado na página 62.

BERG, J. M. *Biochemistry 5th Edition*. 2006. Citado na página 16.

BERMAN, H. et al. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic acids research*, Oxford University Press, v. 35, n. suppl_1, p. D301–D303, 2006. Citado na página 18.

BEST, R. B. et al. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation*, ACS Publications, v. 8, n. 9, p. 3257–3273, 2012. Citado na página 78.

BICKERTON, G. R.; HIGUERUELO, A. P.; BLUNDELL, T. L. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. *BMC bioinformatics*, Springer, v. 12, n. 1, p. 313, 2011. Citado na página 27.

BRANDT, A. M. et al. Exploring the unbinding of l eishmania (l.) amazonensis cpb derived-epitopes from h 2 mhc class i proteins. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 84, n. 4, p. 473–487, 2016. Citado 4 vezes nas páginas 61, 78, 80 e 81.

BROADWAY, R. M. Are insects resistant to plant proteinase inhibitors? *Journal of Insect Physiology*, Elsevier, v. 41, n. 2, p. 107–116, 1995. Citado na página 21.

BUSSI, G.; LAIO, A.; PARRINELLO, M. Equilibrium free energies from nonequilibrium metadynamics. *Physical review letters*, APS, v. 96, n. 9, p. 090601, 2006. Citado na página 78.

CAMACHO, C. et al. Blast+: architecture and applications. *BMC bioinformatics*, Springer, v. 10, n. 1, p. 421, 2009. Citado na página 43.

CAVALLI, A. et al. Investigating drug–target association and dissociation mechanisms using metadynamics-based algorithms. *Accounts of chemical research*, ACS Publications, v. 48, n. 2, p. 277–285, 2015. Citado na página 61.

COCK, P. J. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, Oxford University Press, v. 25, n. 11, p. 1422–1423, 2009. Citado na página 27.

CONSORTIUM, U. Uniprot: a hub for protein information. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D204–D212, 2015. Citado na página 53.

CRAIK, D. J. et al. The future of peptide-based drugs. *Chemical biology & drug design*, Wiley Online Library, v. 81, n. 1, p. 136–147, 2013. Citado 2 vezes nas páginas 18 e 19.

CROOKS, G. E. et al. Weblogo: a sequence logo generator. *Genome research*, Cold Spring Harbor Lab, v. 14, n. 6, p. 1188–1190, 2004. Citado na página 33.

DAS, A. A. et al. Pepbind: a comprehensive database and computational tool for analysis of protein-peptide interactions. *Genomics, proteomics & bioinformatics*, Elsevier, v. 11, n. 4, p. 241–246, 2013. Citado 3 vezes nas páginas 18, 19 e 44.

DELANO, W. L. et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, v. 40, n. 1, p. 82–92, 2002. Citado na página 56.

DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. *science*, American Association for the Advancement of Science, v. 338, n. 6110, p. 1042–1046, 2012. Citado na página 16.

FALQUET, L. et al. The prosite database, its status in 2002. *Nucleic acids research*, Oxford University Press, v. 30, n. 1, p. 235–238, 2002. Citado na página 18.

FASSIO, A. V. et al. napoli: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, 2019. Citado na página 27.

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, Oxford University Press, v. 39, n. suppl_2, p. W29–W37, 2011. Citado na página 31.

FRAPPIER, V.; DURAN, M.; KEATING, A. E. Pixeldb: Protein-peptide complexes annotated with structural conservation of the peptide binding mode. *Protein Science*, Wiley Online Library, v. 27, n. 1, p. 276–285, 2018. Citado 4 vezes nas páginas 18, 19, 20 e 44.

GAUTAM, A. et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D444–D449, 2014. Citado na página 19.

GAUTAM, A. et al. Cppsite: a curated database of cell penetrating peptides. *Database*, Narnia, v. 2012, 2012. Citado na página 19.

GOETZ, D. et al. Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the sars coronavirus. *Biochemistry*, ACS Publications, v. 46, n. 30, p. 8744–8752, 2007. Citado na página 59.

HAGBERG, A.; SWART, P.; CHULT, D. S. *Exploring network structure, dynamics, and function using NetworkX*. [S.l.], 2008. Citado na página 37.

HAMELRYCK, T.; MANDERICK, B. Pdb file parser and structure class implemented in python. *Bioinformatics*, Oxford University Press, v. 19, n. 17, p. 2308–2310, 2003. Citado na página 27.

HUANG, J.; JR, A. D. M. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, Wiley Online Library, v. 34, n. 25, p. 2135–2145, 2013. Citado na página 78.

HUBBARD, S. J.; THORNTON, J. M. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, v. 2, n. 1, 1993. Citado na página 29.

HUMPHREY, W. et al. Vmd: visual molecular dynamics. *Journal of molecular graphics*, Guildford: Butterworth Scientific Limited, c1983-c1996., v. 14, n. 1, p. 33–38, 1996. Citado 2 vezes nas páginas 78 e 80.

JOHANSSON-ÅKHE, I.; MIRABELLO, C.; WALLNER, B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–13, 2019. Citado na página 18.

KARLIN, S.; ALTSCHUL, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 87, n. 6, p. 2264–2268, 1990. Citado na página 40.

KONAGURTHU, A. S. et al. Mustang: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 64, n. 3, p. 559–574, 2006. Citado 3 vezes nas páginas 20, 31 e 33.

KONC, J. et al. Probis-database: precalculated binding site similarities and local pairwise alignments of pdb structures. *Journal of chemical information and modeling*, ACS Publications, v. 52, n. 2, p. 604–612, 2012. Citado na página 42.

KONC, J.; JANEŽIČ, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, Oxford University Press, v. 26, n. 9, p. 1160–1168, 2010. Citado 4 vezes nas páginas 31, 39, 40 e 43.

KREJCI, A. et al. Hammock: a hidden markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinformatics*, Oxford University Press, v. 32, n. 1, p. 9–16, 2016. Citado na página 31.

LAU, J. L.; DUNN, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry*, Elsevier, v. 26, n. 10, p. 2700–2707, 2018. Citado na página 19.

LEE, A. C.-L. et al. A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences*, Multidisciplinary Digital Publishing Institute, v. 20, n. 10, p. 2383, 2019. Citado na página 19.

LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, Elsevier, v. 55, n. 3, p. 379–IN4, 1971. Citado na página 29.

LI, Q. et al. Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *New England Journal of Medicine*, Mass Medical Soc, 2020. Citado na página 57.

LONDON, N.; MOVSHOVITZ-ATTIAS, D.; SCHUELER-FURMAN, O. The structural basis of peptide-protein binding strategies. *Structure*, Elsevier, v. 18, n. 2, p. 188–199, 2010. Citado 3 vezes nas páginas 18, 19 e 44.

MOREIRA, L. et al. Survival and developmental impairment induced by the trypsin inhibitor bis-benzamidine in the velvetbean caterpillar (anticarsia gemmatalis). *Crop Protection*, Elsevier, v. 30, n. 10, p. 1285–1290, 2011. Citado na página 62.

MOSCARDI, F. et al. Soja. manejo integrado de insetos e outros artrópodes-praga. *Artrópodes que atacam as folhas de soja (Hoffmann-Campo CB, Corrêa-Ferreira BS and Moscardi F, eds.). Embrapa, Brasília, DF*, p. 214–334, 2012. Citado na página 61.

NEDUVA, V. et al. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, Public Library of Science, v. 3, n. 12, p. e405, 2005. Citado na página 18.

NOGUCHI, H. et al. Hidden markov model-based prediction of antigenic peptides that interact with mhc class ii molecules. *Journal of bioscience and bioengineering*, Elsevier, v. 94, n. 3, p. 264–270, 2002. Citado na página 31.

OBENAUER, J. C.; CANTLEY, L. C.; YAFFE, M. B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research*, Oxford University Press, v. 31, n. 13, p. 3635–3641, 2003. Citado na página 18.

OTVOS, L. Peptide-based drug design: here and now. In: *Peptide-based drug design*. [S.l.]: Springer, 2008. p. 1–8. Citado na página 18.

PAGE, M. J.; CERA, E. D. Serine peptidases: classification, structure and function. *Cellular and Molecular Life Sciences*, Springer, v. 65, n. 7-8, p. 1220–1236, 2008. Citado na página 63.

PANT, S. et al. Peptide-like and small-molecule inhibitors against covid-19. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, p. 1–15, 2020. Citado na página 19.

PATARROYO-VARGAS, A. M. et al. Kinetic characterization of anticarsia gemmatalis digestive serine-proteases and the inhibitory effect of synthetic peptides. *Protein and peptide letters*, Bentham Science Publishers, v. 24, n. 11, p. 1040–1047, 2017. Citado na página 21.

PERONA, J. J.; CRAIK, C. S. Structural basis of substrate specificity in the serine proteases. *Protein Science*, Wiley Online Library, v. 4, n. 3, p. 337–360, 1995. Citado na página 62.

PETSALAKI, E.; RUSSELL, R. B. Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology*, Elsevier, v. 19, n. 4, p. 344–350, 2008. Citado na página 18.

PHILLIPS, J. C. et al. Scalable molecular dynamics with namd. *Journal of computational chemistry*, Wiley Online Library, v. 26, n. 16, p. 1781–1802, 2005. Citado 2 vezes nas páginas 78 e 79.

PILON, A. M. et al. Protease inhibitory, insecticidal and deterrent effects of the trypsin-inhibitor benzamidine on the velvetbean caterpillar in soybean. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 90, n. 4, p. 3475–3482, 2018. Citado na página 62.

PILON, F.; OLIVEIRA, M. Genbank - anticarsia gemmatalis serine protease mrna, partial cds. 2013. Disponível em: <https://www.ncbi.nlm.nih.gov/nuccore/JX898746.1>. Citado na página 62.

PILON, F. M. et al. Purification and characterization of trypsin produced by gut bacteria from anticarsia gemmatalis. *Archives of insect biochemistry and physiology*, Wiley Online Library, v. 96, n. 2, p. e21407, 2017. Citado na página 21.

PLAXCO, K. W.; SIMONS, K. T.; BAKER, D. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of molecular biology*, Elsevier, v. 277, n. 4, p. 985–994, 1998. Citado na página 27.

PRICE, D. J.; III, C. L. B. A modified tip3p water potential for simulation with ewald summation. *The Journal of chemical physics*, American Institute of Physics, v. 121, n. 20, p. 10096–10103, 2004. Citado na página 78.

PUNTERVOLL, P. et al. Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic acids research*, Oxford University Press, v. 31, n. 13, p. 3625–3630, 2003. Citado na página 18.

QAMAR, M. T. ul et al. Structural basis of sars-cov-2 3clpro and anti-covid-19 drug discovery from medicinal plants. *Journal of pharmaceutical analysis*, Elsevier, 2020. Citado na página 59.

RAITERI, P. et al. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *The journal of physical chemistry B*, ACS Publications, v. 110, n. 8, p. 3533–3539, 2006. Citado na página 81.

RAVEH, B.; LONDON, N.; SCHUELER-FURMAN, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 78, n. 9, p. 2029–2040, 2010. Citado na página 58.

RAWLINGS, N. D.; BARRETT, A. J.; BATEMAN, A. Merops: the peptidase database. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_1, p. D227–D233, 2010. Citado na página 59.

REGO, N.; KOES, D. 3dmol. js: molecular visualization with webgl. *Bioinformatics*, Oxford University Press, v. 31, n. 8, p. 1322–1324, 2014. Citado na página 43.

RODRIGUES, J. P. et al. Clustering biomolecular complexes by residue contacts similarity. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 80, n. 7, p. 1810–1817, 2012. Citado na página 64.

SEEBER, M. et al. Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics*, Oxford University Press, v. 23, n. 19, p. 2625–2627, 2007. Citado na página 80.

SENIOR, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature*, Nature Publishing Group, v. 577, n. 7792, p. 706–710, 2020. Citado na página 17.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, John Wiley & Sons, Ltd, v. 7, n. 1, 2011. Citado na página 31.

SÖDING, J. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, Oxford University Press, v. 21, n. 7, p. 951–960, 2005. Citado na página 31.

SÖLDNER, C. A.; HORN, A. H.; STICHT, H. A metadynamics-based protocol for the determination of gpcr-ligand binding modes. *International journal of molecular sciences*, Multidisciplinary Digital Publishing Institute, v. 20, n. 8, p. 1970, 2019. Citado na página 61.

TRIDGELL, A.; MACKERRAS, P. et al. The rsync algorithm. The Australian National University, 1996. Citado na página 25.

UPTON, D. *CodeIgniter for rapid php application development*. [S.l.]: Packt Publishing Ltd, 2007. Citado na página 43.

VANHEE, P. et al. Pepx: a structural database of non-redundant protein-peptide complexes. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_1, p. D545–D551, 2010. Citado 5 vezes nas páginas 18, 19, 35, 38 e 44.

VIANNA, U. et al. Espécies e/ou linhagens de trichogramma spp. (hymenoptera: Trochogrammatidae) para o controle de anticarsia gemmatalis (lepidoptera: Noctuidae). *Arquivos do Instituto Biológico*, v. 71, p. 81–87, 03 2011. Citado na página 61.

VINOGRADOV, A. A.; YIN, Y.; SUGA, H. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. *Journal of the American Chemical Society*, ACS Publications, v. 141, n. 10, p. 4167–4181, 2019. Citado na página 19.

WALLACE, A. C.; LASKOWSKI, R. A.; THORNTON, J. M. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering, design and selection*, Oxford University Press, v. 8, n. 2, p. 127–134, 1995. Citado na página 85.

WANG, G.; LI, X.; WANG, Z. Apd3: the antimicrobial peptide database as a tool for research and education. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D1087–D1093, 2016. Citado na página 19.

WANG, J. et al. Strapep: a structure database of bioactive peptides. *Database*, Narnia, v. 2018, 2018. Citado na página 19.

WEN, Z. et al. Pepbdb: a comprehensive structural database of biological peptide-protein interactions. *Bioinformatics*, Oxford University Press, v. 35, n. 1, p. 175–177, 2019. Citado 4 vezes nas páginas 18, 19, 20 e 44.

WU, F. et al. A new coronavirus associated with human respiratory disease in china. *Nature*, Nature Publishing Group, v. 579, n. 7798, p. 265–269, 2020. Citado na página 57.

WU, Z. et al. Signal peptides generated by attention-based neural networks. *ACS Synthetic Biology*, ACS Publications, v. 9, n. 8, p. 2154–2161, 2020. Citado na página 21.

XU, X.; ZOU, X. Peppro: A nonredundant structure data set for benchmarking peptide-protein computational docking. *Journal of Computational Chemistry*, Wiley Online Library, 2020. Citado 4 vezes nas páginas 18, 19, 20 e 44.

YANG, J. et al. The i-tasser suite: protein structure and function prediction. *Nature methods*, Nature Publishing Group, v. 12, n. 1, p. 7, 2015. Citado na página 62.

ZHOU, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, Nature Publishing Group, v. 579, n. 7798, p. 270–273, 2020. Citado na página 57.

ZHU, N. et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, Mass Medical Soc, 2020. Citado na página 57.

ZUNDERT, G. V. et al. The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, Elsevier, v. 428, n. 4, p. 720–725, 2016. Citado na página 58.

Produção Acadêmica

O presente trabalho tornou-se um artigo científico, entitulado "*Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm*" e publicado em 2021.

Além disso, durante o período do doutorado (2015-2020), o autor desta tese publicou 4 artigos em revistas científicas indexadas: "*Vermont: a multi-perspective visual interactive platform for mutational analysis*" (2017); "*How to compute protein residue contacts more accurately?*" (2018); "*Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes*" (2019); "*Introducing programming skills for life science students* (2019)"; sendo um como primeiro autor e dois com mesmo nível de colaboração do autor principal.

Apêndices

APÊNDICE A – Protocolos aplicados a metadinâmica

A.1 Seleção de peptídeos para validação de metadinâmica e configuração do sistema

Quatro peptídeos promissores (2q6g, cadeia C; 1uk4, cadeia H; 1lvm, cadeia D; e 1lvb, cadeia D) foram selecionados a partir do resultado do Propedia para a estrutura da Mpro da SARS-CoV-2 para posteriormente serem calculados o ΔG_{bind} por simulação de metadinâmica (MetaD) (BARDUCCI; BONOMI; PARRINELLO, 2011; BUSSI; LAIO; PARRINELLO, 2006; BRANDT et al., 2016), visando a comparação com os valores das métricas retornados pelo Propedia. Os peptídeos foram selecionados procurando uma cobertura máxima da classificação com base tanto nos resultados das métricas de alinhamento e RMSD considerando os resíduos do sítio de ligação (Tabela 10).

As poses iniciais dos peptídeos foram consideradas como sendo aquelas com o melhor resultado obtido pelo Rosetta FlexPepDock. Preservando os estados de protonação previamente atribuídos, os mesmos foram transportados para um procedimento de montagem da topologia, solvatação em uma caixa d'água com preenchimento de 12 Å e adição de Na^+/Cl^- até a neutralização do sistema e força iônica de 0,15 M. Para esses procedimentos, os respectivos pacotes *psfgen*, *solvate* e *ionize*, provindos do *software* VMD/NAMD (HUMPHREY et al., 1996; PHILLIPS et al., 2005), foram utilizados. O campo de força CHARMM36 Huang e Jr (2013), Best et al. (2012) foi aplicado tanto para proteína (receptor), peptídeo, água e íons, quanto o modelo TIP3P para as moléculas de água (PRICE; III, 2004).

	Cadeia	Propedia		MetaD		R^2 ΔG_{bind}	
PDB id	peptídeo	Métrica alin.	RMSD sítio (Å)	ΔG_{bind} (kcal/mol)	σ	Métrica alin.	RMSD sítio
2q6g	C	10.36	0.34	-15.92	±1.34	(-) 0.98	(+) 0.96
1uk4	H	9.47	0.44	-15.97	±4.32		
1lvm	D	5.69	0.84	-4.76	±1.27		
1lvb	D	4.54	1.21	-0.61	±1.57		

Tabela 10 – Correlação entre a energia livre de ligação estimada pela MetaD ΔG_{bind} e seus desvios padrão (σ) e o valores obtidos pelo Propedia (métrica de alinhamento - Métrica alin.) e RMSD do resíduos do sítio de ligação (RMSD sítio). Nas últimas duas colunas, os respectivos coeficientes de correlação negativo e positivo de ΔG_{bind} com cada um dos resultados recuperados do Propedia (Métrica alin. e RMSD sítio).

A.2 Procedimentos de simulação

Todas as simulações foram realizadas no pacote NAMD 2.13 (PHILLIPS et al., 2005), no conjunto NPT *NPT ensemble*, usando o termostato de Langevin e barostato, respectivamente ajustados para 300 K e 1 atm. Foram utilizadas condições de fronteira periódicas (*periodic boundary conditions*, bem como métodos de malha de partículas de Ewald ((*particle mesh Ewald*)) para os cálculos das forças eletrostáticas de longo alcance, com um corte de 12 Å para as interações não-covalentes e um intervalo de tempo de 2 fs. A dinâmica dos átomos de hidrogênio foi restringida e estimada pelo algoritmo SETTLE, implementado para NAMD 2.13 (PHILLIPS et al., 2005).

Antes dos próprios procedimentos de MD, foi aplicado para cada sistema um protocolo metódico de minimização/relaxamento/equilíbrio. Inicialmente, cada sistema foi minimizado em 10.000 passos pelo algoritmo do gradiente conjugado *conjugate gradient algorithm*, também implementado para NAMD 2.13 (PHILLIPS et al., 2005). Em seguida, um protocolo de dinâmica molecular (DM) de relaxamento/equilíbrio de 10 passos, nas condições de simulação listadas anteriormente e com adaptação gradual de restrições harmônicas, conforme as etapas descritas a seguir:

- 500 ps de DM com restrições harmônicas para todos os átomos do receptor e do peptídeo.
- 500 ps de DM com restrições harmônicas apenas para os átomos da cadeia principal *backbone* do receptor e do peptídeo.
- 500 ps de DM com restrições harmônicas apenas para os átomos da cadeia principal do receptor.
- 500 ps de DM sem restrições harmônicas.
- 8 ns de DM sem restrições harmônicas e com reinicialização prévia das velocidades de acordo com um conjunto NPT de 300 K e 1 atm.
- 500 ps de DM reintroduzindo as restrições harmônicas nos átomos da cadeia principal do receptor e do peptídeo.
- 500 ps de DM reintroduzindo as restrições harmônicas em todos os átomos do receptor e do peptídeo.
- 300 ps de DM com remoção das restrições harmônicas das cadeias laterais do peptídeo.
- 300 ps de DM com remoção das restrições harmônicas de todos os átomos do peptídeo.
- 1 ns de DM nas condições acima mencionadas e reiniciando as velocidades para um conjunto NPT compatível com 300 K e 1 atm.

As 5 últimas etapas foram realizadas a fim de preparar o sistema para as condições de restrição utilizadas ao longo do procedimento de DM, ou seja, restrições harmônicas para todo o receptor e liberdade completa apenas para o peptídeo (ver abaixo)).

Após o relaxamento, o último quadro (*frame*)) de cada sistema foi levado a três procedimentos de MetaD independentes de 10 ns. Tais procedimentos foram feitos para desvincular o peptídeo do sítio de ligação, de acordo com o protocolo descrito em Brandt et al. (2016). Basicamente, todos os átomos da Mpro da SARS-CoV-2 foram mantidos harmonicamente restringidos, enquanto foi dado a liberdade completa ao peptídeo, moléculas de água e íons. Duas variáveis coletivas (*collective variables* - CV) foram usadas para descrever o processo de desvinculação do peptídeo da bolsa catalítica. A primeira, CV_{dist}, foi definida como a distância em Å entre os respectivos centros de massa do C145 catalítico da Mpro e do resíduo mais próximo (na pose inicial) do peptídeo. Os respectivos resíduos para cada um dos quatro peptídeos analisados foram: S7 para o 2q6g_C; Q5 para o 1uk4_H; Q307 tanto para 1lvmD_D e 1lvb_D. A segundo CV, CV_{ang}, foi definida como o ângulo em graus (º) determinado pelo centro de massa do resíduo C145 da Mpro, mencionado anteriormente para cada peptídeo e o centro de massa do péptido como um todo. A altura das gaussianas para a MetaD foi definida como 0,02 Kcal/mol e adicionada a cada 2ps com uma largura de 1,77. A CV_{dist} variou entre 0 e 30 Å com uma flutuação de amplitude de 2 Å, enquanto a CV_{ang} variou entre 0º e 180º com uma flutuação de amplitude de 10º. Os panoramas de potenciais de força média (*potential of mean force landscapes*) foram salvos a cada 1 ps. Os resultados foram analisados através de *scripts* desenvolvidos em R e Python, e também com os *softwares* VMD (HUMPHREY et al., 1996) e Wordom (SEEBER et al., 2007).

A.3 Mapas de energia livre e projeções das energias de metadinamicas ao longo das dimensões da CV_{dist} e estimativa da $\triangle G_{bind}$

Para selecionar o número de quadros a serem considerados ao longo da reconstrução do potencial de força média, foi utilizado tanto as energias de interação da mecânica molecular não-covalente entre o peptídeo e a proteína, quanto a distância associada a CV_{dist}. (ou seja, a distância entre o respectivo C145 e os centros de massa de resíduo mais próximos) como uma métrica para observar: 1) o acesso e o preenchimento da energia mínima A, ou seja, a energia mínima para o peptídeo dentro da proteína; 2) o acesso e o preenchimento do mínimo de energia B, ou seja, o mínimo de energia para o peptídeo fora da proteína, no meio aquoso; 3) o evento de re-cruzamento, ou seja, a fase de simulação na qual, uma vez que o sistema atingiu e preencheu completamente ambos os mínimos, o peptídeo ganha maior liberdade dinâmica e vira para visitar ambos os mínimos repetidamente. Seguindo o sugerido na literatura, selecionamos os mapas de potencial de força média salvos até a etapa de simulação imediatamente antes do evento de

re-cruzamento para reconstruir o panorama de energia livre ao longo do evento de desvinculação (BRANDT et al., 2016; RAITERI et al., 2006). Isso é feito para evitar o sobre-preenchimento dos mínimos de energia pelos potenciais gaussianos da MetaD e uma perda de precisão ao longo da reconstrução do panorama de energia livre.

A CV mais diretamente relacionado ao processo de desvinculação é, naturalmente, aquele que descreve a variação da distância entre o peptídeo e o sítio ativo (CV_{dist}). Desta forma, as projeções da energia livre da MetaD na CV_{dist} foi calculado de forma semelhante a Brandt et al. (2016), conforme a seguinte equação:

$$-\beta G_{CV_{ang}}(CV_{dist}) = ln\frac{\int e^{-\beta G(CV_{dist},\, CV_{ang})} dCV_{ang}}{\int e^{-\beta G(CV_{dist},\, CV_{ang})} dCV_{ang} dCV_{dist}} \tag{A.1}$$

onde β = 1/k_bT, sendo k_b a constante de Boltzmann (1,9858 x 10^{-3}·kcal·mol^{-1}·K^{-1}); T = 300 K; e G(CV_{dist}, CV_{ang}) para o valor de energia livre na posição (CV_{dist}, CV_{ang}).

Para a estimativa de ΔG_{bind} de cada réplica da MetaD, o valor mínimo de $G_{CV_{ang}}(CV_{dist})$ em uma CV_{dist} compatível com a completa independência do ambiente do peptídeo da influência da proteína (ou seja, $CV_{dist} \geq 25$ Å) foi diminuída do valor mínimo desta medida dentro de uma distância compatível com o peptídeo ligado (especificamente ou não) ao sítio ativo da proteína (ou seja, $CV_{dist} \leq 20$ Å) de acordo com a equação:

$$\Delta G_{bind} = G'_{CV_{ang}}(CV_{dist}) - G''_{CV_{ang}}(CV_{dist}) \tag{A.2}$$

onde $G'_{CV_{ang}}(CV_{dist})$ é o valor mínimo dentro da proteína, enquanto $G''_{CV_{ang}}(CV_{dist})$ é o mínimo fora da proteína. Para os casos em que dois ou mais mínimos com favorabilidade semelhante foram encontrados dentro da proteína, ambos os mínimos foram igualmente ponderados em um valor global de $G'_{CV_{ang}}(CV_{dist})$ (i.e.,$G_{CV_{ang}}(CV_{dist})^{Min}{}_{inside}$) de acordo com a equação:

$$-\beta G_{CV_{Ang.}}(CV_{Dist.})^{Min.}_{inside} = ln\frac{\sum_{i=1}^{n} e^{-\beta G'_{CV_{Ang.}}(CV_{Dist.})}}{\int e^{-\beta G_{Ang.}(\, CV_{dist})} dCV_{Dist.}} \tag{A.3}$$

onde i = 1, 2,..n, sendo o número de mínimos equivalentes em diferentes CV_{dist} valores dentro da proteína. Finalmente, a precisão do Propedia foi testada pela análise de MetaD medindo a correlação dos valores ΔG_{bind} estimados de acordo com as equações (A.1, A.2, A.3) e os respectivos valores das métricas de alinhamento e o RMSD do sítio de ligação recuperado pelo Propedia para cada um dos quatro peptídeos escolhidos para validação.

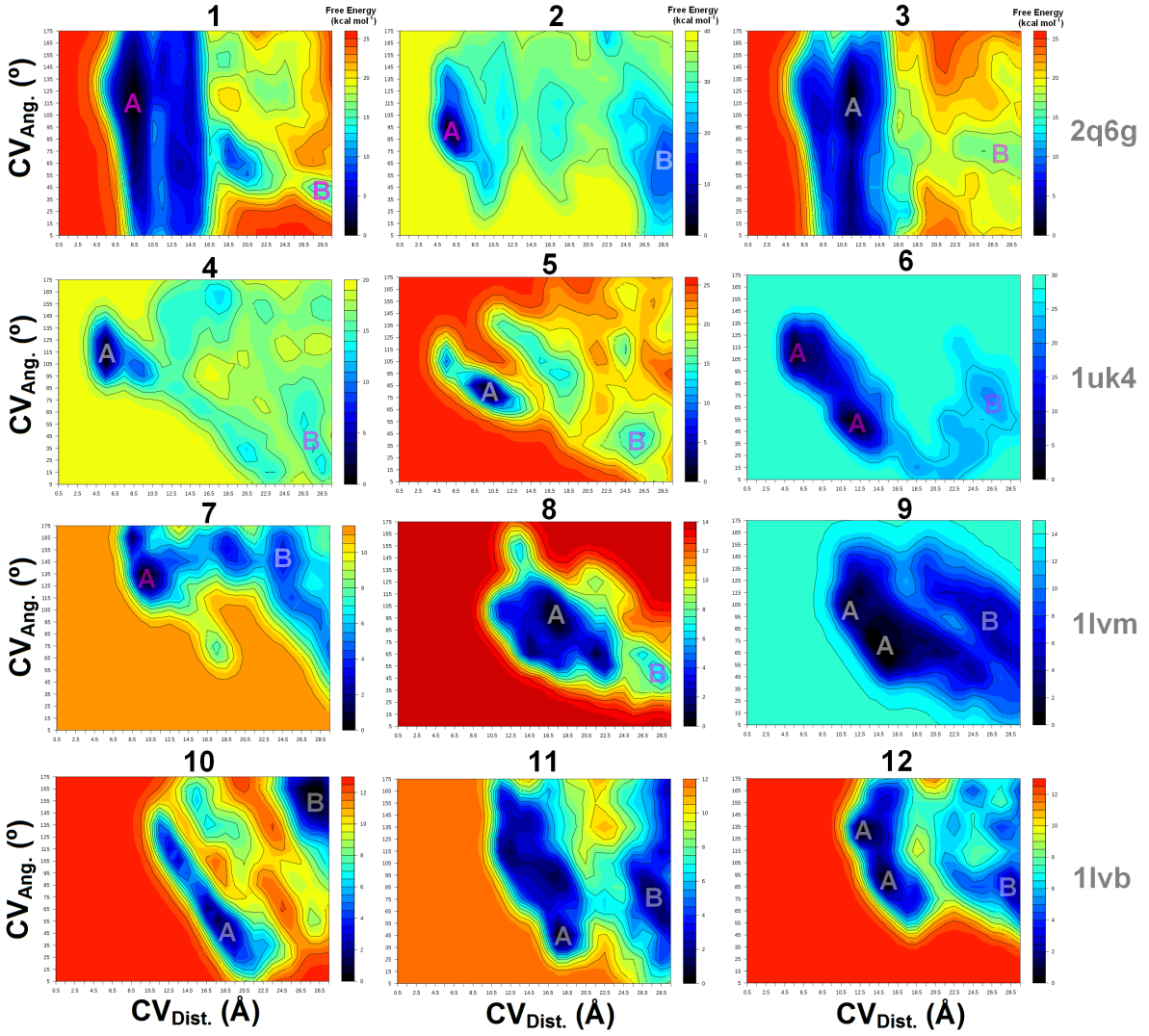


Figura 33 – Triplicatas dos panoramas de energia livre para desvincular os peptídeos do sítio de ligação da Mpro da SARS-CoV-2. PDB e cadeias do peptídeos: 2q6g, cadeia C; 1uk4, cadeia H; 1lvm, cadeia D; e 1lvb, cadeia D. Em cada sistema, o A representa os mínimos da proteína, enquanto B, os mínimos no meio aquoso.

APÊNDICE B – *Anticarsia gemmatalis* Hübner

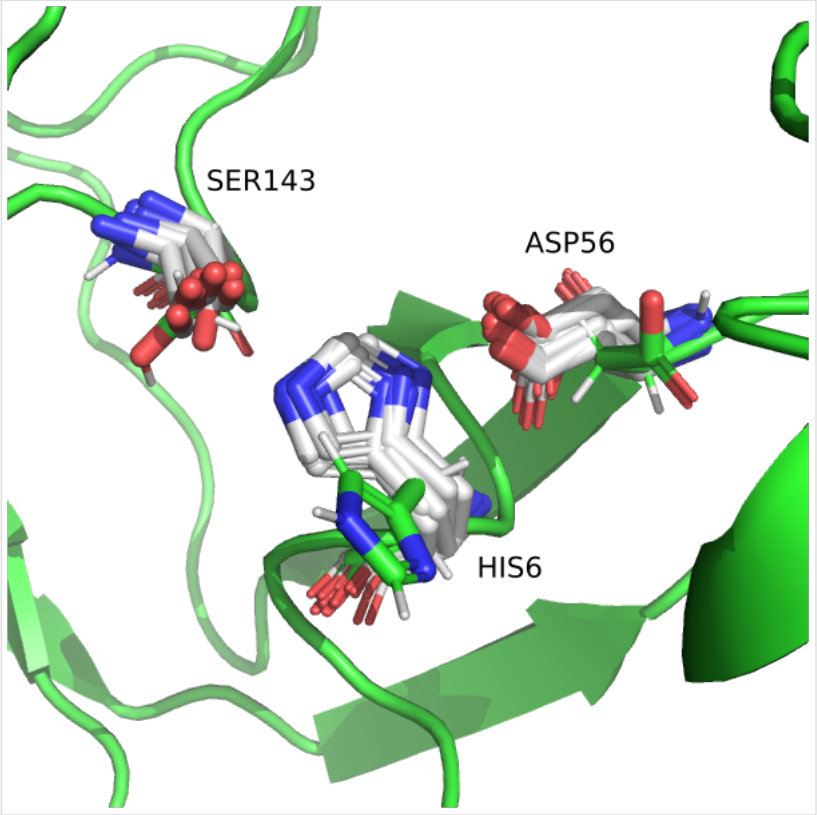
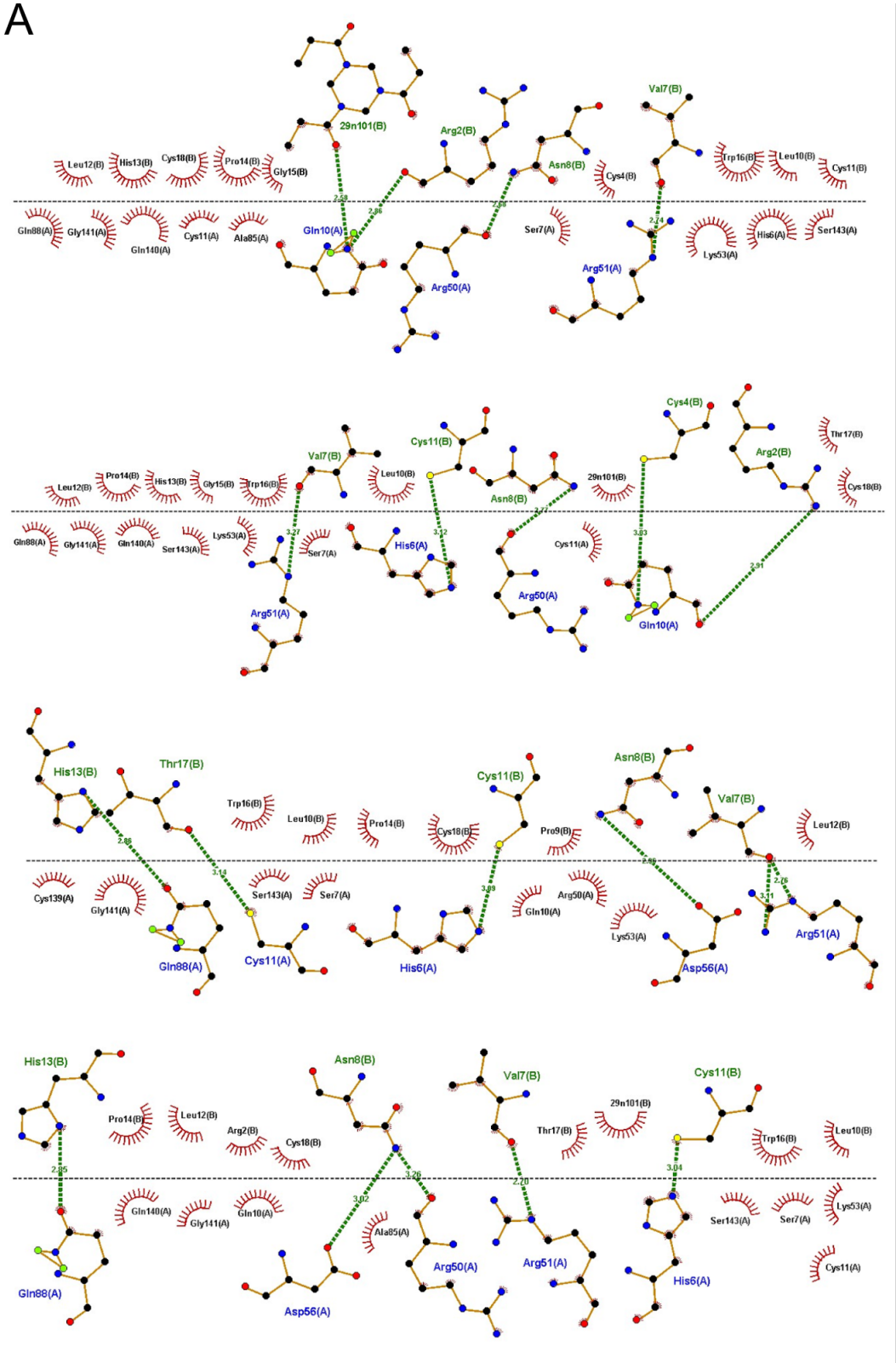
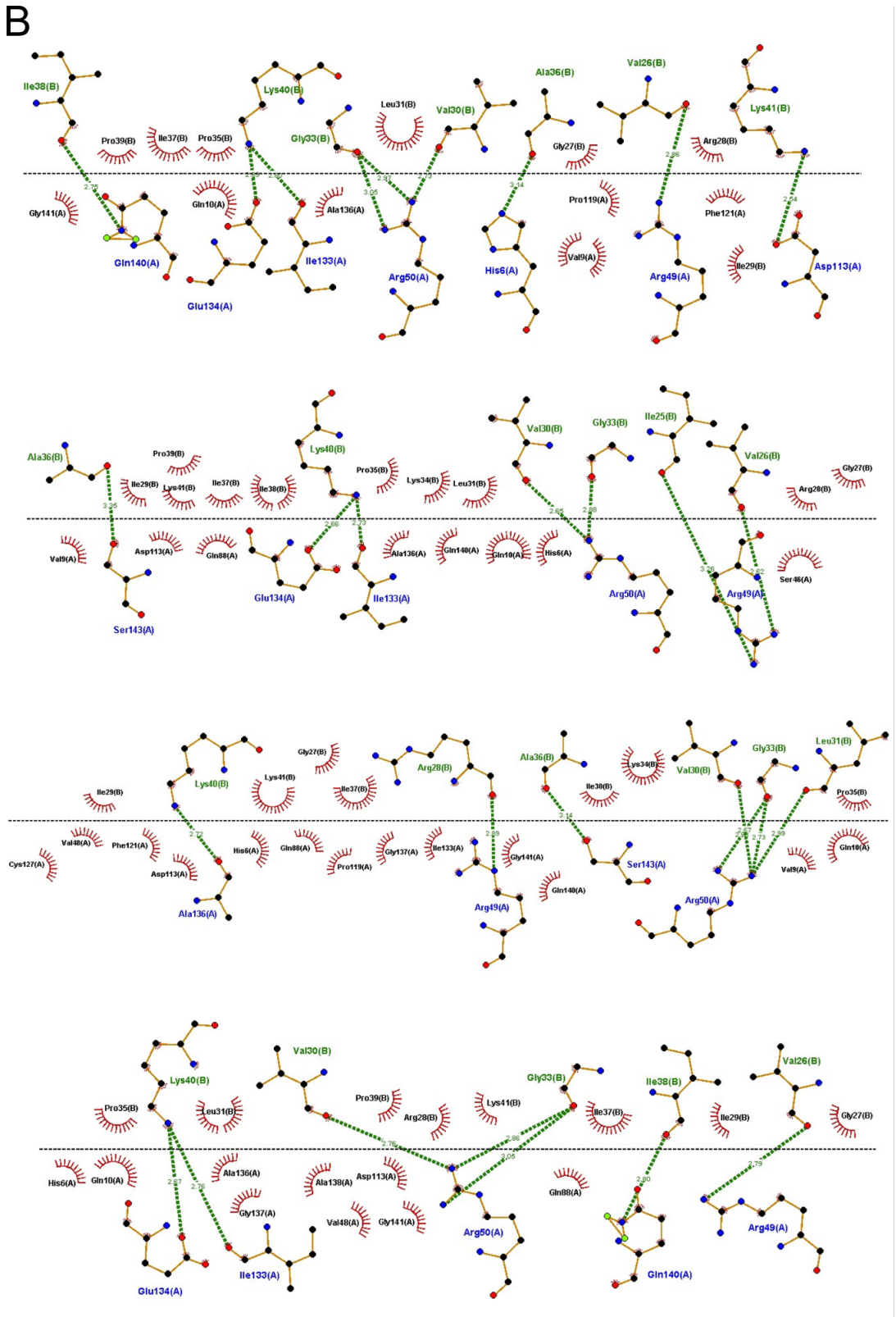


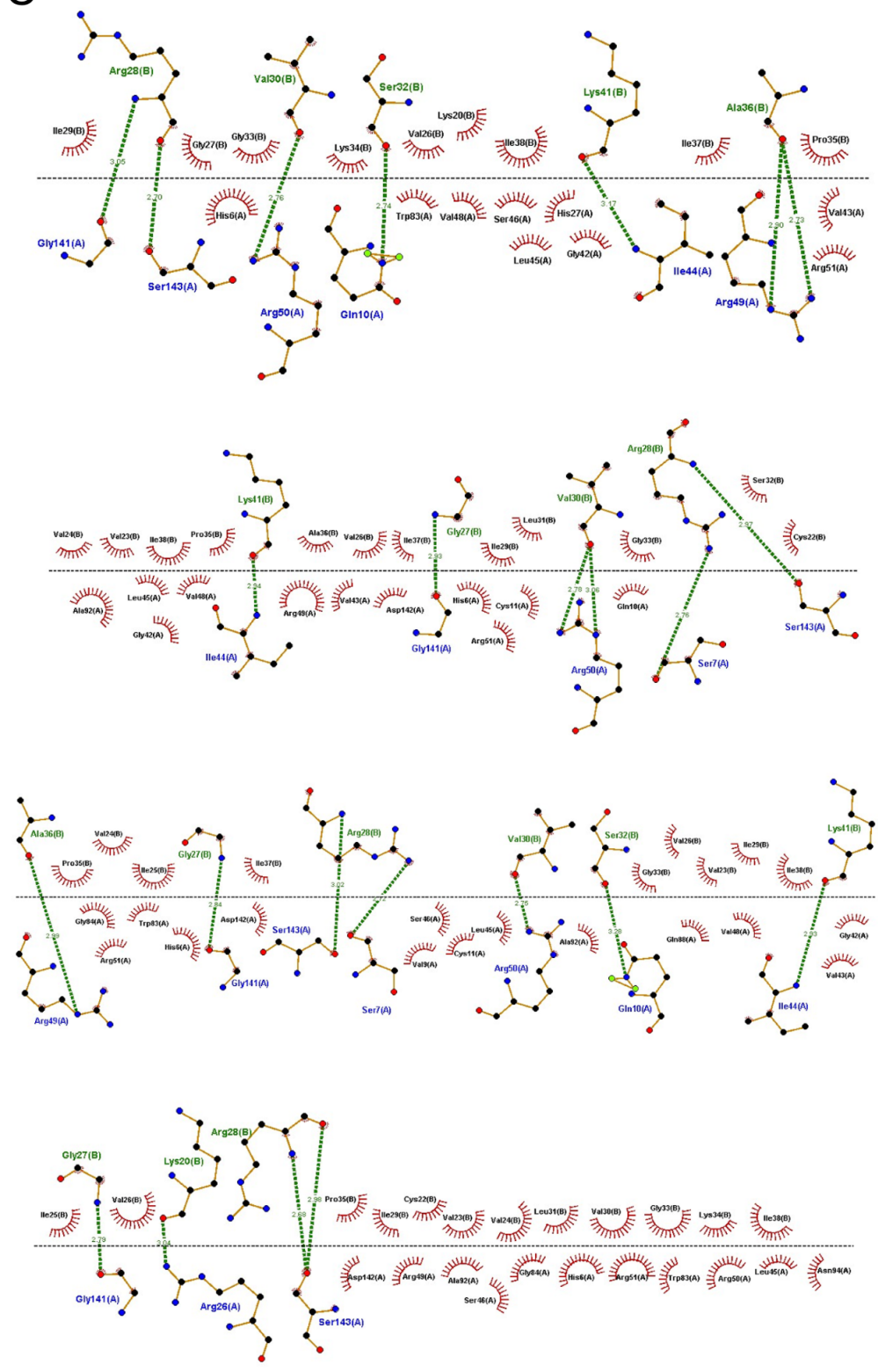
Figura 34 – Sobreposição do alinhamento estrutural do modelo de protease de AG (em verde) e os modelos melhores classificados usados no procedimento de modelagem. Os resíduos representados em bastões (*sticks*) são os resíduos do modelo correspondentes à tríade catalítica

Figura 35 – Mapas de interação para as três melhores resultados de FFC, dos complexos 6rw2_B_A (A), 3kn2_B_C (B) e 2obq_B_C (C). Os resíduos dos receptores estão marcados com seus nomes em azul enquanto os resíduos dos peptídeos estão em verde. As ligações de hidrogênio são destacadas em linhas verdes tracejadas e as interações hidrofóbicas são indicadas pelos arcos vermelhos. Figura gerada pelo *software* LIGPLOT (WALLACE; LASKOWSKI; THORNTON, 1995).




B
Ile38(B)
Pro39(B)
Ile37(B)
Pro35(B)
Lys40(B)
Gly33(B)
Leu31(B)
Val30(B)
Ala36(B)
Gly27(B)
Val26(B)
Arg28(B)
Lys41(B)
Gly141(A)
Gln10(A)
Ala136(A)
Pro119(A)
Phe121(A)
Gln140(A)
Ile133(A)
Arg50(A)
His6(A)
Val9(A)
Arg49(A)
Ile29(B)
Asp113(A)
Glu134(A)
Ala36(B)
Pro39(B)
Ile29(B)
Lys41(B)
Ile37(B)
Ile38(B)
Lys40(B)
Pro35(B)
Lys34(B)
Leu31(B)
Val30(B)
Gly33(B)
Ile25(B)
Val26(B)
Gly27(B)
Arg28(B)
Val9(A)
Asp113(A)
Gln88(A)
Ala136(A)
Gln140(A)
Gln10(A)
His6(A)
Ser46(A)
Ser143(A)
Glu134(A)
Ile133(A)
Arg50(A)
Arg49(A)
Gly27(B)
Lys40(B)
Lys41(B)
Arg28(B)
Ala36(B)
Lys34(B)
Val30(B)
Gly33(B)
Leu31(B)
Ile29(B)
Ile37(B)
Ile38(B)
Pro35(B)
Val48(A)
Phe121(A)
His6(A)
Gln88(A)
Gly137(A)
Ile133(A)
Gly141(A)
Gln10(A)
Cys127(A)
Asp113(A)
Pro119(A)
Ser143(A)
Val9(A)
Ala136(A)
Arg49(A)
Gln140(A)
Arg50(A)
Val30(B)
Gly33(B)
Ile38(B)
Val26(B)
Lys40(B)
Pro39(B)
Lys41(B)
Ile37(B)
Ile29(B)
Gly27(B)
Pro35(B)
Leu31(B)
Arg28(B)
His6(A)
Gln10(A)
Ala136(A)
Ala138(A)
Asp113(A)
Gln88(A)
Gly137(A)
Val48(A)
Gly141(A)
Arg50(A)
Arg49(A)
Glu134(A)
Ile133(A)
Gln140(A)

C


Arg28(B)
Val30(B)
Ser32(B)
Lys41(B)
Ala36(B)
Ile29(B)
Gly27(B)
Gly33(B)
Lys34(B)
Lys20(B)
Val26(B)
Ile38(B)
Ile37(B)
Pro35(B)
His6(A)
Trp83(A)
Val48(A)
Ser46(A)
His27(A)
Val43(A)
Leu45(A)
Gly42(A)
Arg51(A)
Gly141(A)
Ser143(A)
Arg50(A)
Gln10(A)
Ile44(A)
Arg49(A)
Lys41(B)
Val24(B)
Val23(B)
Leu31(B)
Cys22(B)
Ala92(A)
Val43(A)
Asp142(A)
Cys11(A)
Gln10(A)
Ser7(A)
Ala36(B)
Gly84(A)
Asn94(A)
Gln88(A)
Gly42(A)
Lys20(B)
Arg26(A)

APÊNDICE C – Artigos Publicados

C.1 Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm

Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm

Pedro M. Martins, Lucianna H. Santos, Diego Mariano, Felippe C. Queiroz, Luana L. Bastos, Isabela de S. Gomes, Pedro H. C. Fischer, Rafael E. O. Rocha, Sabrina A. Silveira, Leonardo H. F. de Lima, Mariana T. Q. de Magalhães, Maria G. A. Oliveira & Raquel C. de Melo-Minardi

BMC Bioinformatics **22**, Article number: 1 (2021) | Cite this article

11k Accesses | **59** Citations | **15** Altmetric | Metrics

Abstract

Background

Protein–peptide interactions play a fundamental role in a wide variety of biological processes, such as cell signaling, regulatory networks, immune responses, and enzyme inhibition. Peptides are characterized by low toxicity and small interface areas; therefore, they are good targets for therapeutic strategies, rational drug planning and protein inhibition. Approximately 10% of the ethical pharmaceutical market is protein/peptide-based. Furthermore, it is estimated that 40% of protein interactions are mediated by peptides. Despite the fast increase in the volume of biological data, particularly on sequences and structures, there remains a lack of broad and comprehensive protein–peptide databases and tools that allow the retrieval, characterization and understanding of protein–peptide recognition and consequently support peptide design.

C.2 Vermont: a multi-perspective visual interactive platform for mutational analysis

The Author(s) *BMC Bioinformatics* 2017, **18**(Suppl 10):403
DOI 10.1186/s12859-017-1789-3

BMC Bioinformatics

RESEARCH Open Access

Vermont: a multi-perspective visual interactive platform for mutational analysis

Alexandre V. Fassio[1,2*†], Pedro M. Martins[1,2†], Samuel da S. Guimarães[3], Sócrates S. A. Junior[3], Vagner S. Ribeiro[3], Raquel C. de Melo-Minardi[1] and Sabrina de A. Silveira[3]

From Symposium on Biological Data Visualization (BioVis) 2017
Prague, Czech Republic. 24 July 17

Abstract

Background: A huge amount of data about genomes and sequence variation is available and continues to grow on a large scale, which makes experimentally characterizing these mutations infeasible regarding disease association and effects on protein structure and function. Therefore, reliable computational approaches are needed to support the understanding of mutations and their impacts. Here, we present VERMONT 2.0, a visual interactive platform that combines sequence and structural parameters with interactive visualizations to make the impact of protein point mutations more understandable.

Results: We aimed to contribute a novel visual analytics oriented method to analyze and gain insight on the impact of protein point mutations. To assess the ability of VERMONT to do this, we visually examined a set of mutations that were experimentally characterized to determine if VERMONT could identify damaging mutations and why they can be considered so.

Conclusions: VERMONT allowed us to understand mutations by interpreting position-specific structural and physicochemical properties. Additionally, we note some specific positions we believe have an impact on protein function/structure in the case of mutation.

Keywords: Point mutation, Visual analytics platform, Intramolecular network, Complex network, Mutational analysis

C.3 Introducing programming skills for life science students

Article

Introducing Programming Skills for Life Science Students

Diego Mariano*
Pedro Martins
Lucianna Helene Santos
Raquel Cardoso de Melo-Minardi

From the Laboratory of Bioinformatics and Systems (LBS), Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract

The advent of the high-throughput next-generation sequencing produced a large number of biological data. Knowledge discovery from the huge amount of available biological data requires researchers to develop solid skills in biology and computer science. As the majority of the Bioinformatics professionals are either computer science or life sciences graduates, to teach biology skills to computer science students and computational skills to life science students has become usual. In this article, we reported the experience of teaching programming for life science students. Our strategy is composed by explaining basic concepts of algorithms, abstraction of biological problems, and script programming using Python language. Based on the student's answers to an assessment questionnaire, we conclude that the course achieved positive results. They reported an improvement in their skills in programming and bioinformatics. Furthermore, the students approved the didactic adopted in the classes and evaluation methods (programming exercises and final presentation). This article is useful for other professors who want to implement an initial bioinformatics training for undergraduate or graduate students in life sciences. We believe that the strategies here demonstrated could be reproduced, which could help in the formation of a new generation of bioinformaticians with hybrid abilities in computation and biology. © 2019 International Union of Biochemistry and Molecular Biology, 47(3):288–295, 2019.

Keywords: *Bioinformatics education; life science students; Python programming language*

C.4 Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes

Entertainment Computing 29 (2019) 31–42

Contents lists available at ScienceDirect

Entertainment Computing

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

Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes

Marcos F.M. Silva[a,*], Pedro M. Martins[a], Diego C.B. Mariano[a], Lucianna Helene Santos[a], Isabela Pastorini[a], Naiara Pantuza[a], Cristiane N. Nobre[b], Raquel C. de Melo-Minardi[a], Proteingo Players

[a] Department of Computer Science – Federal University of Minas Gerais (UFMG), Belo Horizonte, MG Brazil
[b] Post-Graduation Program in Informatics – Pontifical Catholic University of Minas Gerais (PUC-MG), Belo Horizonte, MG Brazil

ARTICLE INFO

Keywords:
Evaluation methodologies
Human-computer interface
Interactive learning environments
Interdisciplinary projects
Teaching/learning strategies
Serious games
Non-covalent interactions

2017 MSC:
00-01
99-00

ABSTRACT

Lately, numerous works have characterized the effectiveness of games in the process of learning. Benefits such as pleasure, stimulation, creativity, and enthusiasm have captivated people's interest in science through interactive games. Hence, many games for teaching subjects in biochemistry have been developed. However, understanding molecular interactions in proteins can still be a difficult task for students in biochemistry classes. One of the main issues reported by them is the fact they cannot visually see an interaction that happens between proteins. More specifically, the atomic structure, the particulate nature of matter, the molecule, and the chemical interactions are considered abstract concepts by teachers and students. In this paper, we present the conception, development, and evaluation of Proteingo: a game created to build a crowdsourced database of protein-protein interactions. An experiment was conducted during the X-meeting 2016 Congress in Brazil, where 27 users played the Proteingo game. The development of the players along the plays, together with the answers of a post-test questionnaire, give substantial evidence of user's incidental learning of chemical interactions through the Proteingo game. Also, 96.3% of the players consider the use of the game in classrooms as a didactic-pedagogical resource. Proteingo is available at http://bioinfo.dcc.ufmg.br/proteingo/.

C.5 How to compute protein residue contacts more accurately?

How to compute protein residue contacts more accurately?

Pedro M. Martins
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
pmartins@dcc.ufmg.br

Vinícius D. Mayrink
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
vdm@est.ufmg.br

Sabrina de A. Silveira
Universidade Federal de Viçosa
Viçosa, Minas Gerais, Brazil
sabrina@ufv.br

Carlos H. da Silveira
Universidade Federal de Itajubá
Itabira, Minas Gerais, Brazil
carlos.silveira@gmail.com

Leonardo H.F. de Lima
Universidade Federal de São João del-Rei
Sete Lagoas, Minas Gerais, Brazil
leofrancalima@ufsj.edu.br

Raquel C. de Melo- Minardi
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
raquelcm@dcc.ufmg.br

ABSTRACT

Computing contacts in proteins is important to several types of studies from Bioinformatics to Structural Biology. An accurate computation of contacts is essential to the correctness and reliability of applications involving folding prediction, protein structure prediction, quality assessment of protein structures, network contacts analysis, thermodynamic stability prediction, protein-protein and protein-ligand interactions, docking and so forth. In this work, we built an extensive database of contacts using about 45,000 structures from PDB to compare three paradigms for contact prospection at the atomic level: distance-based only, distance-geometry-based and distance-angulation-based.

The main contribution of this paper is a critical evaluation of three different paradigms that can be used to compute contacts between protein atoms. We focused on protein-protein interfaces and analyzed four types of contacts, namely hydrogen bonds, aromatic stackings, hydrophobic and ionic (attractive) interactions. We scanned for possible contacts in the range from 0 to 7 Å. Our database with all computed contacts as well as the source code used to populate this database is freely available at bioinfo.dcc.ufmg.br/capri Our data showed the importance of a geometric approach to filter out spurious occluded contacts after about 3.5 Å for aromatic stackings, hydrophobic and ionic interactions. For hydrogen bonds, to filter out spurious contacts, we need to consider the angles involved in the interactions.

CCS CONCEPTS

• **Applied computing** → **Molecular structural biology**;

KEYWORDS

protein-protein interaction; protein interface; protein residue contact; contact computation; database

SAC'18, April 9-13, 2018, Pau,France

ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/10.1145/3167132.3167136

ACM Reference format:
Pedro M. Martins, Vinícius D. Mayrink, Sabrina de A. Silveira, Carlos H. da Silveira, Leonardo H.F. de Lima, and Raquel C. de Melo- Minardi. 2018. How to compute protein residue contacts more accurately?. In *Proceedings of ACM SAC Conference, Pau,France, April 9-13, 2018 (SAC'18)*, 8 pages.
https://doi.org/10.1145/3167132.3167136

1 BACKGROUND

The computation of protein interactions is important for a variety of studies in Bioinformatics and Structural Biology. An accurate computation of contacts is essential to the correctness and reliability of applications involving folding prediction, protein structure prediction, quality assessment of protein structures, network contacts analysis, thermodynamic stability prediction, protein-protein and protein-ligand interactions, docking, among others. According to Silveira et al. [21], there are diverse forms to define a contact. A traditional and straightforward manner is to establish a threshold distance, also called cutoff. Considering two given points i and j, i is in contact with j if $d(i, j) < r$, where $d(i, j)$ is the Euclidean distance between i and j and r is the predefined threshold. According to the authors, the challenge is to define the correct cutoff since there is a broad range of options in the literature: 3.8 Å [17], 4.5 Å [11], 5.0 Å [7], 5.5 Å [9], 6.0 Å [20], 6.5 Å [19], 7.0 Å [1], 8.0 Å [16], 9.0 Å [15]. Another strategy to compute contacts is the *Voronoi tessellation* or *Delaunay triangulation*, its dual problem. It consists of filling a Euclidean space R^d with a collection of polytopes with no overlaps and no gaps between them. It is interesting because each point is connected only with the closest neighborhood points. In the case of proteins studied at the atomic level, atoms are the points and edges are interactions between such points and their first shell of contacting atoms.

In a general manner, authors classify methods for contact prospection as *cutoff dependent* (CD) and *cutoff free* (which we named DT, from Delaunay triangulation, which is an example of a cutoff free method). They also classify the existing methods according to their granularity: *fine-grained*, that works at an atomic level and *coarse-grained*, that works at the residue level. A coarse-grained analysis may use a simplification of the residue considering only their C_α (of the main chain), the geometric center or barycenter or even the last heavy atom of the side chain.

From our point of view, the work from Silveira and colleagues [21] represent a considerable advance to answer questions involved

APÊNDICE D – Prêmios

D.1 Best Poster Award X-Meeting 2016 na categoria “Software Development and Databases”

X-meeting 2016

Belo Horizonte | November 16th to 18th

12th International Conference of the AB3C

Best Poster Award

This certifies that the Best Poster Award is given to the work intitled "*An approach for constructing a database of manually curated contacts in proteins*", authored by **Marcos F M Silva, Pedro M Martins,Diego César Batista Mariano,Isabela Pastorini, Naiara Pantuza and Raquel Melo Minardi**, at the poster session (Software Development and Databases) of the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

Nicole Scherer
Nicole Scherer
Poster Session Chair

Mainá Bitar
Mainá Bitar
Poster Session Co-Chair

Glória Franco
Glória Franco
AB3C President

90 ANOS UFMG 1927 - 2017

AB3C