BIHARCK MUNIZ ARAÚJO

GAPIN: UMA FERRAMENTA PARA VISUALIZAÇÃO E ANÁLISE DE REDES DE INTERAÇÕES ATÔMICAS INTERMOLECULARES

Belo Horizonte
Julho de 2019

BIHARCK MUNIZ ARAÚJO

GAPIN: UMA FERRAMENTA PARA VISUALIZAÇÃO E ANÁLISE DE REDES DE INTERAÇÕES ATÔMICAS INTERMOLECULARES

Tese apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Dr. Carlos Henrique da Silveira

Belo Horizonte

Julho de 2019

Muniz Araújo, Biharck

GAPIN: Uma ferramenta para visualização e análise de redes de interações atômicas intermoleculares / Biharck Muniz Araújo. — Belo Horizonte, 2019
xxx, 103 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientadores: Dr. Carlos Henrique da Silveira

1. Protein-Protein-Interface. 2. Macromolecules. 3. PPI. I. Tï£¡tulo.

CDU

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE

Biharck Muniz Araújo

107/2019
entrada
2°/2014
CPF:
015.649.546-50

Às quatorze horas do dia **05 de julho de 2019**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Gapin: Uma Ferramenta para visualização e análise de redes de interações atomicas intermoleculares"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Carlos Henrique da Silveira**, 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	CPF	Indicação
Dr. Carlos Henrique da Silveira	UNIFEI	58794549672	APROVADO
Dr. Aristóteles Góes Neto	UFMG	54434882520	APROVADO
Dr. José Miguel Ortega	UFMG	05950126807	Aprovado
Dr. Wagner Meira Junior	UFMG	50996037691	Aprovado
Dr. Daniel Cristian Ferreira Soares	UNIFEI	041961156-32	Aprovado
Dr. Gerd Bruno da Rocha	UFPB	837504014-20	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 05 de julho de 2019.

Dr. Carlos Henrique da Silveira - Orientador _____
Dr. Aristóteles Góes Neto _____
Dr. José Miguel Ortega _____
Dr. Wagner Meira Junior _____
Dr. Daniel Cristian Ferreira Soares _____
Dr. Gerd Bruno da Rocha _____

Dedico essa tese principalmente a minhas amadas Aline Lopes Coelho e Liz Lopes Muniz por estarem sempre ao meu lado e me fazerem ser essa marido e pai super orgulhose de ter vencido na vida ao lado de vocês. Obrigado por fazerem parte de minha vida. Amo vocês.

Agradecimentos

Aproximadamente, trinta e três anos depois estou eu aqui escrevendo uma seção de agradecimentos de uma tese de doutorado... Nunca poderia imaginar tamanha responsabilidade! Não sei se aqui é o melhor local para contar um pouquinho de minha vida... Todavia, pensando na hipótese de que tudo que tenho e conquistei na vida foi porque fui presenteado pela companhia de pessoas que me ajudaram e acreditaram em minhas loucuras; o agradecimento parece ser um lugar apropriado.

Difícil começar sem antes agradecer aos meus pais e, principalmente, à minha mãe, Maria das Graças Muniz Araújo por tudo. E quando falo tudo, é tudo mesmo! Por abrir mão de sua vida para cuidar dos filhos. Saímos de diversas situações complexas de vida, moramos com minha avó Maria Júlia Gonçalves que infelizmente nos deixou em 2018. Uma fase um tanto complicada; mas, mesmo assim, minha mãe amada sempre cuidou para que nunca descobríssemos as dificuldades que estávamos passando. Só hoje, depois de alguns anos, é que consigo entender tudo. Por mais que eu agradeça o resto da minha vida, esse tempo não será o suficiente. Afinal de contas, aquele café da tarde com café puro e mandioca cozida não era atoa. Durante várias anos, minha mãe me levava e buscava todos os dias na escola. Fazia sol, fazia chuva, no interior ou na capital, onde quer que seja. Por isso, mãe, sou o homem que sou; graças a tudo que senhora fez por mim. Nesse momento, deixo aqui o meu mais sincero obrigado.

Agradeço infinitamente, também, à minha esposa amada e querida Aline Lopes Coelho. São quatorze anos de companheirismo, amor, dedicação e delírio por acreditar nas minhas insanidades. A primeira a aceitar estar ao meu lado com esse amor incondicional. Não poderia estar aqui hoje escrevendo esse agradecimento se não fosse um carro que ela comprou na minha mão para que eu pudesse pagar a faculdade! Acredite, se não fosse aquele momento, nunca teria me formado. Enfrentamos muita coisa juntos... Nesses 14 anos, posso dizer seguramente que vivemos de fato uns 90 anos de um casal comum. Foram tantos desafios, morando em cidades diferentes para realizar nossos sonhos, empreender, mudar de profissão... Aline sempre apoiou meus estudos, os quais aconteciam durante madrugadas, finais de semana, feriados. Obrigado por

emprestar seu Pentium III 700 mhz para que eu pudesse fazer os trabalhos de AEDS. Muitas vezes, apoiou meu choro de desespero quando não aguentava mais, querendo largar tudo pois precisava descansar para trabalhar no dia seguinte como uma pessoa normal. Aliás, pessoa normal não cabe no nosso dicionário. Enquanto nossos amigos saíam, iam para festas, nós estávamos em casa planejando nossa vida juntos, finanças, sonhos e metas. Traçando planos para esquivar das surpresas, algumas vezes, desagradáveis da vida. Planejando como conciliar os 3 empregos, a universidade, o mestrado e o doutorado. Nossa! Aline, como é incrível como você cuida bem de mim e me dá energia pra seguir em frente. Muito obrigado por tudo.

Como consequência desse amor, nasceu nossa "metadinha": Liz Lopes Muniz em junho de 2018. Antes mesmo dela entender o que é o mundo, já estava em meu colo enquanto eu programava o GAPIN. Às vezes, oferecendo seu bico (o que acredito que, de alguma forma, era para me dar conforto). Olhar para aquele rostinho era mais uma fonte de energia para seguir em frente.

Claro que não poderia deixar de agradecer ao meu orientador, ou melhor, ao meu grande amigo Carlos Henrique da Silveira. Poderia inclusive dizer em caixa alta - "GRANDE AMIGO- se não deixasse de ser polido . Nunca poderia imaginar que essa relação aluno e professor se tornaria uma parceria inacreditável! Ainda, de alguma forma, sinto que os nossos "times"de vida não foram muito bem sincronizados... Se não fosse isso, meu amigo Carlos, teríamos a empresa de tecnologia mais incrível do mundo! Ou seríamos os maiores lunáticos da face da terra (se é que já não somos!). Acompanhei de perto algumas de suas dores e pode ter certeza que sofri junto. Sei que o senhor também acompanhou algumas minhas e também sofreu junto. Se isso não é amizade eu não tenho outro adjetivo para descrever esse sentimento. Muito obrigado por estar comigo desde o começo e acreditar que eu poderia fazer algo incrível. Todas as vezes que me sentia mal, era só conversar com o senhor que saía da conversa acreditando que eu realmente era competente. Aliás, sabia que até hoje tenho aquela folha de caderno com marca de caneca de café de um de nossos encontros? Dentre viagens a trabalho nos falamos em diversos fusos-horários distintos, durante finais de semana, feriados, madrugadas, dias, tardes, via Skype, hangouts, WhatsApp, sinais de fumaça, diário, Bitbucket. Qualquer mecanismo de comunicação era o suficiente para nos encontrarmos.

Espero um dia ser para os meu alunos, 10 por cento do orientador que o senhor foi pra mim, assim, sei que já serei um professor realizado. Nunca vi alguém se entregar tanto assim em um trabalho. Obrigado Carlos e obrigado a sua família por entender e o apoiar nossos delírios.

Agradeço, também, ao meu querido amigo de doutorado Wandré Veloso que

esteve comigo desde o dia -1. Nos encontramos na porta do processo seletivo e mal sabíamos que nos tornaríamos amigos de ir na festinha das crianças!

De fato, a vida poderia ser um pouquinho mais fácil... Mas, se fosse tão fácil eu não teria chegado onde estou hoje e não estaria cercado das pessoas que amo.

"It's easy to play any musical instrument: all you have to do is touch the right key at the right time and the instrument will play itself."
(Johann Sebastian Bach)

Resumo

Vários métodos computacionais, bases de dados e ferramentas inovadoras têm sido propostas para visualização e análise de dados de interações biomoleculares, com o objetivo de extrair conhecimento útil principalmente através de representações das interfaces em redes de contatos. No entanto, tais iniciativas tendem a ser especializadas em determinado tipo de interface (proteína-proteína ou proteína-ligante, por exemplo), e não se generalizam facilmente para qualquer interfaceamento, independente das biomoléculas envolvidas. E acima de tudo: como melhor visualizar tais redes de interações, tendo o desenho de candidatos a fármacos em mente? No intuito de somar esforços a esses desafios, propõem-se aqui a ferramenta **GAPIN** - ***G****rouped and* ***A****ligned* ***P****rotein* ***I****nterface* ***N****etworks*, uma aplicação 100% web, com toda a imediata disponibilidade, portabilidade, usabilidade e conveniência que só os modernos navegadores podem oferecer. GAPIN tem como principal entrada de dados os arquivos PDBs, definindo interfaces entre biomoléculas como grafos em nível atômico. Uma granularidade nesse nível permite visualizações independente das biomoléculas envolvidas, se entre proteínas, ácidos nucléicos, carboidratos, lipídios, ligantes, íons ou mesmo águas. GAPIN é capaz de contrastar as estruturas PDB renderizadas com os respectivos grafos, em dois níveis de granularidade: a primeira, com nós representando átomos da interface; a segunda, com nós representando comunidades de átomos (fruto do agrupamento em grafo) conforme a densidade das arestas, formando grafos modularizados ou de alto nível. GAPIN também disponibiliza uma opção para alinhamento e similaridade dos grafos modularizados. Mostramos neste trabalho que grafos de alto nível podem ajudar na identificação e caracterização de *Spots*, regiões nas interfaces proteínas-proteínas que agrupam resíduos com relevante contribuição ao ΔG de *binding*. Há fortes evidências na literatura que tais regiões tendem a ser potenciais alvos para candidatos a fármacos. Visando dar destaque ao amplo espectro de ação do GAPIN, dois experimentos foram conduzidos: um envolvendo o alinhamento de regiões hidrofóbicas numa base de dados de diferentes serino-peptidases e inibidores; outro, envolvendo a caracterização de *Spots* numa base de dados de mutações por alanina. Mostramos que GAPIN foi

capaz de identificar, por alinhamento, padrões hidrofóbicos não triviais na primeira base de dados. Também foi capaz de revelar, na segunda base de dados, uma curiosa correlação entre as áreas de contatos dos nós em grafos modularizados e a presença de *Spots* energéticos. Espera-se que neste trabalho tenha sido possível demonstrar a versatilidade visual e a potencialidade analítica da ferramenta GAPIN, para estudos de uma grande variedade de interfaces intermoleculares, com efetivo poder de auxiliar pesquisadores de diversas especialidades a entenderem melhor as propriedades topológicas e físico-químicas de potenciais alvos terapêuticos na busca por fármacos inovadores.

Abstract

Aiming to extract useful knowledge mainly through representations of the interfaces in networks of contacts, several computational methods, databases, and innovative tools have been proposed for visualization and analysis of data of biomolecular interactions. However, such initiatives tend to be specialized in a particular type of interface, (such as protein-protein or protein-ligand, for instance), and do not easily generalize to any kind of interfaces, regardless of the biomolecules involved. And it is more than that: What is a good way to visualize such interactions networks, focusing on the design of drug candidates? In order to add efforts to these challenges, we propose a tool called **GAPIN** - ***G****rouped and* ***A****ligned* ***P****rotein* ***I****nterface* ***N****etworks*, which is a 100 % web application, with high availability, portability, usability and convenience that modern browsers can offer. The main data input for GAPIN is PDB files. GAPIN will define interfaces between biomolecules as graphs at the atomic level. A granularity at this level allows independent visualizations of the involved biomolecules, whether among proteins, nucleic acids, carbohydrates, lipids, ligands, ions or even waters. GAPIN is able to contrast the PDB structures rendered with its respective graphs, in two levels of granularity: the first one, with nodes representing atoms of the interface; the second, with nodes representing communities of atoms (the result of graph clustering) according to the density of the edges, forming modular or higher-level graphs. GAPIN also provides an option for alignment and similarity of modularized graphs. We show in this work that high-level graphs can help in the identification and characterization of Spots, regions at the protein-protein interfaces that group together residues with a relevant contribution to the ΔG of binding. Literature shows strong evidence that such regions tend to be potential targets for drug candidates. Aiming to highlight overal spectrum of GAPIN, two experiments were developed: the first one involving the alignment of hydrophobic regions in a database of different serine-peptidases and inhibitors; the other one, involving the characterization of Spots in an alanine mutagenesis data set. We have shown that GAPIN was able to identify, by alignment, non-trivial hydrophobic patterns in the first database. It was also able to reveal, in the second da-

tabase, a curious correlation between the contact areas of nodes in modularized graphs and the presence of energetic Spots. The outcome of this work expects the possibility to demonstrate the visual versatility and analytical potentiality of the GAPIN tool for the study of a huge variety of intermolecular interfaces, with effective power to help researchers from different specialties to get a better understanding about the topological and physicochemical properties of potential therapeutic targets for innovative drugs.

Lista de Figuras

1.1 Superfícies de Connolly (Connolly [1983]) das integrinas αIIb (azul) e β3 (branco), evidenciando o fármaco ligante Tirofibana em *stick* (laranja). A Tirofibana mimetiza o tripeptídeo ARG-GLY-ASP presente numa alça do fibrinogênio, ponto principal de interação deste com o complexo das integrinas αIIbβ3. PDBid: 2VDM. Imagem gerada no Pymol versão 1.7.2.1 (DeLano [2002]). 3

1.2 Tela de apresentação do GAPIN, com um menu destacado na tarja superior, entrada do PDBid no canto superior direito. Na parte central da página, divulgação de algumas funcionalidades, bem como um vídeo promocional. . 7

1.3 Exemplo de tela após entrada de um PDBid, como 1PPF. No lado esquerdo vemos os grafos em duas granularidades, fina ou baixo nível (abaixo); e grossa ou alto nível (acima), sendo este último resultado do agrupamento do primeiro. No lado direito, a estrutura renderizada da 1PPF, com *surface* de Connolly em uma das cadeias, *spacefill* dos átomos da interface, coloridos conforme o agrupamento do grafo. 8

3.1 Instruções para instalação do *RINalyzer* conforme disposto no site https://rinalyzer.de/docu/install.php 15

3.2 UCSF - Chimera (janela esquerda) e *RINalyzer Cytoscape* (janela direita) para 1PPF, destacando interface entre uma Elastase de Leucócito Humano e um Inibidor de Ovomudoide de Peru. A interface na estrutura e rede de contatos estão sincronizadas visualmente, em azul. 17

3.3 Tela do NAPS após escolha de *Protein Complex* para PDBid 1PPF - Elastase de Leucócito Humano e um Inibidor de Ovomudoide de Peru. Enzima em tons azuis, inibidor em tons verde, interface entre eles com arestas em laranja, arestas intracadeia em cinza. Centroides em carbonos alfas. O nó em vermelho representando o resíduo (I:LEU:18) no lado inibidor foi posto em destaque tanto na rede quanto estrutura. Esse resíduo ocupa o sítio de especificidade na elastase. 18

3.4 Montagem de telas do PDBePISA para 1PPF. A) Resultado do *Interface* após entrada da 1PPF. B) Resultado após escolha de *Details* em alguma linha de interação em A). C) Visualização de estruturas com destaque da interface cadeia-cadeia. D) Visualização de estruturas com destaque de uma interface cadeia-ligante, com átomos da cadeia em *spacefill* verde e do ligante (NAG - *N-Acetyl-D-Glucosamine*) em *wireframe* vermelho. 19

3.5 Tela capturada do *Protein Contacts Atlas*. Como não se conseguiu construir uma rede de contatos para interfaces cadeia-cadeia, segue um exemplo dos contatos cadeia-ligante para 1TEC - Subtilisina Termitase da bactéria *T. vulgaris* com Inibidor Eglina C de Sanguessuga. À esquerda, um *asteroid plots* evidencia contatos de 1a e 2a vizinhança (ordem) com o cálcio 343 da cadeia E (enzima). À direita, os contatos refletidos na estrutura, em *spacefill*, com o cálcio ao centro. 20

3.6 Tela capturada (e adaptada) do STRING. A) Resultado da consulta para o UniProtKB id P08246, que refere-se Elastase de Neutrófilo Humano (ELANE). B) Resultado de um duplo *clique* no nó ELANE, trazendo algumas informações anotadas. C) Ao clicar na imagem da estrutura, abre-se uma tela do PDBSum ((Laskowski et al. [1997]). Na aba *Prot-Prot* vem uma imagem estática contendo as interações da interface da ELANE (cadeia E) com inibidor antileucoproteinase SLPI - *Secretory Leukocyte Protease Inhibitor* (cadeia I) 22

3.7 Tela capturada (e adaptada para melhor visualização) do *webservice Interactoma3D*, indicando a interação entre uma elastase de neutrófilo humano e um inibidor antileucoproteinase SLPI - *Secretory Leukocyte Protease Inhibitor* - SLPI, destacados em quadrados vermelhos na rede do interactoma. Vê-se que tanto a elastase quanto o SLPI fazem interações com uma série de outras proteínas, nem todas com estruturas resolvidas ainda. PDB id: 2Z7F. 23

3.8 Tela capturada (e adaptada) do STRING. A) Resultado da consulta para o UniProtKB id P08246, que refere-se Elastase de Neutrófilo Humano (ELANE). B) Retorna basicamente o mesmo resultado da figura 3.6B. C) Ao clicar na imagem da estrutura, abre-se uma tela do PDBSum. Na aba *Ligands* vem uma imagem estática contendo as interações da interface da ELANE (cadeia E) com dois N-Acetyl-D-Glucosamina (NAG) e um Alfa-L-Fucose (FUC), ligados covalentemente entre si, e também com a enzima em E-ASN-159 24

3.9 Tela capturada do LigPlot+ para interações intercadeia usando DIMPLOT para 1PPF. Foi usado $*E$ e $*I$ para *DOMAIN1* e *DOMAIN2* respectivamente, e resto default. Ativação de *Hydrophobic interactions* e na distâncias em *Runtime parameters* $->$ *Non-bonded contact parameters* não surtiram efeito na imagem. 25

3.10 Tabela comparativa das ferramentas analisadas. Por essa tabela é possível analisar a complexidade de instalação das ferramentas, sua utilização e principais características 27

4.1 Área de contato tal qual definida pela metodologia BARS. A) Dada uma sonda de raio p, e raios de van der Waals r_i e r_j, a área de contato $Ac_{i,j}$ entre eles é aquela que a sonda não toca. B) Se a distância for maior que os raios de van der Waals mais o diâmetro da sonda ($d_{i,j} >= r_i + r_j + 2p$), $Ac_{i,j}$ será zero, configurando a possibilidade de uma cavidade entre átomos i e j. (figura adaptada de (Alves [2015]), com permissão.) 30

4.2 Interfaces intermoleculares em 1PPF, a partir de contatos em que $Ac \geq Ac_{min}$. A) Elastase e seu Inibidor Ovomucoide destacados em *surface* tipo Connolly, com os elementos da interface cadeia-cadeia em *ball+stick* coloridos conforme a cadeia. B) Mesmo anterior, mas átomos em *spacefill*. C) Foco na interface cadeia-cadeia, sem *surfaces*. D) Foco agora nas interfaces cadeias-ligantes. Os ligantes são oligossacarídeos (em verde) e águas (em azul). Como são muitas, uma boa extensão da superfície do complexo enzima-inibidor produziu $Ac \geq Ac_{min}$. As demais cores são das cadeias enzima e inibidor. E) Destaque em *surface* para os oligossacarídeos. F) Destaque em *surface* para as águas. 32

4.3 A) Representação de uma matriz de adjacências k-partida $\mathbf{M}$ para k cadeias. U_i representa o conjunto de átomos de cada cadeia i. Não há arestas ($m_{i,j} = 0$) entre átomos intracadeias (entre U_i e U_i). B) Exemplo de um rede envolvendo 3 cadeias U_1, U_2 e U_3, coloridas por cores diferentes, com átomos simbolizados por círculos (nós) e arestas indicando um Ac diferente de zero. Nota-se a presença de contatos ternários ao centro, envolvendo átomos de 3 cadeias diferentes, interligados entre si. 35
4.4 Exemplo de grafo de alto nível construído a partir de um grafo de baixo nível. A) Mesmo grafo de baixo nível da figura (4.3B), pondo em evidência 3 grupos pelas 3 cores diferentes. Os traços pretos indicam onde houve o corte (*cut*) das arestas de modo a produzir 3 partições no grafo. Os números próximos às arestas indicam as respectivas áreas de contato (Ac). B) O grafo de alto nível derivado do grafo de baixo nível. Os rótulos nos nós indicam o somatório das áreas de contatos (pesos) da arestas internas aos grupos. Os rótulos nas arestas indicam o volume do corte, ou o somatório das arestas cortadas para produzir as partições. Graficamente, tanto nós quanto arestas são proporcionais aos respectivos valores dos rótulos. 36
4.5 A) Figura de um grafo de alto nível de interfaces cadeia-cadeia apolares com 6 grupos da 1PPF. O rótulo do pequeno nó em bege mais à direita é de 29Å^2. B) Respectiva tabela com a matriz de adjacências. Valores indicam áreas de contato em Å^2, exceção à coluna $\boldsymbol{Q}\%$ que representa a qualidade do nó, em porcentagem. 39
4.6 Varredura de agrupamentos para interfaces cadeia-cadeia com interações apolares em 1PPF, variando k de 3 a 6. Cores em marrom-vermelho indicam nós mais preservados que em amarelo-branco. 40
4.7 Resultado do alinhamento de grafos de alto nível para interfaces cadeia-cadeia apolares entre 1PPF (complexo entre Elastase de Leucócito Humano e Inibidor Ovomucoide de Peru) com 1TEC (complexo entre Subtilisina Termitase da bactéria *T. vulgaris* com Inibidor Eglina C de sanguessunga). À esquerda, grafos de baixo nível; ao centro, grafos de alto nível; à direita, estruturas renderizadas. Nós alinhados têm as mesmas cores, bem como respectivos átomos nas estruturas associados aos nós. 44
4.8 Visão geral do projeto seguindo as técnicas de *Continuous Deployment* com *Bitbucket Pipelines* em que uma vez que alguma alteração é enviada ao Bitbucket, um pipeline é iniciado rodando uma bateria de testes a fim de garantir que as alterações não afetem o código em ambiente produtivo e dê um feedback rápido antes do deployment 46

4.9 Arquitetura *Back end* do GAPIN, com um exemplo de conteúdo (alguns dados da 1TEC) no formato JSON armazenado no MongoDB 47

5.1 Página inicial do GAPIN em que o usuário busca por um PDBid que não existe ainda na base de dados 50
5.2 Página de importação de PDBid do GAPIN. Do lado direito a opção de importar o PDB pela base do PDB e do lado direito a opção de realizar (*upload*) de um arquivo PDBid customizado. 50
5.3 Após o processo de importação ser iniciado, o usuário poderá acompanhar o andamento da importação pela tela de (*jobs*) 51
5.4 Início do processo de importação de qualquer PDB 51
5.5 Primeiros processos já finalizados para a etapa de importação. 51
5.6 Tela principal do GAPIN exibindo a 1PPF. A) Grafo de alto nível o qual mostra os agrupamentos. Nesse caso pode-se ver a divisão em 5 grupos sendo 4 contectados e um desconexo. B) Grafo de baixo nível ou nível atômico o qual cada nó do grafo é um átomo. C) estrutura em 3D interativa da proteína em questão. 52
5.7 *Mouse hover* sobre átomos no grafo da rede de interações. 53
5.8 Interação do usuário com a rede de nível atômica do lado inferior esquerdo refletindo a seleção na estrutura do lado direito 54
5.9 Interação do usuário com (*mouse hover*) do grafo em nível de grupos . . . 54
5.10 Interação do usuário com a seleção de grupos refletindo na estrutura do lado direito e na rede a nível atômico do lado inferior esquerdo 55
5.11 Representações em dois formatos distintos da 1PPF. A) Representação em *ball+stick* B) Representação em *spacefill* 55
5.12 Representação das cores no GAPIN. A) Divisão de cores por grupos. B) Divisão de cores por cadeias. C) Divisão de cores por elementos. D) Divisão de cores por polaridade. E) Sem divisão de cores 56
5.13 Representação das combinações de visualização da biomolécula. A) Somente ANY sem os ligantes também com a estrutura completa dos resíduos B) Somente os ligantes com a estrutura completa dos resíduos C) Somente os átomos que fazem parte do contato D) Somente os átomos que fazem parte do contato sem a visualização da estrutura completa E) Somente os átomos que fazem parte do contato sem os resíduos 57
5.14 A) Representação das combinações de *surface* de uma biomolécula pelo GAPIN. B) Cadeias e ligantes B) Cadeias e águas C) Somente ligantes . . 58
5.15 Representação das interfaces da 1HMD 59

5.16 Representação proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface com o Inibidor em formato $Ball+Stick$ coloridos por uma distribuição de 5 grupos. 60

5.17 Representação proteína 1PPF dando destaque a enzima e inibidor em formato *surface* e os átomos que fazem parte da interface $Ball+Stick$ coloridos por uma distribuição de 5 grupos. 60

5.18 Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a Enzima em formato (*surface*) e os átomos que fazem parte da interface em formato (*Ball Stick*) contendo a estrutura completa do resíduo em que o átomo pertence 61

5.19 Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato *Spacefill* contendo a estrutura completa do resíduo em que o átomo pertence. 62

5.20 Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato $BallStick$) 63

5.21 Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato *spacefill*. 63

5.22 Representação das cores da interface selecionada no GAPIN 64

5.23 Modelo de seleção manual de átomos. A). Caso o usuário queira visualizar todo o resíduo, basta digitar I.LEU18 e o a Leucina 18 do inibidor será colocada em evidência conforme figura B) Pode-se também selecionar apenas o nome de uma das cadeias e todos os elementos que pertencerem a essa cadeia serão selecionados C) O fato de já existir uma seleção pré-existente não impede a adição de novos elementos conform D) em que já havia uma seleção em torno do elemento I.GLU19.CB e suas conexões de primeiro nível e fora adicionado o resíduo I.LEU18. 66

5.24 Opção de seleção de cores pelo GAPIN para o grafo de contatos e a estrutura. 67

5.25 Fluxo de adicionar uma nova biomolécula no GAPIN para visualização. . . 68

5.26 Tela contendo duas biomoléculas para análise 68

5.27 Processo de sincronização dos grafos em relação a estrutura. A) Pré sincronização. B) Pós sincronização 69

5.28 Visualização dos *SPOTS* ao longo do aumento da distribuição do número de grupos da 1PPF. Da esquerda pra direita de cima para baixo vemos a distribuição de grupos da 1PPF Apolar,Apolar Chain vs Chain começando 2 grupos até 7. 70
5.29 Tela de seleção de alinhamento com outra biomolécula 71
5.30 Tela de alinhamentos da 1PPF com 1TEC. 72
5.31 Tela de visualização dos alinhamentos selecionados pelo usuário entre a 1PPF e 1TEC 73
5.32 Representação 2D da ASA - *Accessible Surface Area* tal qual definidos pioneiramente por Lee & Richards (Lee & Richards [1971]). Uma sonda (*probe*) é rolada entorno dos átomos, cujo volume é definido pelos respectivos raios de van der Waals, computando uma unidade de área nas regiões de contato.(figura adaptada de Keith Callenberg, Wikipedia commons, CC BY-SA 3.0). 74
5.33 Gráficos de correlação BARS x ΔASA envolvendo 68 complexos filtrados do *Affinity Database 2.0* com I-RMSD$<$ 1.5. 77
5.34 Gráfico da distribuição do tamanho dos grupos envolvendo 15 complexos da base de treinamento do programa de predição de *Spots* APIS (Xia et al. [2010]). O tamanho do agrupamento foi normalizado por *z-score* para cada complexo. Constata-se que quanto mais energético é um *Spot*, maior é o tamanho relativo dos grupos em que ele está. 79
5.35 Análise de *Spots* da 2PTC, uma Tripsina Bovina em complexo com seu principal inibidor BPTI - *Bovine Pancreatic Trypsin Inhibitor*. (A) Grafo de baixo nível com 5 partições, coloridas de forma a discriminá-las, compreendendo todos os tipos de contatos (apolares e polares). (B) Grafo de alto nível. O alvo da mutação na cadeia BPTI (LYS15) pela alanina está no maior grupo, em verde claro. (C) Grafo de baixo nível no contexto geral da tripsina em $surface$ Connolly. (D) No grafo de alto nível, se o usuário clicar no nó verde claro , esse grupo pode ser isolado dos demais. (E) Grupo destacado colorido por cadeia (Tripsina em vermelho, BPTI em azul) e com representação de $surface$ do inibidor. É possível ver melhor agora que este grupo capturou centro ativo da tripsina, o que inclui o bolsão de especificidade. (F) Foco nos átomos com interações polares. (G) Foco em átomos com interações apolares. Nota-se que o bolsão é essencialmente polar. **(H)** Dois importantes resíduos da interface formando uma ponte salina: ASP189 (Tripsina) e LYS15 (BTPI), destacando a propensão das tripsinas em receber resíduos carregados positivamente no bolsão catalítico. 81

5.36 Complexo cadeia-ligante (ANY-LIG) do PDBid 3Q70 - Aspártico Protease do vírus da AIDS (HIV) e o inibidor péptido-mimético Ritanovir, particionado em 8 grupos. A) Grafo de baixo nível colorido conforme grupos. B) Grafo de alto nível, com o nó de maior tamanho em roxo. C) Estrutura sincronizada aos grafos, com a protease renderizada em *surface* e *ball+stick*, estando o inibidor Ritanovir em *spacefill*. O grupo ao qual o Ritanovir pertence é o maior (em roxo). 82
5.37 Alinhamento da 1PPF com 8 complexos da tabela (5.3). O alinhamento da 1PPF com 1TEC foi exibido na figura (4.7). O alinhamento foi feito pelo algoritmo *Topos*, com ajustes manuais para melhor visualização. Isso inclui também o ajuste manual de cores entre nós correspondentes, mas tudo feito através do GAPIN. Um editor de imagens foi usado apenas para compor a figura final. 84
5.38 Gráfico da distribuição de RMSDs calculados sobre as coordenadas dos centróides dos grafos de alto nível da figura (5.37). 85

Lista de Tabelas

4.1 Tabela exemplificando o cálculo do índice de preservação (Pre_{index}). Na comparação $3 \Rightarrow 4$, sobrepõe-se o grafo de 3 partições contra o de 4. Vê-se pelas áreas $A3$ e $A4$ que as sobreposições de nós $1-1$ e $3-3$ atendem ao critério de preservação ($Dist < 1.4$ e $r > 0.85$). Mas, isso não acontece para $2-2$. Na Comparação $4 \Rightarrow 5$, todos atendem ao critério de preservação. Mas, a sobreposição $2-2$ advém de uma sobreposição não preservada na comparação anterior. Logo, é feita a média: $(0.00 + 0.97)/2 = 0.48$. 41

4.2 Classificação apolar x polar de interações atômicas utilizadas no GAPIN. . 42

5.1 Comparação de parâmetros termodinâmicos para Elastase Pancreática de Porco com o Inibidor Ovomucoide de Peru. Parte experimental feita com ITC - *Isothermical Titration Calorimety*, 25° C. Parte computacional feita a partir de ASA polar e apolar da estrutura PDBid 3EST superimposta à 1PPF. Foi feito assim porque não havia estrutura resolvida do complexo da Elastase de Porco com Inibidor Ovomucoide. Para detalhes metodológicos vide (Baker & Murphy [1997]). 75

5.2 Testes estatísticos para as distribuições de tamanho dos grupos (*p-values*) referentes aos dados da figura (5.34) 80

5.3 Complexos serino-peptidase e inibidores, conforme (Gonçalves-Almeida et al. [2011]). *Clan* refere-se às estruturas terciárias relacionadas e *family* às sequências relacionadas, da classificação de peptidases do MEROPS (Rawlings et al. [2018]). *Class* e *Fold* são classificações estruturais do SCOP (Murzin et al. [1995])). 83

Sumário

Agradecimentos **ix**

Resumo **xv**

Abstract **xvii**

Lista de Figuras **xix**

Lista de Tabelas **xxvii**

1 Introdução **1**

2 Objetivos **11**

2.1 Objetivo Geral 11

2.2 Objetivos Específicos 11

3 Revisão da Literatura **13**

4 Materiais e Métodos **29**

4.1 Cálculo da área de contato 29

4.2 Definição das interfaces moleculares 31

4.3 Grafos das interfaces moleculares 33

4.3.1 Grafos de baixo nível 34

4.3.2 Grafos de alto nível 35

4.3.3 Grafos em interfaces cadeia-cadeia e cadeia-ligante 37

4.4 Agrupamento em grafos 37

4.5 Qualidade dos agrupamentos 39

4.6 Varredura de agrupamentos 40

4.7 Índice de preservação do nó 41

4.8 Tipos de interações 41

4.9 Alinhamento de grafos de alto nível 42
4.10 Sincronização visual 45
4.11 GAPIN - Engenharia de Software 45
4.12 GAPIN - Arquitetura 46
4.12.1 *Front end* 46
4.12.2 *Back end* 46

5 Resultados e Discussões **49**
5.1 GAPIN - Usabilidade 49
5.1.1 GAPIN - Fluxo de Utilização 49
5.2 Experimentos 73
5.2.1 Áreas de Contato 73
5.2.2 Spots 77
5.2.3 Alinhamentos 83

6 Conclusões e Perspectivas **87**
6.1 Perspectivas 90

Referências Bibliográficas **93**

Capítulo 1

Introdução

Para alguns, vivemos uma mudança de paradigma na forma como pensamos e interpretamos os fenômenos biológicos (Marcum [2008];Noble [2010]). Para esses defensores, a visão que orientou as biociências por quase todo o século XX foi em boa medida reducionista, no sentido de tentar entender a complexidade da vida decompondo-a em subproblemas, que foram decrescendo em escala conforme os avanços das técnicas experimentais e dos modelos teórico-computacionais, até chegar ao nível atômico. Como peças de um gigantesco quebra-cabeça multidimensional, as partes seriam remontadas e encaixadas umas às outras novamente, de maneira a gerar componentes maiores, em níveis crescente de complexidade, até que um todo, compreensível, se revelasse por si só.

Não há como negar a importância desse modelo mecanicista e seu mapeamento de componentes, sejam eles genes, transcritos, proteínas ou qualquer outra molécula bioativa. Em uma boa sorte de casos ajudou a elucidar doenças, compor diagnósticos confiáveis e terapias efetivas (Green et al. [2017]). Mas, parece, não foi suficiente para quebrar o hermetismo do todo, nem eficiente no enfrentamento de certas enfermidades complexas, como o câncer, diabetes e certas neurodesordens (Yan et al. [2017]).

Nesse sentido, outras abordagens foram ganhando força, resgatando visões mais sistêmicas e holísticas da vida (Bernard [1865];Von Bertalanffy [1950]), ainda que o debate epistemológico pareça não ter fim sobre as reais fronteiras e contrastes entre reducionismo e holismo (Gatherer [2010]). Entre as muitas versões dessas abordagens, talvez a que tenha chamado mais atenção atenda pelo nome de Biologia de Sistemas (Kitano [2002])). Nela, os componentes cedem importância para a rede de interações entre eles. Mapear essas redes em interactomas é só um ponto de partida (Cafarelli et al. [2017]). É preciso entender a sua organização estrutural e espacial, bem como sua dinâmica, seus mecanismos de retroalimentação e controle, no tempo, frente diferentes

escalas e contextos.

Não por menos, o estudo das redes complexas, no seu sentido mais genérico, foi elevado à categoria de disciplina, pelas *US National Academies*, num relatório histórico de 2005 (Council et al. [2005]). Fala-se agora numa Ciência das Redes (*Network Science*) (Barabasi [2016]), abrindo espaço, na área biomédica, para subdisciplinas como a Medicina das Redes (*Network Medicine*) (Barabási [2007]) e a Farmacologia das Redes (*Network Pharmacology*) (Hopkins [2008]).

Essa mudança de perspectiva abre novos caminhos para fármacos inovadores, que ao invés de intervir classicamente pela competição com substratos ou pequenas moléculas endógenas, o faz pela interferência com as interfaces cadeias-cadeias (Hopkins [2008]). A indústria farmacêutica acompanha com atenção essas iniciativas, pois reconhece que está cada vez mais difícil encontrar ligantes candidatos que sobrevivam a todas as exigências do complexo processo de descoberta e desenvolvimento de fármacos (Arrowsmith [2012]).

Um notório caso de sucesso deste tipo de fármaco que intervem nas interações cadeia-cadeia em proteínas é a Tirofibana (Hartman et al. [1992]). Aprovada em testes clínicos e já em uso comercial, ela modula a agregação das integrinas αIIb e β3 (duas glicoproteínas que formam um receptor transmembrana em plaquetas) ao fibrinogênio, evento necessário para disparar a agregação plaquetária na formação de coágulos (Figura 1.1). Logo, este fármaco é usado como antitrombótico, especialmente indicado para tratamento (e profilaxia) de pacientes que sofreram (ou estão sujeitos a) isquemia cardíaca (Arkin & Wells [2004]).

Esse envolvimento de integrinas e fibrinogênio nas complexas vias que modulam o fenômeno da coagulação sanguínea exemplifica como que a fisiologia celular pode ser entendida como resultado de uma miríade de proteínas interagindo entre si e com outras biomoléculas de modo a fazer emergir toda uma gama de funcionalidades possíveis em diferentes tipos de células (Huttlin et al. [2017]). Os avanços nesse mapeamento vêm revelando uma rede de incrível complexidade (Cafarelli et al. [2017]). Mas, para surpresa dos holistas, tais redes podem comportar certas características simplificadoras, com presença de elementos hierárquicos e modulares (Ravasz et al. [2002]). Assim, determinados subsistemas parecem aninhados recursivamente em outros subsistemas, resultando numa topologia que lembra um fractal, multinível, de diferentes escalas de interações e granulosidades. Uma rede numa escala mais baixa pode virar um componente numa escala mais alta.

Por exemplo, apesar de um ribossomo poder ser visto, no nível molecular, como uma "máquina" que opera sob uma rede dinâmica e complexa de interações entre proteínas, RNAs, ligantes, íons e água, no nível celular, pode-se olhá-lo como um componente

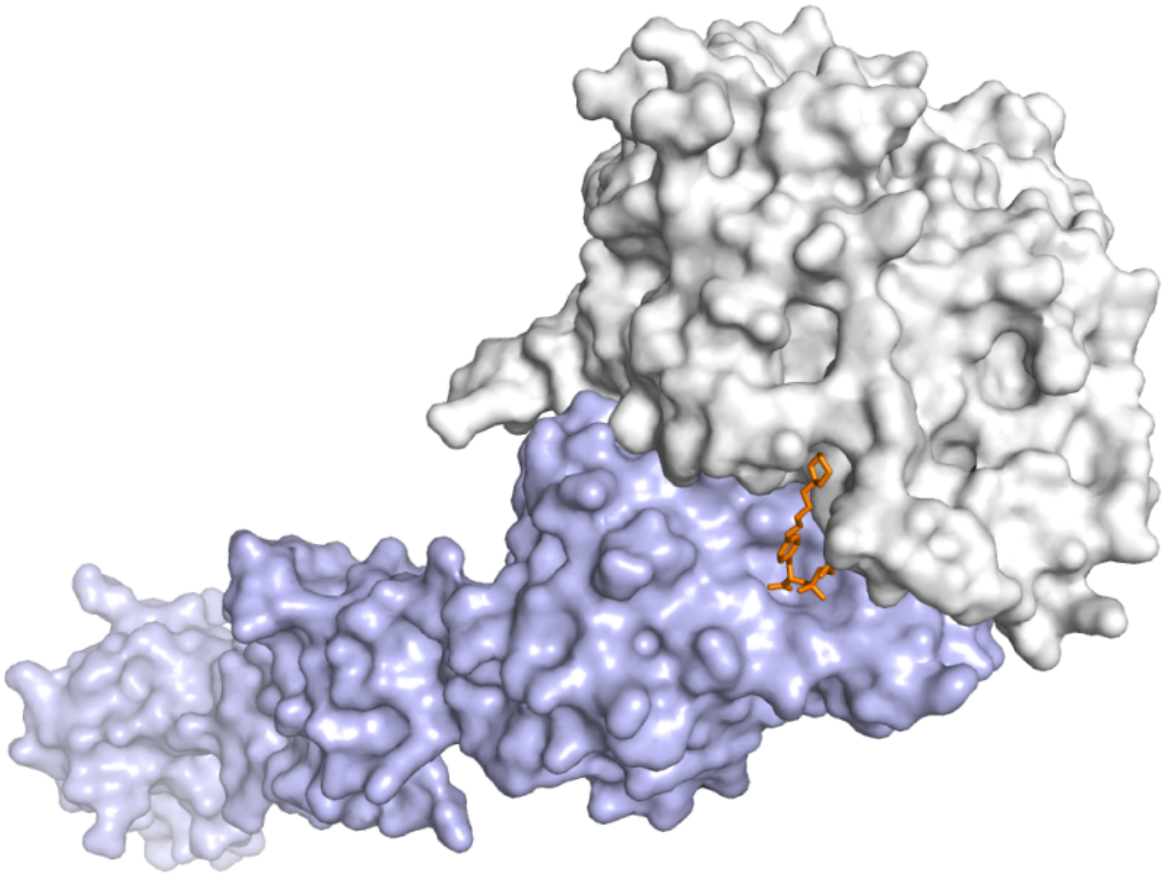


Figura 1.1: Superfícies de Connolly (Connolly [1983]) das integrinas αIIb (azul) e β3 (branco), evidenciando o fármaco ligante Tirofibana em *stick* (laranja). A Tirofibana mimetiza o tripeptídeo ARG-GLY-ASP presente numa alça do fibrinogênio, ponto principal de interação deste com o complexo das integrinas αIIbβ3. PDBid: 2VDM. Imagem gerada no Pymol versão 1.7.2.1 (DeLano [2002]).

funcional unitário em uma rede onde os nós agora são átomos ou moléculas agrupados em módulos. O mesmo pode-se pensar de outros agrupamentos triviais (neste aspecto) em escala celular, como aqueles compartimentalizados em organelas nos eucariotos: mitocôndrias, cloroplastos, sistemas flagelares etc. Mas, as redes celulares metabólicas também podem apresentar agrupamentos não triviais, sem compartimentalização física aparente, descorrentes exclusivamente da topologia modularizada da rede em si. Isto é particularmente evidente em procariotos. Em bactérias, as reações das grandes classes de metabólitos/substratos (carboidratos, lipídeos, ácidos nucleicos, aminoácidos/peptídeos) formam grupos com relativa modularidade dentro da rede metabólica que, de alguma forma, se auto-organiza em seu citoplasma (Ravasz et al. [2002]).

Dados os avanços crescentes das técnicas experimentais, do poder de processamento, transferência e armazenamento de dados dos computadores, e da efetividade

das simulações, já não é mais tão ficção a ideia de um microscópio computacional (Dror et al. [2012]), uma sofisticada aplicação capaz de perscrutar virtualmente e analiticamente um fenômeno biomolecular/celular em diferentes *zooms* espaciais e temporais. Com apoio de recursos avançados em computação gráfica e inteligência artificial, seria possível "navegar" entre as estruturas e simulações, usando diferentes abstrações, filtros, granularidades e escalas, do nível celular ao atômico (quem sabe do anatômico/fisiológico ao subatômico/quântico).

Do ponto de vista das abstrações, poderia-se passar por diferentes *zooms* entre redes de redes de interações recursivamente aninhadas, dada a formação de comunidades de comunidades de nós entrepostas. Por exemplo: a partir de uma rede de baixo nível envolvendo interações atômicas, comporia-se uma nova rede de nível mais alto, onde os nós seriam comunidades de átomos agrupados; a partir dessa rede, aplicando-se outro agrupamento levando em conta as interações intermoleculares, substratos, ligantes e anotações funcionais, teria-se outra rede de nível mais alto, definindo partes do interactoma, em seu sentido amplo, de forma a abarcar também o metaboloma e as interações gênicas (Vidal et al. [2011]). Novo agrupamento a partir desses delimitaria outra rede de nível mais elevado ainda, onde nós seriam comunidades de interações intermoleculares, evidenciando vias de sinalização, vias metabólicas, vias regulatórias etc. E assim, sucessivamente, sempre agregando, na medida do possível, outros dados e metadados a cada recursão.

Não é preciso ser um especialista em computação para antever os enormes desafios de engenharia de *software* e *hardware* que o projeto, desenvolvimento, implantação e manutenção de uma sistema tão complexo quanto um microscópio computacional demandaria. Nesse particular, é provável que os desafios de *software* sejam bem maiores que os de *hardware*, no momento.

No que tange ao software, um grande desafio envolveria a visualização de dados num contexto *big data*. Conforme bem expressaram O'Donoghue e coautores: *how to benefit from this data deluge without being overwhelmed by it* [1] (Wong et al. [2010]). Isso implica também formas visualmente criativas de promover a integração de dados de diferentes fontes. O'Donoghue e coautores acrescentam ainda os seguintes desafios (Wong et al. [2010]):

- usabilidade: como promover uma experiência de instalação, utilização e atualização de um sistema que não seja custosa e traumática ao usuário?

- análise visual: como nem toda análise de dados complexos pode ser automatizada, parte do papel da visualização é permitir um juízo humano sobre os resultados.

[1] Em tradução livre: como se beneficiar do dilúvio de dados sem ser soterrado por ele?

O desafio é encontrar um balanço produtivo entre a automação e a avaliação humana.

- representação multiescala: como melhor definir e navegar por entre diferentes escalas ou níveis de informação?

- representações inovativas: como criar ou escolher as melhores metáforas ou abstrações visuais?

- padronizações: adoção de padrões visuais contribui para a usabilidade, embora isso possa inibir a inovação visual. Como encontrar um equilíbrio ou os devidos contextos para um ou outro?

- terceira dimensão: dependendo da complexidade dos dados, uma terceira dimensão pode ajudar no juízo humano sobre os resultados, embora isso possa exigir interatividade ou visualização *estereoscópica*.

- computação aumentada: uso de camadas visuais de informações, acrescidas conforme a interação e demandas do usuário.

No que tange ao hardware, os desafios envolveriam a concepção de uma infraestrutura de servidores e de rede de dados, segura, balanceada e de fácil escalabilidade, que possam dar suporte às demandas do *front end*. Isso implica em um sistema de computação intensiva, paralela e/ou distribuída, seja local ou em nuvens (Bell et al. [2006]).

Cabe destaque o crescente desempenho das GPUs (*Graphics Processing Units*), dada sua alta capacidade de paralelização de algumas operações recorrentes em simulações e mineração de dados (como operações matriciais) num único *chip*. Uma única NVIDIA™ Tesla V100 GPU *Accelerator* chega a comportar 5120 núcleos cuda, com um desempenho de até 15 TFLOPS em precisão simples (Smith [2017]). O problema tem sido como integrar eficientemente várias GPUs em *clusters*, dada a comunicação relativamente lenta entre nós quando comparada com sistemas de alto desempenho desenhados para aplicações específicas ou ASICs (*Application Specific Integrated Circuits*) (Dror et al. [2012]).

Nesse sentido, despontam os ASICs Anton™, projetados especificamente para simulações moleculares, da empresa D E Shaw *Research* (Shaw et al. [2014]). Já em sua segunda geração, um *cluster* de 512 nós Anton2, com 33792 núcleos, foi capaz de atingir uma taxa de simulação de 85 μs/dia para uma proteína como a dihidrofolato redutase (DHFR), com 23558 átomos. Mesmo um ribossomo inteiro, da ordem de 2.2

milhões de átomos, pôde rodar uma dinâmica molecular à taxa de 3.6 μs/dia (Pan et al. [2019]). E uma simulação do enovelamento de um domínio WW com cerca de 10 mil átomos ultrapassou a barreira de 1 milissegundo em 2 semanas de máquina (Shaw et al. [2010]).

É crescente a quantidade de métodos computacionais, bases de dados e ferramentas inovadoras propostos para a visualização e análise de dados de interações biomoleculares (O 'donoghue et al. [2018]). No entanto, tais iniciativas ainda estão longe da integração necessária a um microscópio computacional, pois tendem a focar em determinados domínios do interactoma, como nas *gene regulatory networks* (GRN) (Hecker et al. [2009])[2], nas redes metabólicas (Kanehisa et al. [2017]), nas *protein-protein interactions* (PPIs)(Vidal et al. [2011]), nas *structural interactions networks* (SINs)(Kim et al. [2006]), nas redes entre proteínas e pequenos ligantes químicos (Szklarczyk et al. [2016]), nas redes de águas (Brini et al. [2017]), para citar apenas alguns. Mesmo quando a ênfase está nas interações de proteínas com alguma coisa, as aplicações mantêm especificidades, não se generalizando facilmente para qualquer interfaceamento, independente das outras biomoléculas envolvidas.

Todos esses desafios estão na base das motivações para esta tese. A ideia de microscópio computacional serviu-nos de inspiração para um sistema que pudesse navegar visualmente e analiticamente em múltiplas escalas, trabalhando diferentes granularidades de redes, através do agrupamento em comunidades de nós, com foco inicial restrito às interfaces intermoleculares de estruturas resolvidas e depositadas no PDB (*Protein Data Bank*) (Berman et al. [2000]). Tal sistema exigiu inovações de projeto que conciliassem as demandas de visualização/análise no *front end* e processamento/armazenamento no *back end*, bem como a realidade da infraestrutura de *hardware* e rede disponíveis nas instituições envolvidas, bem distantes das sofisticações de um *cluster* Anton. Tudo isso tendo sempre em mente que o sistema deveria integrar generalidade e efetividade da forma mais sinérgica possível, permitindo ao usuário a identificação e caracterização de contextos topológicos e químicos das redes de interações entre quaisquer biomoléculas do PDB. Tal apropriação de contextos poderia facilitar investigações mais detalhadas do usuário a possíveis candidatos a fármacos, dado um conjunto de alvos terapêuticos de interesse.

No intuito de somar esforços a esses desafios, propõem-se nesta tese o sistema **GAPIN**, acrônimo para ***G**rouped and **A**ligned **P**rotein **I**nterface **N**etworks*, uma aplicação 100% web, com toda a imediata disponibilidade, portabilidade, usabilidade e conveniência que só os modernos navegadores podem oferecer (Figura 1.2). Pode ser

[2]Para siglas consolidadas na literatura, preferiu-se manter a original em inglês

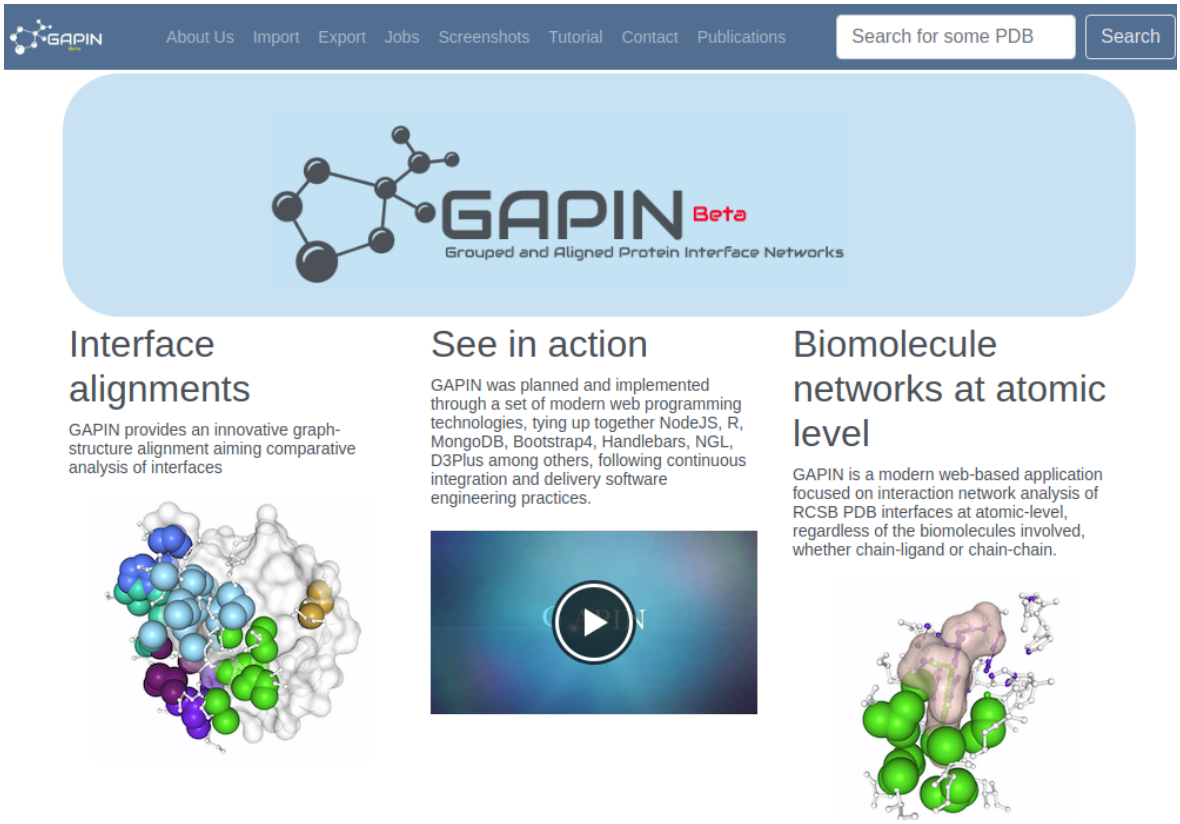


Figura 1.2: Tela de apresentação do GAPIN, com um menu destacado na tarja superior, entrada do PDBid no canto superior direito. Na parte central da página, divulgação de algumas funcionalidades, bem como um vídeo promocional.

acessada pelo endereço http://gapin.unifei.edu.br.

Seu propósito maior é permitir uma análise visual das interfaces entre biomoléculas do PDB, decompondo-a em dois tipos de representação ou abstração: de um lado, as redes de contatos intermoleculares; de outro, as estruturas atômicas renderizadas; mas com ambos os lados mantendo sincronismo visual e de operações. Para tanto, as redes são definidas como grafos bipartidos ou k-partidos não direcionados em que nós podem ser átomos ou grupos de átomos (conforme a granularidade), e as arestas têm pesos definidos pelas áreas de contatos interatômicos. O grafo é bipartido ou k-partido porque somente são permitidas arestas entre átomos de cadeias diferentes (Figura 1.3).

Ao longo da pesquisa e desenvolvimento do GAPIN, uma série de inovações algorítmicas e de projeto foram sendo incorporadas, com geração de alguns resultados inéditos e inesperados.

Uma primeira inovação diz respeito à definição de interface entre biomoléculas. GAPIN tem como principal entrada de dados os arquivos PDBs, definindo interfaces entre biomoléculas como redes em nível atômico. O mais comum na literatura,

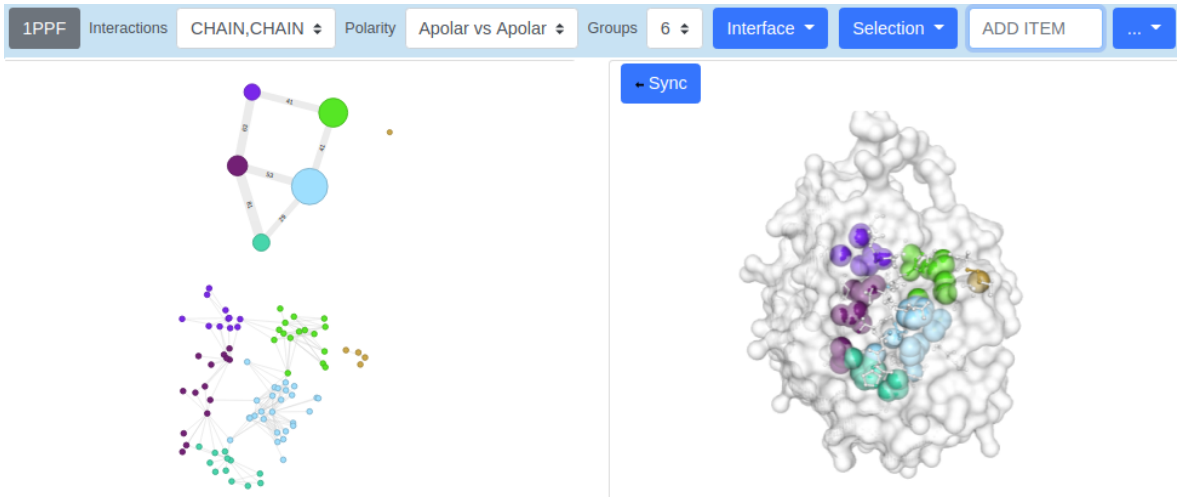


Figura 1.3: Exemplo de tela após entrada de um PDBid, como 1PPF. No lado esquerdo vemos os grafos em duas granularidades, fina ou baixo nível (abaixo); e grossa ou alto nível (acima), sendo este último resultado do agrupamento do primeiro. No lado direito, a estrutura renderizada da 1PPF, com *surface* de Connolly em uma das cadeias, *spacefill* dos átomos da interface, coloridos conforme o agrupamento do grafo.

principalmente em PPINs - *Protein-Protein Interaction Networks* é um mapeamento no nível dos resíduos (McGillivray et al. [2018]). Mas, uma granularidade no nível atômico generaliza as interações para além de proteínas-proteínas, habilitando visualizações e análises independentes das biomoléculas envolvidas, se entre proteínas, ácidos nucléicos, carboidratos, lipídios, ligantes, íons ou mesmo águas. Logo, GAPIN permite mapear a interação de qualquer-átomo com qualquer-átomo, para qualquer biomolécula resolvida no PDB, seja entre cadeia-cadeia ou cadeia-ligante.

Uma segunda inovação é que GAPIN é capaz de contrastar as estruturas PDB renderizadas com os respectivos grafos, em dois níveis de granularidade: a primeira, de baixo nível, com nós representando átomos da interface; a segunda, com nós representando comunidades de átomos (fruto do agrupamento em grafo) conforme a densidade das arestas, formando grafos modularizados ou de alto nível. Mesmo que tais representações por grafos não-modularizados e modularizados não sejam inéditas em estudos de PPINs (detalharemos isso mais adiante, na revisão da literatura), o seu uso em redes generalizadas para quaisquer biomoléculas no nível atômico é bem provável que seja.

Uma terceira, diz respeito à forma como os grafos são visualmente exibidos. Optou-se por uma representação estruturada, definindo-se *layouts* de grafos em que cada nó tem uma coordenada espacial 2D como centroide. Nos grafos de baixo nível (granularidade fina), cada nó é um átomo, e os centroides são as coordenadas 3D dos respectivos átomos (mas projetados no plano XY). Nos grafos de alto nível (granularidade grossa), cada nó tem como centroide o centro geométrico dos átomos agrupados

(conforme resultado da *clusterização*). Dessa forma, foi possível construir uma relação interativa entre estruturas renderizadas do PDB e os grafos: manipulações com o *mouse* na estrutura poderiam ter efeitos síncronos nos grafos, de modo que rotacionando a primeira, reposicionaria também o segundo.

Todas essas inovações estão amparadas por várias opções de seleção de objetos e representações visuais, seja no lado dos grafos, seja no lado das estruturas. No lado dos grafos, tanto baixo quanto alto nível, é possível selecionar e destacar nós. No lado das estruturas, é possível selecionar cadeias, resíduos, átomos, com diversas opções de: representação: *ball+stick*, *spacefill*, *surface*; e de cores: por grupos, cadeias, elementos, interações. Tudo isso aplicável ao complexo como um todo ou somente à seleção feita previamente.

GAPIN agrega ainda duas ferramentas auxiliares de análise. A primeira chamada de *Spots*, usa uma metodologia de varredura (da Silveira et al. [2009a]) de agrupamentos, sob número crescente de grupos, no intuito de demarcar e permitir a investigação de partições do grafo que sejam mais densamente conectadas. De forma inesperada, descobriu-se no desenrolar desta tese que o volume das partições correlaciona-se com a probabilidade da mesma poder abrigar *Hot Spots*, termo consagrado na literatura pelos estudos mutagênicos pioneiros de Clackson & Wells na interface entre hormônio de crescimento humano (hGH) e seu receptor (hGHbp) (Clackson & Wells [1995]), referindo-se a grupo de resíduos que tenham significativa contribuição à variação da variação da energia livre de Gibbs nas interações cadeia-cadeia ($\Delta\Delta G$ de *binding*). Quanto mais energético é um *Spot*, mais chance da região em que ele está servir como um alvo drogável referencial para candidatos a fármacos (Cukuroglu et al. [2014]).

A segunda ferramenta permite aferir similaridade de grafos por meio de alinhamentos. Tal alinhamento pode ser feito de forma manual ou automática. Para esta última, criou-se um algoritmo inédito chamado *Topos* que mede a similaridade de grafos levando em conta tanto as suas topologias (suas formas) quanto as posições dos centroides de cada nó. Será mostrado como esta ferramenta permite comparar interfaces de complexos envolvendo diferentes peptidases e inibidores, com foco apenas no perfil das hidrofobicidades, em consonância com resultados publicados anteriormente por nosso grupo de pesquisa (Alves [2015],Gonçalves-Almeida et al. [2011]). Cabe destacar que na corrente versão do GAPIN, o algoritmo *Topos* trabalha apenas com alinhamentos de interfaces cadeia-cadeia.

Um aprofundamento dessas (e outras) inovações e resultados experimentais será apresentado nos demais capítulos. Esta tese conta com a seguinte organização: o capítulo II formalizará os objetivos; o Capítulo III, a revisão da literatura; o Capítulo IV, o detalhamento metodológico; o Capítulo V, resultados dos experimentos feitos e

discussões; o Capítulo VI, conclusões e perspectivas.

Capítulo 2

Objetivos

2.1 Objetivo Geral

Pesquisar, projetar e desenvolver métodos computacionais através da disponibilização de um sistema WEB para identificação e caracterização de interfaces entre quaisquer biomoléculas existentes no PDB, seja cadeia-cadeia ou cadeia-ligante, de forma a prospectar padrões de interações que possam auxiliar na delimitação de alvos terapêuticos e indicação de candidatos a fármacos.

2.2 Objetivos Específicos

- Desenvolver um modelo robusto de definição de interfaces entre biomoléculas do PDB, no nível atômico, levando em conta o caracter polar/apolar de cada átomo;
- Desenvolver modelos de agrupamento em grafos para definições de regiões *spots* nas interfaces, bem como métricas de avaliação dos agrupamentos e resistência a reparticionamentos;
- Desenvolver modelos de similaridade de grafos e alinhamento de estruturas (par a par), visando análise comparativa de interfaces
- Incorporar desenvolvimentos anteriores em uma interface WEB que permita ao usuário interações com estruturas e respectivos grafos de contatos;
- Habilitar pelo sistema WEB a possibilidade de importação personalizada de arquivos PDB e exportação das matrizes de adjacências dos grafos;

Capítulo 3

Revisão da Literatura

Nos dias de hoje, o gerenciamento da bibliografia de uma tese pode ser feito por ferramentas colaborativas capazes de organizar e compartilhar documentos em formato PDF. Nesta tese, foi usado do Mendeley (Mendeley [2019]). Estão destacados em *favoritos* 212 documentos, compondo a base da leitura de artigos que fundamentaram esta tese, compartilhados entre orientador e doutorando. Esse número não inclui leituras de documentos não-PDFs lidos diretamente na Internet. No entanto, para este tópico de revisão da literatura, serão destacados apenas os considerados mais relevantes em termos de ferramentas e métodos concorrentes.

Conforme apresentado na Introdução, GAPIN almeja o uso de redes de contatos para delimitar interfaces entre quaisquer biomoléculas do PDB (não somente entre proteínas), sejam elas cadeia-cadeia ou cadeia-ligante. Essa abrangência, num contexto de redes de contatos, foi uma nas inovações desta tese. Em nossa revisão da literatura, não foi encontrada nenhuma ferramenta com essa generalidade de escopo, embora GAPIN possa ter interseções com várias delas.

Dentre as que fazem uso de redes de contatos nas delimitações de interfaces, uma das mais importantes e conhecidas envolvendo PPIN foi construída com base em *plug-in's* da plataforma *Cytoscape* (Shannon et al. [2003]): *NetworkAnalyser* e *RINalyzer* (Doncheva et al. [2012]).

Embora originalmente desenhada para trabalhar com redes biológicas, por pesquisadores do *Institute for Systems Biology*, Seattle/USA, *Cytoscape* é hoje uma plataforma genérica, de código aberto, para análise e visualização de quaisquer tipos de redes complexas, mantido e desenvolvido por um consorcio internacional (Shannon et al. [2003]).

Cytoscape é subdividida em duas plataformas. A primeira é a versão *desktop* desenvolvida por meio da linguagem Java, hoje limitada na versão 8 lançada oficialmente

em março de 2014 e suporte previsto para 2019/2020. Atualmente a versão Java está na *release* 12.0.1 o que pode implicar não só evoluções da linguagem como pacotes de correções de segurança.

A utilização da versão *desktop* do *Cytoscape* exige uma configuração avançada do usuário por meio do terminal *shell* em que são necessários alguns conjuntos de *scripts* a serem executados a fim de instalar a ferramenta. Como a versão *desktop* precisa de um Sistema Operacional robusto para sua execução, não é possível por exemplo sua utilização através de *tablets* ou qualquer outro dispositivo móvel como *smartphones*.

O grupo também disponibiliza uma biblioteca em Javascript para visualização e análise de redes. Todavia, essa biblioteca não é um sistema WEB e sim um conjunto de APIs que podem ser utilizadas em uma aplicação para construção de um sistema de visualização de redes.

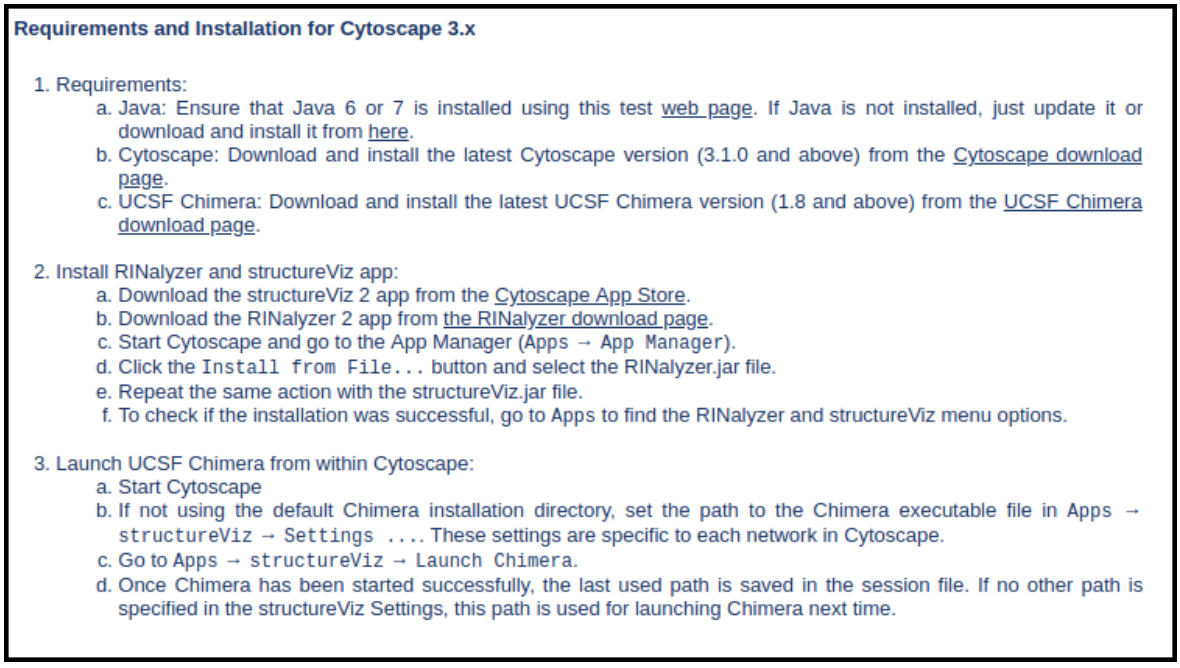
Em tempo de definição arquitetural do GAPIN, tomou-se a decisão de não utilizar o *Cytoscape* e derivados devido ao alto tempo de renderização em comparação ao NGL e ao D3 Plus JS, bem como a facilidade de integração com PDBs já nativos do NGL. Outro ponto fundamental durante a tomada de decisão foi a adesão da comunidade na utilização de ferramentas de visualização de proteínas e redes de interações.

NetworkAnalyzer é um *plug-in* do *Cytoscape* capaz de computar e exibir uma série de parâmetros topológicos de importância em redes biológicas, incluindo também caracterizações e visualizações gráficas conforme diversos modelos estatísticos de redes, tais como sem escala (*scale free*) ou mundo pequeno (*small word*) (Assenov et al. [2008]).

RINalyzer, juntamente com *RINerator* são duas ferramentas baseadas em *Cytoscape* capazes de montar representações 2D das redes de interações inter-resíduos (RIN - *Residue Interaction Networks*) tendo como referência estruturas 3D de proteínas no PDB (Doncheva et al. [2011]). Enquanto *RINerator* gera as RINs, *RINalyzer* monta visualizações simultâneas entre as RINs e estruturas renderizadas de proteínas do PDB, e o faz isso tendo com visualizador dessas estruturas o UCSF - Chimera (Pettersen et al. [2004]). *RINalyzer* tem como centro de desenvolvimento o *Max Planck Institute for Informatics*, na Alemanha.

O primeiro ponto a destacar nessa ferramenta é o relativamente complexo processo de instalação. A Figura 3.1 mostra os passos necessários. Já que é baseado em *Cytoscape*, ele precisa de Java (atualmente, na versão 8; ainda não homologado para a versão 9 mais recente)[1]. Após o Java, tem que instalar o *Cytoscape* (foi utilizado a versão 3.7.0). A seguir, os *plug-in's RINalyzer* e *structureViz 2*. Por fim, instalar o

[1]As instruções do site *RINalyzer* pedem Java versão 6 ou 7, mas funcionou bem em nossos testes na versão 8, em Ubuntu 16.04.



Requirements and Installation for Cytoscape 3.x

1. Requirements:
   a. Java: Ensure that Java 6 or 7 is installed using this test web page. If Java is not installed, just update it or download and install it from here.
   b. Cytoscape: Download and install the latest Cytoscape version (3.1.0 and above) from the Cytoscape download page.
   c. UCSF Chimera: Download and install the latest UCSF Chimera version (1.8 and above) from the UCSF Chimera download page.

2. Install RINalyzer and structureViz app:
   a. Download the structureViz 2 app from the Cytoscape App Store.
   b. Download the RINalyzer 2 app from the RINalyzer download page.
   c. Start Cytoscape and go to the App Manager (`Apps → App Manager`).
   d. Click the `Install from File...` button and select the RINalyzer.jar file.
   e. Repeat the same action with the structureViz.jar file.
   f. To check if the installation was successful, go to Apps to find the RINalyzer and structureViz menu options.

3. Launch UCSF Chimera from within Cytoscape:
   a. Start Cytoscape
   b. If not using the default Chimera installation directory, set the path to the Chimera executable file in `Apps → structureViz → Settings ...`. These settings are specific to each network in Cytoscape.
   c. Go to `Apps → structureViz → Launch Chimera`.
   d. Once Chimera has been started successfully, the last used path is saved in the session file. If no other path is specified in the structureViz Settings, this path is used for launching Chimera next time.

Figura 3.1: Instruções para instalação do *RINalyzer* conforme disposto no site https://rinalyzer.de/docu/install.php

UCSF - Chimera. Para cada passo, há de se checar se a instalação ocorreu sem erros, o que não costuma ser a norma em alguns contextos de máquinas.

Após enfrentar a instalação, ocorrendo tudo bem, pode-se iniciar o processo de uso da dupla *RINalyzer* e UCSF - Chimera (Figura 3.2). A curva de aprendizagem parece longa, pois envolve familiarização com dois sistemas projetados independentemente, e juntados num só. Isoladamente, são sistemas bem completos e eficientes, próximos do estado da arte no domínio particular de cada um. Mas, a junção de ambos num sistema só não parece se integrar de forma harmônica, comprometendo usabilidade em nome da versatilidade. Cada um opera numa janela em separado, oferecendo uma gama de opções nativas muito além do que o usuário necessitaria para o foco principal de se delimitar e estudar interfaces. Não por menos, os autores publicaram um artigo de 16 páginas no *Nature Protocols* com um longo tutorial passo-a-passo de uso da ferramenta (Doncheva et al. [2012]), com um tempo estimado de execução de 2 horas (não inclusos instalação).

RINalyzer, como o próprio acrônimo diz, trabalha com RINs - ***Residue*** *Interaction Networks*), operando redes de contatos no nível de resíduos. Se por um lado tal granularidade permite algumas análises consistentes quando os resíduos são mais homogêneos em suas propriedades químicas, ela dificulta em caso de heterogeneidade. Por exemplo: leucina tem uma cadeia lateral toda hidrofóbica. Essa uniformidade implica quase indiferença se em nível atômico ou de resíduos, dependendo da vizinhança.

Mas para lisina já não é assim, pois tem uma cadeia carbônica alifática ligada a um grupo amino polar ou carregado positivamente (dependendo do pH). Trabalhar no nível atômico facilita enxergar essa pluralidade farmacofórica de resíduos como a lisina, e compor redes de contatos quimicamente e topologicamente mais informativas. Já com *RINalyzer*, como opera no nível de resíduos, há uma dificuldade maior de entender ou separar esses acoplamentos heterogêneos. A última versão do *RINalyzer* disponível no *App Store* da *Cytoscape* é a 2.0.0, de 2014, o que leva à especulação de que a ferramenta pode ter tido seu desenvolvimento interrompido ou descontinuado.

Como visto anteriormente, GAPIN opera de forma bem diferente do *RINalyzer*. A começar pela instalação. GAPIN, como uma aplicação *web*, não demanda qualquer atividade de instalação, e seu uso é imediato a partir de um navegador moderno. Outras diferenças: GAPIN foi projetado desde a concepção para ser uma aplicação integrativa e sinérgica entre redes de contatos e visualização estrutural de interfaces, configurando-se num *front end* enxuto e com recursos focados nesse tipo de análise. Ao contrário de *RINalyzer*, GAPIN constrói redes de contatos em nível atômico, oferecendo uma análise mais rica dos contatos. Também as arestas no GAPIN tem como peso as áreas de contato, ao contrário do *RINalyzer* que usa critérios de distância sem pesos para as arestas (pelo menos, na condição default). Já destacamos dois recursos importantes do GAPIN: agrupamento gerando *Spots* e alinhamento de interfaces. Embora não pareça impossível compor e/ou agregar *plug-in's* do *Cytoscape* para atividades semelhantes, não é algo trivial de ser feito, especialmente para alguém sem habilidades em programação. Por fim, mas não menos importante, GAPIN é uma ferramenta cliente-servidor colaborativa no sentido de que quanto mais é usado, mais completo fica. *RINalyzer* opera por default só no lado cliente, a menos que seja adaptado algum *plug-in* de colaboração.

Uma outra ferramenta que faz um bom uso do paralelismo de análise entre redes de contatos e estruturas PDBs é o NAPS - *Network Analysis of Protein Structures* (Chakrabarty & Parekh [2016]), desenvolvido por pesquidores do *International Institute of Information Technology*, em Hyderabad, India. NAPS tem um foco maior na parte de análise das rede de contatos, mas somente no nível de resíduos. Os contatos são todos definidos por critérios de distância, com opção de escolha de diferentes centroides, com carbonos alfas, betas, átomos mais próximos entre dois resíduos, centro de massa. NAPS computa contatos tanto intra quanto intercadeias, seja para proteína-proteína, proteína-RNA ou proteína-DNA. Escolhida a rede, é possível investigá-la sob vários parâmetros tais como: grau dos nós (e grau médio da vizinhança dos nós), diversas medidas de centralidade (*closeness*, *betweeness*, *eigenvector*), de agrupamento (*clustering coefficient*), excentricidade, força (*strength*) etc. Tudo isso pode ser visto

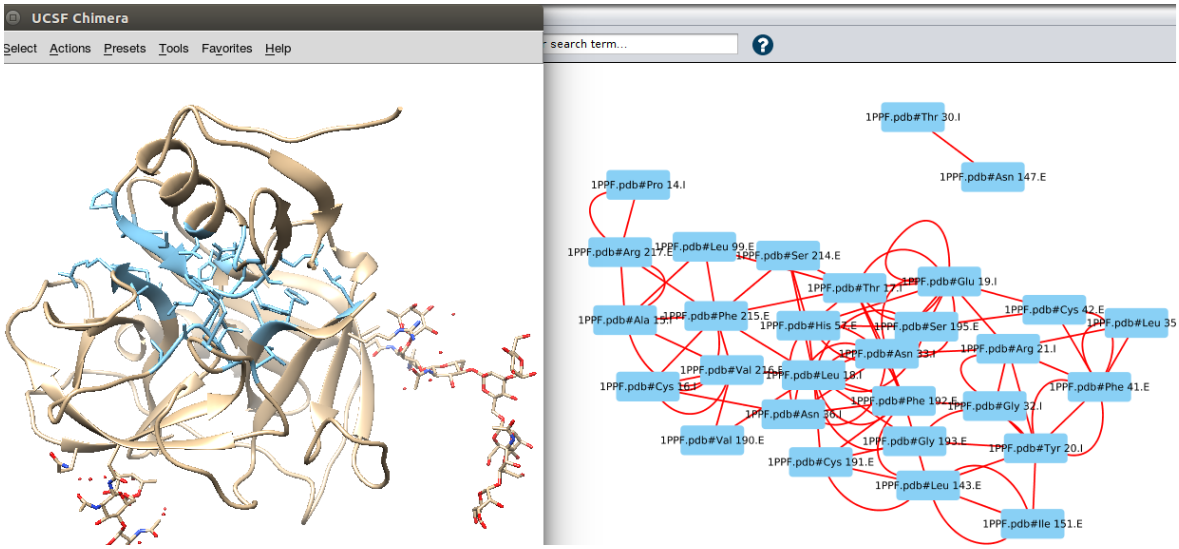


Figura 3.2: UCSF - Chimera (janela esquerda) e *RINalyzer Cytoscape* (janela direita) para 1PPF, destacando interface entre uma Elastase de Leucócito Humano e um Inibidor de Ovomudoide de Peru. A interface na estrutura e rede de contatos estão sincronizadas visualmente, em azul.

por gráfico ao longo da estrutura primária. Tal como GAPIN, a rede de contatos é estruturada por um *layout* que dá posição aos nós num espaço 3D. Tal como GAPIN também, NAPS permite visualizar contatos polares, apolares e todos, com o adicional de também carregados. Uma amostra de sua tela pode ser visto na figura 3.3.

Mas, as similaridades (parciais ou totais) entre NAPS e GAPIN param por aí. Algumas diferenças mais marcantes: GAPIN trabalha com dois níveis de redes, usando área de contatos para definir arestas. Nas redes de baixo nível, os nós estão representando átomos e não resíduos; nas redes de alto nível, agrupamento de átomos. GAPIN permite manter sincronizado rede e estrutura; em NAPS, elas estão assíncronas. Também permite selecionar nós em ambas as redes, com opção de isolar os nós selecionados na estrutura. Trabalhar no nível atômico permite a GAPIN generalizar o estudo de interfaces para além de cadeia-cadeia, seja entre proteínas ou não, mas também cadeia-ligantes, o que inclui águas. GAPIN tem recursos para analisar *Spots* e alinhar interfaces. Por fim, o foco maior de GAPIN nas análises das interfaces intermoleculares está mais nas estruturas que nas redes, tendo em mente sempre a possibilidade de avaliar contextos topológicos e farmacofóricos de alvos terapêuticos. A parte de rede é uma abstração auxiliar usada principalmente como forma de identificar agrupamentos de átomos densamente conectados. Foi possível indicar por estudos estatísticos que tais regiões podem abrigar *Hot-Spots*, reconhecidos na literatura como potenciais alvos de candidatos a fármacos, algo que será descrito com detalhes no capítulo - V. Por

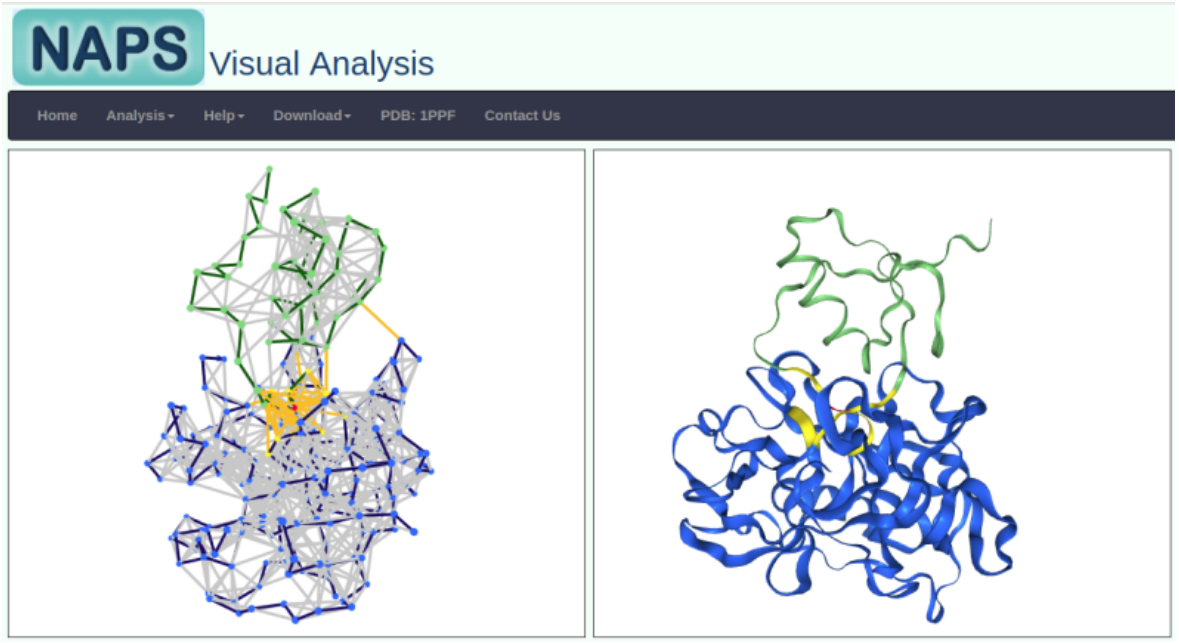


Figura 3.3: Tela do NAPS após escolha de *Protein Complex* para PDBid 1PPF - Elastase de Leucócito Humano e um Inibidor de Ovomudoide de Peru. Enzima em tons azuis, inibidor em tons verde, interface entre eles com arestas em laranja, arestas intracadeia em cinza. Centroides em carbonos alfas. O nó em vermelho representando o resíduo (I:LEU:18) no lado inibidor foi posto em destaque tanto na rede quanto estrutura. Esse resíduo ocupa o sítio de especificidade na elastase.

conta disso, GAPIN oferece uma série de opções de visualização para as estruturas, tais quais já descritas no capítulo I - Introdução. NAPS, com seu foco maior nas redes de contatos, oferece apenas uma opção de visualização de estruturas.

Outro sistema correlato é o *webservice* PDBePISA (*Proteins, Interfaces, Structures and Assemblies* - PISA), mantido e hospedado pelo *European Molecular Biology Laboratory - European Bioinformatics Institute* (EMBL-EBI) (Krissinel & Henrick [2007]). PDBePISA apresenta um *front end* interativo para exploração de interfaces intermoleculares, ofertando uma série de informações físico-químicas e termodinâmicas dos complexos, tais como: energia livre de formação, energia de solvatação, áreas acessíveis e não acessíveis ao solvente, ligações de hidrogênio, pontes salinas, interações hidrofóbicas etc. Mas, talvez seu recurso mais poderoso seja a previsão de estruturas quaternárias e montagens biológicas multiestruturais (*biological assemblies*) a partir de monômeros ou oligômeros, tendo como referência a otimização da energia livre de associação (Krissinel [2011]). Outras virtudes do PDBePISA incluem: um banco de dados passível de consulta, com informações estruturais e físico-químicas, tais como: estado oligomérico, grupos espaciais e de simetria, número de pontes salinas e dissulfetos, composição, parâmetros termodinâmicos etc; um sistema de busca e *ranking* de interfaces

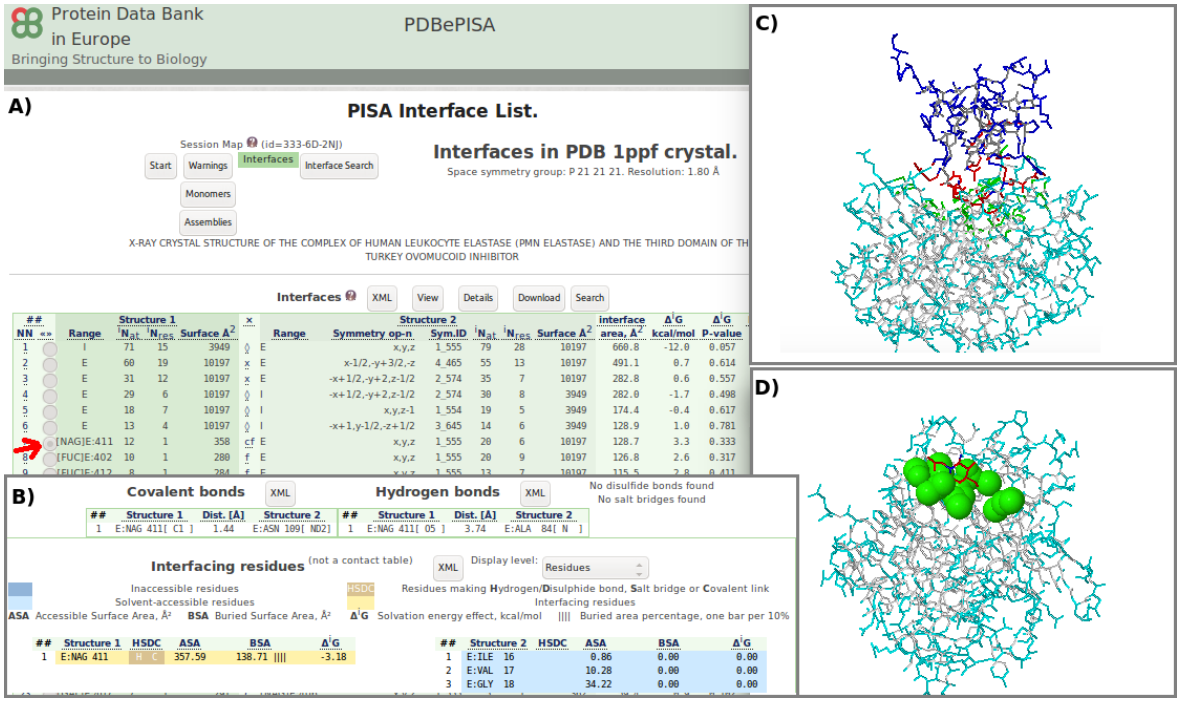


Figura 3.4: Montagem de telas do PDBePISA para 1PPF. A) Resultado do *Interface* após entrada da 1PPF. B) Resultado após escolha de *Details* em alguma linha de interação em A). C) Visualização de estruturas com destaque da interface cadeia-cadeia. D) Visualização de estruturas com destaque de uma interface cadeia-ligante, com átomos da cadeia em *spacefill* verde e do ligante (NAG - *N-Acetyl-D-Glucosamine*) em *wireframe* vermelho.

por grau de similaridade, conforme seus descritivos estruturais e físico-químicos.

Se por um lado, o grande detalhamento das informações físico-químicas e termodinâmicas oferece possibilidade de uma análise mais aprofundada, por outro lado o usuário pode facilmente se sentir perdido em meio a tanta informação. Talvez contribua para isso a apresentação dessas informações em longas e monótonas tabelas. É provável que ficasse mais "limpo" e visualmente eficiente se a informação fosse apresentada ao usuário sob demanda, e não tudo de uma vez, conforme prega o princípio da progressividade em interfaces humano-computador (Nielsen & Loranger [2006]).

Como ferramenta de análise (e previsão) de interfaces intermoleculares, PDBePISA parece bem efetivo, ainda que ele não agregue informações de redes de contatos nas interfaces, nem trabalhe com agrupamentos desses contatos. Mas, tecnologicamente, trata-se de uma aplicação que pode entrar em desuso, a se manter o *front end* como está. Além das questões de usabilidade citadas, ele emprega *Applets* Java para as visualizações de estruturas em navegadores. O problema é que, por questões de segurança e com o advento do HTML5.0, navegadores modernos como Chrome,

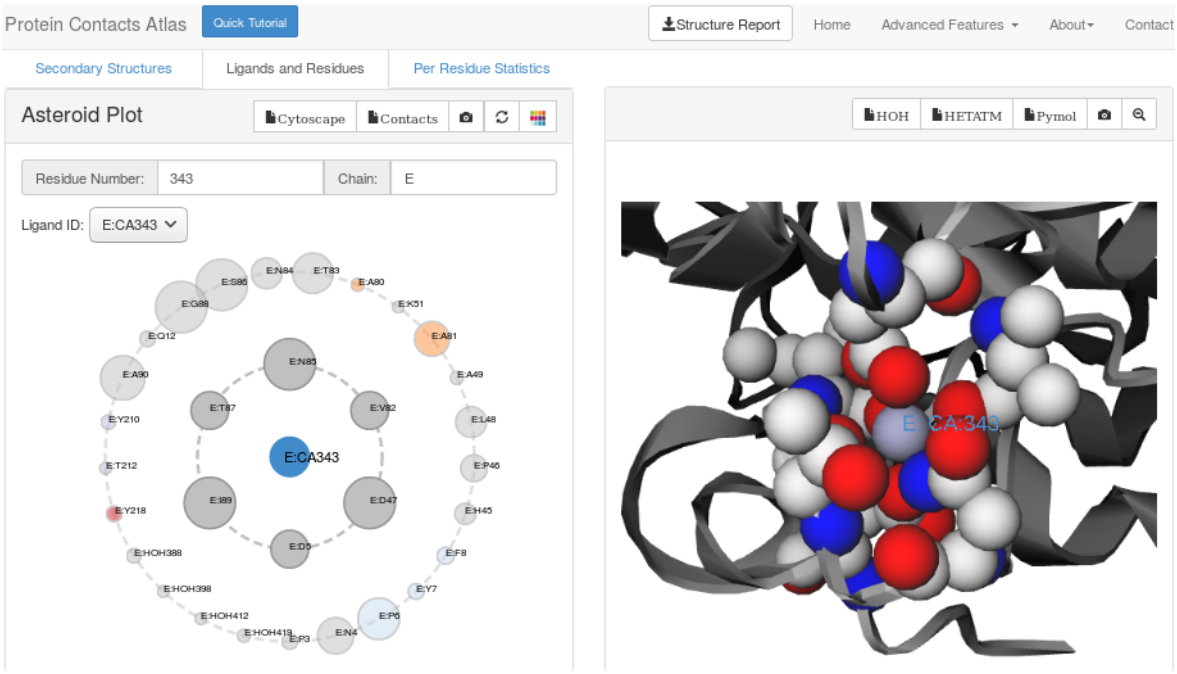


Figura 3.5: Tela capturada do *Protein Contacts Atlas*. Como não se conseguiu construir uma rede de contatos para interfaces cadeia-cadeia, segue um exemplo dos contatos cadeia-ligante para 1TEC - Subtilisina Termitase da bactéria *T. vulgaris* com Inibidor Eglina C de Sanguessuga. À esquerda, um *asteroid plots* evidencia contatos de 1a e 2a vizinhança (ordem) com o cálcio 343 da cadeia E (enzima). À direita, os contatos refletidos na estrutura, em *spacefill*, com o cálcio ao centro.

Firefox, Safari[2] e outros não mais suportam o NPAPI - *Netscape Plugin Application Programming Interface*, necessário para se instanciar as extensões Java (Hruska [2016]).

Como exemplo de um *webservice* mais moderno nessa temática tem-se o *Protein Contacts Atlas* (Kayikci et al. [2018]), hospedado no *Medical Research Council*, *Laboratory of Molecular Biology* (MRC-LMB), em Cambridge, Inglaterra. Impossível não dizer que um dos autores do referido artigo do *Protein Contacts Atlas* é o indiano, Nobel de química de 2009[3], Venkatraman "Venki" Ramakrishnan, premiado pela primeira resolução estrutural da subunidade 30S do ribossomo procarioto(MRC-LMB [2019]).

Ao abrir o *Protein Contacts Atlas* num navegador, depara-se com uma interface limpa, moderna e bem desenhada. Belas ilustrações evidenciam o poder de visualização da ferramenta, lançando mão de conceitos gráficos de última geração, como *chord plots* e *asteroid plots*. Mas, ao começar a trabalhar com a ferramenta, fica claro que os autores deram foco (talvez demasiado) à análise dos contatos a partir das estruturas secundárias. Com o *chord plots*, por exemplo, só se consegue operar se houver estrutura

[2]Para capturarmos as telas do PDBePISA exibidas aqui, tivemos que rodá-lo num navegador mais antigo e simples, como o Konqueror.

[3]Juntamente com Thomas A. Steitz e Ada Yonath.

secundária definida. Em algumas interfaces, como entre Elastase de Neutrófilo Humano e Inibidor Ovomucoid de Peru (1PPF), não existem estruturas secundárias intercadeias, o que faz com que o *chord plots* fique em branco. *Protein Contacts Atlas* permite também avaliar contatos do tipo cadeia-ligantes, sendo onde funciona melhor o seu *asteroid plots*. A parte de elementos visuais da estrutura tem limitações: não foi possível ver superfícies, não parece existir um controle de cores, e mesmo após muitas tentativas, não se teve sucesso em fazer que com a ferramenta gerasse uma rede de contatos da interface, mesmo com um PDB-id que apresenta elementos secundários na interface enzima-inibidor, como na 1TEC (Subtilisina Termitase da bactéria *T. vulgaris* com Inibidor Eglina C de Sanguessuga). Enfim, apesar de toda a primeira boa impressão da estética, não se percebeu o *Protein Contacts Atlas* como uma ferramenta efetiva na análise de contatos em interfaces intermoleculares. Um exemplo de tela do pode ser visto na figura 3.5.

No que diz respeito à base de dados de interactomas com suporte integrado de visualização das redes de interações, são muitos os *webservices* disponíveis. Uma revisão feita por pesquisadores austríacos da *Medical University Graz* (Jeanquartier et al. [2015]) avaliou 11 dessas bases com *front end web*, tendo em vista os seguintes critérios: suporte multiplataforma, interoperabilidade, integração de dados, número de interações possíveis, qualidade das visualizações e cobertura dos dados. Após avaliação, os 3 mais bem colocados foram STRING - *Search Tool for the Retrieval of Interacting Genes/-Proteins* (Szklarczyk et al. [2017]), IntAct - *Molecular Interaction Database*[4] (Orchard et al. [2014]) e CPDB - *Consensus PathDB* (Kamburov et al. [2013]). Como há razoável sobreposição de objetivos e recursos entre as 3, destacaremos somente a primeira colocada.

STRING, de fato, aparece como estado-da-arte em base de dados e recursos *web* em algumas revisões em PPI (Jeanquartier et al. [2015], Cafarelli et al. [2017]). É mantido por um consórcio multinacional, tendo como instituições pilares: o SIB - *Swiss Institute of Bioinformatics*, na Suiça; o CPR - NNF *Center for Protein Research - Novo Nordisk Foundation*, ligado à *University of Copenhagen*, na Dinamarca; e o intergovernamental EMBL - *European Molecular Biology Laboratory*. Trata-se de um grande integrador e consolidador de dados de PPIs, sejam indicados por aparatos experimentais ou preditos, envolvendo vasto número de organismos. Eis algumas estatísticas da sua base de dados, encontradas em seu site na internet: 24.5 milhões de proteínas; 52.9 milhões de interações com *score* maior que 0.9; 3.0 bilhões de interações com *score* maior que 0.1; 5090 organismos, sendo 477 eucariotos.

[4]IntAct fundiu-se a outro data base de interações chamada MINT (Licata et al. [2012]), resultando no MIntAct (Orchard et al. [2014])

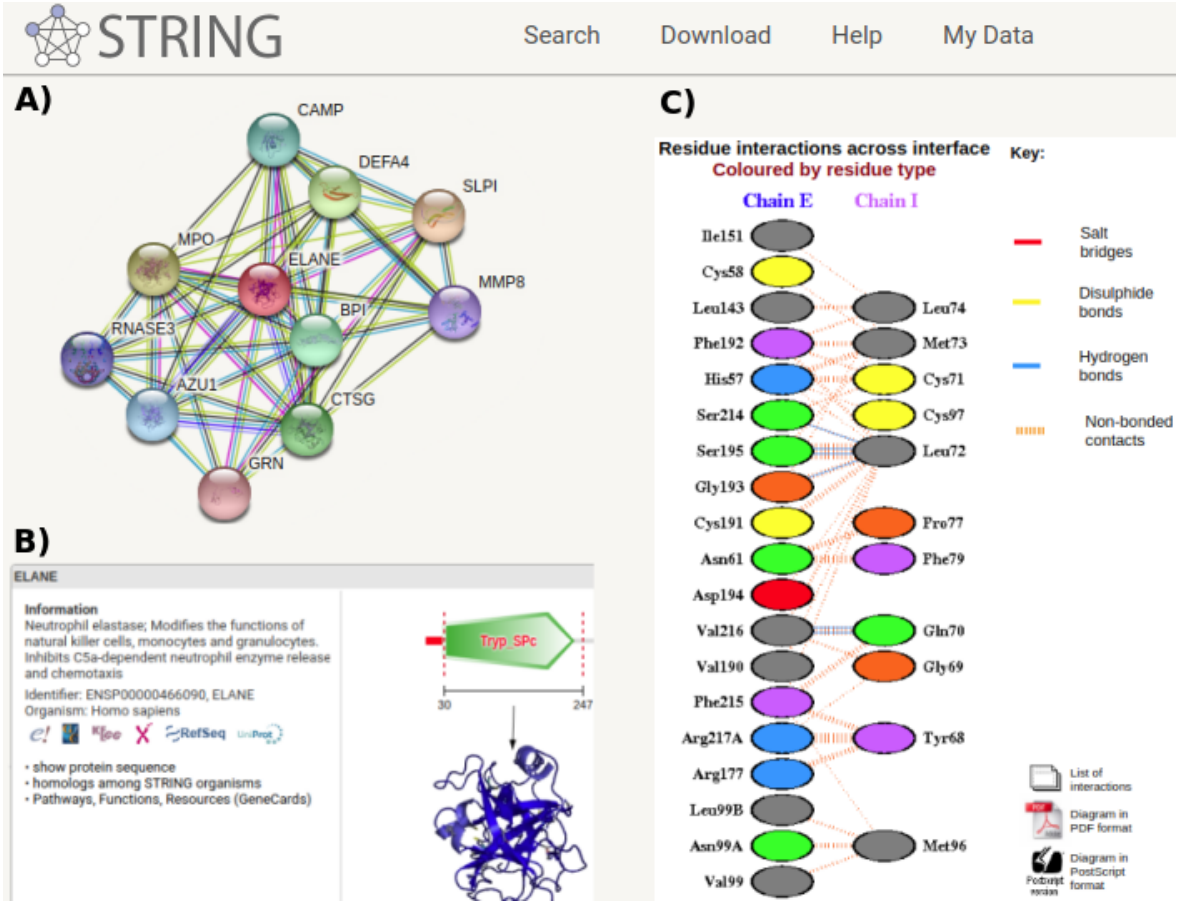


Figura 3.6: Tela capturada (e adaptada) do STRING. A) Resultado da consulta para o UniProtKB id P08246, que refere-se Elastase de Neutrófilo Humano (ELANE). B) Resultado de um duplo *clique* no nó ELANE, trazendo algumas informações anotadas. C) Ao clicar na imagem da estrutura, abre-se uma tela do PDBSum ((Laskowski et al. [1997]). Na aba *Prot-Prot* vem uma imagem estática contendo as interações da interface da ELANE (cadeia E) com inibidor antileucoproteinase SLPI - *Secretory Leukocyte Protease Inhibitor* (cadeia I)

A parte que interessa a esta tese é verificar se um sistema tão complexo e abrangente quanto o STRING concebe algum tipo de *zoom* em sua rede de interações, do nível mais alto (interactoma) ao nível mais baixo (resíduos/átomo) nas interfaces cadeia-cadeia. A figura 3.6 mostra o resultado de uma consulta com o UniProtKB id P08246, referente à Elastase de Neutrófilo Humano da nossa já conhecida 1PPF. Na figura 3.6A é possível ver a rede no contexto geral do interactoma, como um agrupamento local, o que inclui ao centro a referida elastase com a sigla ELANE. Clicando no nó ELANE, vai-se para a janela da figura 3.6B, com algumas informações das anotações feitas. Um outro clique na imagem da estrutura em destaque, já nos leva para fora do ambiente STRING, surpreendentemente para uma nova janela com resultados do PDBSum do

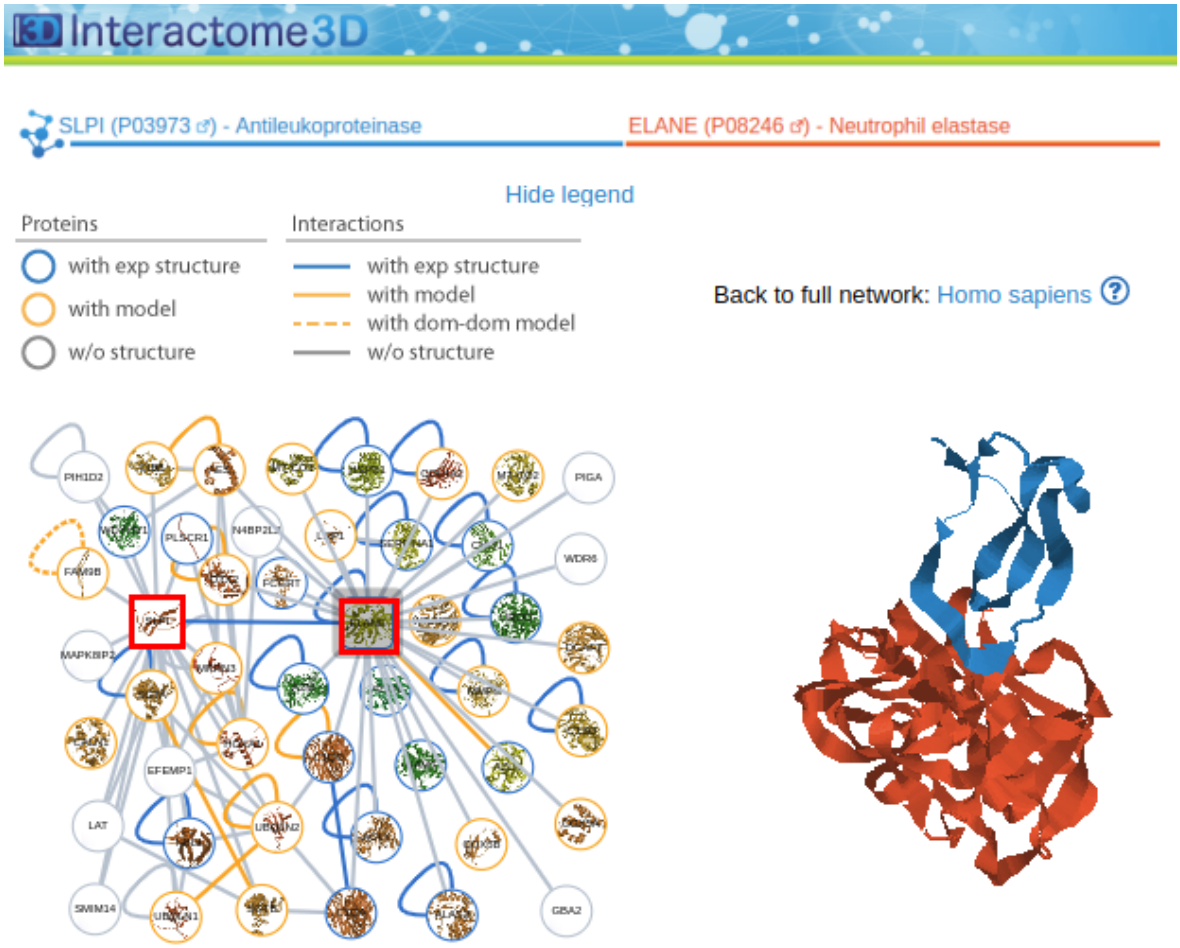


Figura 3.7: Tela capturada (e adaptada para melhor visualização) do *webservice Interactoma3D*, indicando a interação entre uma elastase de neutrófilo humano e um inibidor antileucoproteinase SLPI - *Secretory Leukocyte Protease Inhibitor* - SLPI, destacados em quadrados vermelhos na rede do interactoma. Vê-se que tanto a elastase quanto o SLPI fazem interações com uma série de outras proteínas, nem todas com estruturas resolvidas ainda. PDB id: 2Z7F.

PDBid: 2ZTF. PDBSum (Laskowski et al. [1997]) é um *webservice* antigo, lançado em 1997, e retorna um sumário consolidado das informações estruturais de biomoléculas depositadas no PDB. Na aba Prot-Prot do PDBSum é possível ver a imagem estática na figura 3.6C, sem nenhuma interatividade, envolvendo a interface da ELANE com o inibidor antileucoproteinase SLPI - *Secretory Leukocyte Protease Inhibitor*, no nível de resíduos. Algo realmente bem limitado e (talvez) ultrapassado. Como STRING faz conexões com muitos outras bases de dados e serviços *web*, tentou-se encontrar outras formas de visualizar e analisar interfaces no nível resíduo/atômico, sem sucesso. Nada que chegasse próximo do que GAPIN faz nas interfaces cadeia-cadeia.

Mesmo não estando entre as bases de dados avaliadas pela revisão mencionada

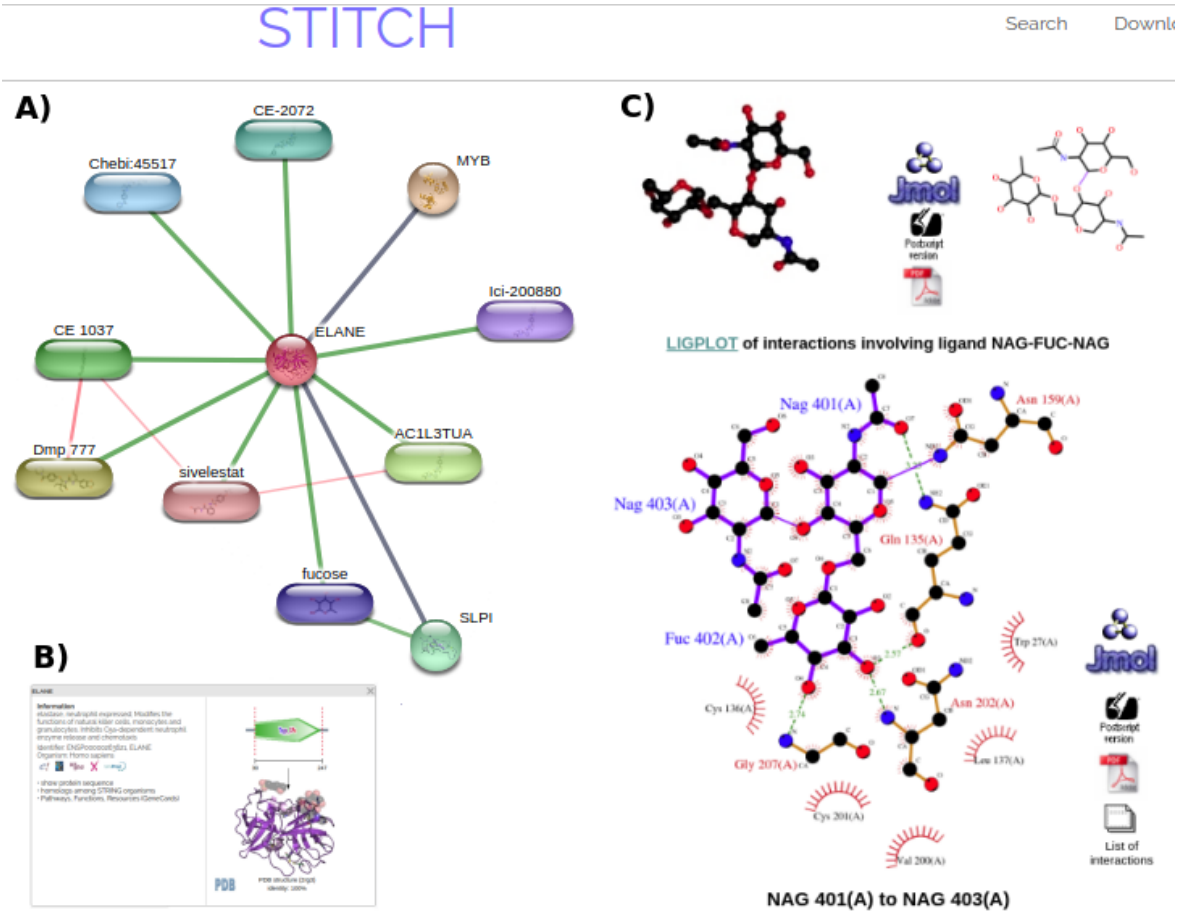


Figura 3.8: Tela capturada (e adaptada) do STRING. A) Resultado da consulta para o UniProtKB id P08246, que refere-se Elastase de Neutrófilo Humano (ELANE). B) Retorna basicamente o mesmo resultado da figura 3.6B. C) Ao clicar na imagem da estrutura, abre-se uma tela do PDBSum. Na aba *Ligands* vem uma imagem estática contendo as interações da interface da ELANE (cadeia E) com dois N-Acetyl-D-Glucosamina (NAG) e um Alfa-L-Fucose (FUC), ligados covalentemente entre si, e também com a enzima em E-ASN-159

.

acima, o *Interactoma3D* merece este parágrafo. É um *webservice* desenvolvido pelo *Institute for Research in Biomedicine*, ligado à Universidade de Barcelona, Espanha (Mosca et al. [2013]). Oferece visualizações de mais de 12 mil PPIs a partir de banco de dados de interactomas, mas (como STRING) com preocupação em dar destaque àquelas com estruturas resolvidas, sejam experimentalmente, depositadas no PDB; ou modeladas e depositadas em bases de modelos, como o Modbase (Pieper et al. [2014]). O usuário pode também inserir interações e estruturas não depositadas ainda. Apesar da abrangência ao nível de interactomas completos, evidenciando o contexto maior das cadeias proteicas em investigação, sua análise das interações estruturais resume-se a um visualização mais simples em Jmol, como exemplificado na Figura 3.7.

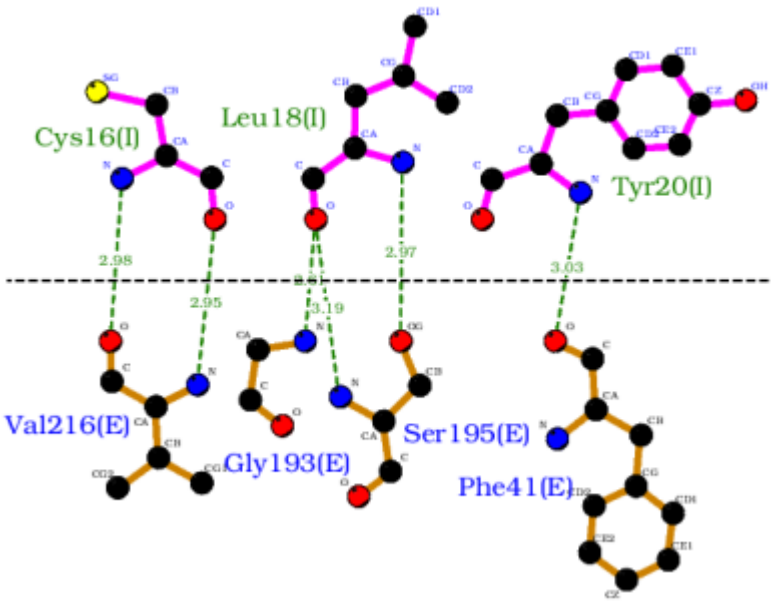


Figura 3.9: Tela capturada do LigPlot+ para interações intercadeia usando DIMPLOT para 1PPF. Foi usado $*E$ e $*I$ para *DOMAIN1* e *DOMAIN2* respectivamente, e resto default. Ativação de *Hydrophobic interactions* e na distâncias em *Runtime parameters* $->$ *Non-bonded contact parameters* não surtiram efeito na imagem.

Finalmente, há de se mencionar o STITCH - *Search Tool for Interacting Chemicals*, também suportado pelo mesmo consórcio do STRING envolvendo SIB, CPR-NNF e EMBL (Szklarczyk et al. [2016]). Seu foco está nas interações cadeia-ligantes. No site do STITCH é possível encontrar estatísticas tais como: 0.5 milhões de ligantes, 9.6 milhões de proteínas, 1.6 bilhões de interações, em 2031 organismos. O *front end* do STITCH parece fazer uso do mesmo *framework* do STRING. Ao fazer o mesmo teste de entrada com UniProtKB id P08246, referente à ELANE, retorna uma tela com mesmo layout do teste em STRING (figura 3.8). Clicando na imagem da estrutura, abre-se mais uma vez uma janela do PDBSum, só que dessa vez, o resultado que nos interessa está na aba *Ligands*. É exibido novamente uma imagem estática, não interativa, mas agora no formato LIGPLOT. Essa ferramenta foi criada em 1995 pelo grupo da Janet Thornton, hoje no EMBL-EBI, na Inglaterra (Wallace et al. [1995]), como um programa que gera automaticamente esquemas 2D de complexos proteína-ligantes a partir do PDB. Apesar de uma atualização para LigPlot+ em 2011[5] (Laskowski & Swindells [2011]), tais esquemas continuam pouco interativos, ainda que ele tenha opções de exportar para RASMOL ou PYMOL. O LigPlot+ vem com opção de montar esses esquemas 2D para interfaces cadeia-cadeia, através da aba DIMPLOT. Fez-se

[5] A versão corrente é a v.2.1, de 2015

um teste com a 1PPF, e ele conseguiu mapear apenas algumas ligações de hidrogênio, não assinalando nenhum contato hidrofóbico, mesmo com mudanças nos parâmetros de distâncias (figura 3.9).

Por fim, a tabela 3.10 mostra um comparativo entre as ferramentas analisadas nesse trabalho incluíndo o GAPIN.

Em futuros desenvolvimentos e ampliações do GAPIN, espera-se incorporar redes de interações intermoleculares no nível interactoma, nos moldes do STRING, Interactome3D e STITCH, mesmo quando não houver dados estruturais (mais detalhes, no capítulo VI - Conclusões e Perspectivas).

Nome	Utilização	Instalação	Principais Características
Cytoscape	Desktop, Java 8 sem suporte para versões mais recentes	Muito complexa e demorada. Exige que o usuário tenha bastante conhecimento sobre instalações em shell	Atua como módulos que podem ser incorporados como plugins por outras ferramentas e possivelmente disponibilizadas pela Cytoscape store.
Cytoscape JS & Javascript API	Javascript API	Não exige nenhuma instalação por ser uma API em Javascript	API opensource para visualização e análises de grafos
RINalyzer	Plugin baseado no Cytoscape	O mesmo que o Cytoscape com a adição do plugin do RINalyzer	opera em redes de contatos no nível de resíduo
NAPS	Aplicação WEB	Não se aplica	computa contatos tanto intra quanto intercadeias
PDBePISA	AppletsJava	Depende de Java e browser antigo	uma série de informações físico-químicas e termodinâmicas dos complexo
Protein Contacts Atlas	Aplicação WEB	Não se aplica	Permite avaliar contatos do tipo cadeia-ligantes
STRING	Aplicação WEB	Não se aplica	Sua base contém mais de 24.5 milhões de proteínas; 52.9 milhões de interações com score maior que 0.9; 3.0 bilhões de interações com score maior que 0.1; 5090 organismos, sendo 477 eucariotos
Interactoma3D	Aplicação WEB	Não se aplica	Oferece visualizações de mais de 12 mil PPIs a partir de banco de dados de interactomas
STITCH	Aplicação WEB	Não se aplica	Foco nas interações cadeia-ligante
LigPlot+	Aplicação WEB	Não se aplica	Interações intercadeia
GAPIN	Aplicação WEB	Não se aplica	Alinhamentos manuais e automáticos, definição de spots, visualização em 2D no formato de redes de contatos e 3D no formato da estrutura, aceita importação de arquivos PDB customizados

Figura 3.10: Tabela comparativa das ferramentas analisadas. Por essa tabela é possível analisar a complexidade de instalação das ferramentas, sua utilização e principais características

Capítulo 4

Materiais e Métodos

4.1 Cálculo da área de contato

O método clássico usado para cálculo de superfícies biomoleculares disponíveis para interações é o ASA ou SASA - *Accessible Surface Area ou Solvent-Accessible Surface Area* (Lee & Richards [1971]), que computa a área acessível de cada átomo à uma molécula de água. Apesar de ser amplamente usado na literatura e dispor de muitos métodos diferentes para o seu cálculo (Ali et al. [2014]), nos complexos cadeia-cadeia ou cadeia-ligante ASA computa áreas acessíveis e áreas das interfaces sem definir que resíduo ou átomo estabeleceu contato com quais outros.

Adotou-se aqui um método de cálculo de área específico para trabalhar com interfaces intermoleculares que tem a vantagem de mapear também contatos. Foi usado pela primeira vez na tese de doutorado de uma egressa do Programa Interunidades de Pós-Graduação em Bioinformática da UFMG (Alves [2015]), e aprimorado nesta tese. Foi nomeado como método BARS, em alusão aos autores envolvidos na sua criação e desenvolvimento (Biharck, Alves, Romanelli, Silveira).

BARS calcula uma área diferente do ASA, mas mantém ainda boa correlação com ela (vide Capítulo V - Resultados e Discussões). Enquanto ASA computa a área exposta ao solvente, BARS calcula a área **não** exposta ao solvente, dado dois átomos isolados. Ou seja, seria uma área decorrente exclusivamente da proximidade de dois átomos (área de contato) e não acessível ao solvente (figura (4.1).

Importante frisar que BARS opera heuristicamente, sempre isolando dois átomos, mesmo que eles estejam empacotados contra uma miríade de átomos como é o caso de proteínas e outras biomoléculas. Nesse sentido, trata-se de uma área de contato aproximada, pois não leva em consideração a influência de outros átomos na vizinhança imediata. Mesmo com essa simplificação, conforme dito anteriormente, ainda guarda

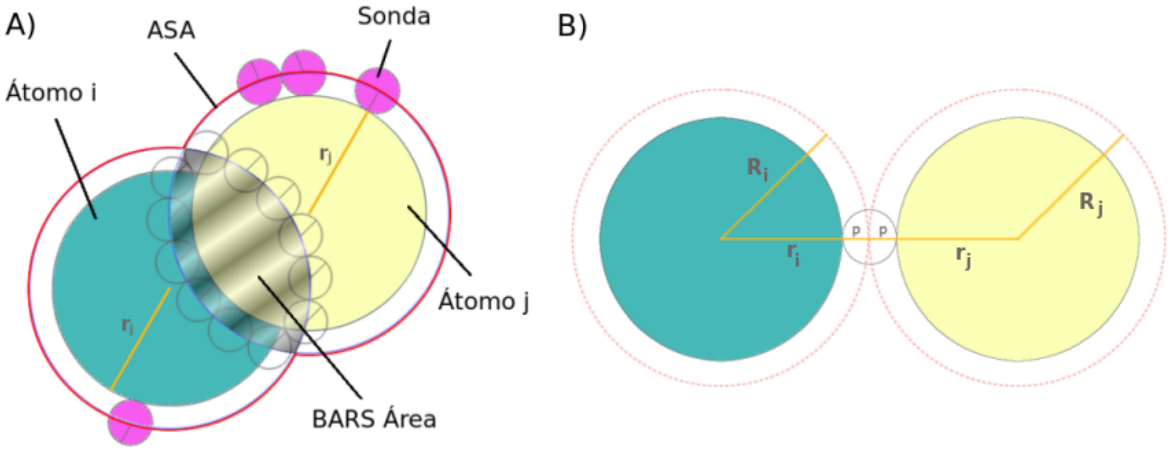


Figura 4.1: Área de contato tal qual definida pela metodologia BARS. A) Dada uma sonda de raio p, e raios de van der Waals r_i e r_j, a área de contato $Ac_{i,j}$ entre eles é aquela que a sonda não toca. B) Se a distância for maior que os raios de van der Waals mais o diâmetro da sonda ($d_{i,j} >= r_i + r_j + 2p$), $Ac_{i,j}$ será zero, configurando a possibilidade de uma cavidade entre átomos i e j. (figura adaptada de (Alves [2015]), com permissão.)

correlação com ASA. Um dos objetivos é simplificar o cálculo, ganhando significativo tempo computacional. Isso porque trabalhando apenas com dois átomos é possível calcular a área de contato analiticamente, em ordem de um ($O(1)$), através da equação[1] (4.1).

$$A_c(R_i, R_j, d_{i,j}) = 2\pi({R_i}^2 + {R_j}^2) - \pi(R_i + R_j)d_{i,j}\left[1 + \left(\frac{R_i - R_j}{d_{i,j}}\right)^2\right] \tag{4.1}$$

Sendo:

$\mathrm{R}_i = \mathrm{r}_i + \mathrm{p}$
$\mathrm{R}_j = \mathrm{r}_j + \mathrm{p}$
r_i = raio de van der Waals do átomo i
r_j = raio de van der Waals do átomo j
p = raio da sonda (molécula de água)
$\mathrm{d}_{i,j}$ = distância entre o $i°$ e o $j°$ átomo

Além do ganho no tempo computacional, há ainda outros efeitos colaterais bons. Olhando a equação (4.1) e a figura (4.1) percebe-se que se a distância entre os átomos i e j for maior que os respectivos raios de van der Waals e o diâmetro da sonda, a

[1] A dedução dessa equação encontra-se em (Alves [2015]).

área de contato é zerada. Tal como em ASA, a sonda usada aqui foi uma molécula de água simplificada, assumindo-a esférica e isotrópica, com raio de van der Waals aproximado para o raio do oxigênio (1.4Å). Logo, a área de contato será zero sempre que for possível intervir uma ou mais moléculas de água entre dois átomos, configurando a possibilidade de uma cavidade em termos da metodologia BARS. E se há condição para uma cavidade, assume-se a improbabilidade de contato direto entre esses átomos[2].

E isso pode ter sentido termodinâmico, principalmente para os contatos hidrofóbicos intermoleculares. Se a distância entre dois átomos na interação entre duas biomoléculas diferentes é tal que não permite mais a interveniência de uma molécula de água, podemos pressupor que houve exclusão de água nesse ponto da interface. Ou seja, um processo de dessovatação. Tal processo é parte fundamental da energia livre de ligação entre duas biomoléculas. A liberação de moléculas de águas em regiões hidrofóbicas para o solvente contribui para o aumento da entropia geral do sistema, e isso é um dos fatores que orienta a espontaneidade do fenômeno de aproximação e ancoragem entre duas biomoléculas (Chandler [2005]).

4.2 Definição das interfaces moleculares

A metodologia BARS de cálculo de área de contato tem outro efeito colateral bom: ela permite definir e encontrar os átomos que estão participando da região de interface entre biomoléculas. A definição é simples: um átomo i de a uma biomolécula M será considerado da interface se existe uma área de contato não nula ($Ac \neq 0$) com qualquer outro átomo não pertencente à biomolécula M. Na verdade, pode-se estabelecer um parâmetro Ac_{min} e considerar apenas contatos em que $Ac \geq Ac_{min}$. Para esta tese, foi usado um $Ac_{min} = 5\text{Å}^2$. Esse número foi estipulado de forma empírica. Fará parte dos trabalhos futuros desta tese encontrar uma racionalização para a escolha do Ac_{min}.

Para o mapeamento dos átomos das interfaces, são computados primeiro as distâncias Euclideanas de todos os átomos contra todos átomos (de cadeias diferentes). De posse deste parâmetro, são calculadas as Ac, preservando apenas as com $Ac \geq Ac_{min}$.

Um exemplo de interface cadeia-cadeia e cadeia-ligante com todos os tipos de átomos para 1PPF pode ser visto na figura (4.2).

[2]O contato indireto, pela intermediação de um ou mais átomos entre eles ainda continua existindo, mas o foco aqui está no contato direto.

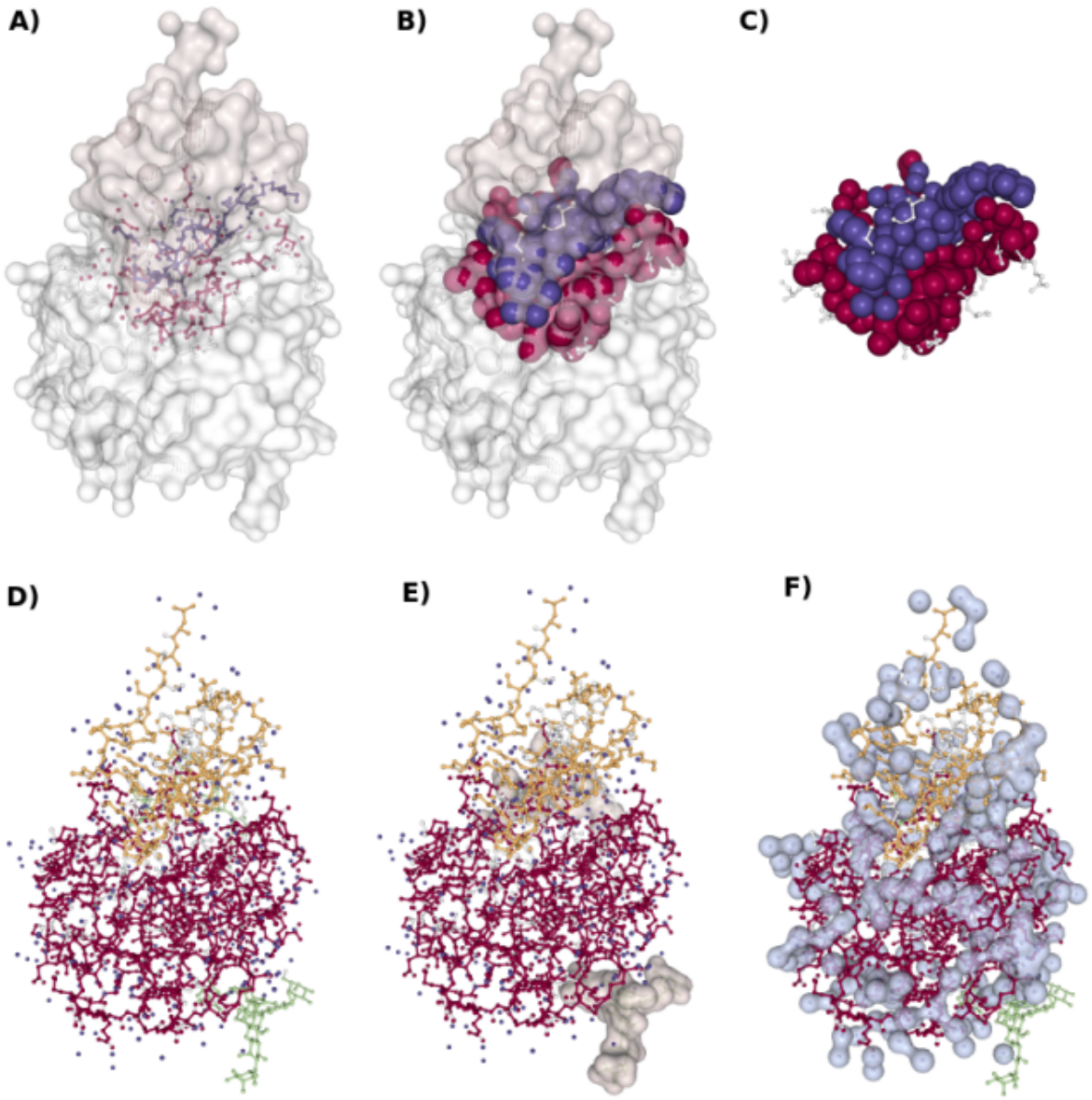


Figura 4.2: Interfaces intermoleculares em 1PPF, a partir de contatos em que $Ac \geq Ac_{min}$. A) Elastase e seu Inibidor Ovomucoide destacados em *surface* tipo Connolly, com os elementos da interface cadeia-cadeia em *ball+stick* coloridos conforme a cadeia. B) Mesmo anterior, mas átomos em *spacefill*. C) Foco na interface cadeia-cadeia, sem *surfaces*. D) Foco agora nas interfaces cadeias-ligantes. Os ligantes são oligossacarídeos (em verde) e águas (em azul). Como são muitas, uma boa extensão da superfície do complexo enzima-inibidor produziu $Ac \geq Ac_{min}$. As demais cores são das cadeias enzima e inibidor. E) Destaque em *surface* para os oligossacarídeos. F) Destaque em *surface* para as águas.

4.3 Grafos das interfaces moleculares

BARS vai além de só encontrar as interfaces intermoleculares, listando resíduos ou átomos com $Ac \geq Ac_{min}$. Ele permite mapear que átomo ou resíduo faz contato com que outro dentro da interface. Ou seja, pode-se construir uma rede de contatos, seja entre cadeia-cadeia ou cadeia-ligantes, tendo as áreas Ac como peso das arestas. Um átomo i terá uma aresta com peso Ac com outro átomo j se e somente se $Ac \geq Ac_{min}$. Nesse sentido, esses pesos dirão o quão fortes serão os contatos, pois quanto mais bem empacotados estiverem os átomos das interfaces, mais densa ela será, mais próximos estarão os átomos entre si, e maiores serão suas áreas de contatos.

Para representar a rede de contatos nas interfaces usou-se o conceito de grafos. Matematicamente, um grafo G é definido por uma função $G = (V, E)$, onde V é um conjunto finito e não vazio de nós ou vértices e E é um conjunto de ligações ou arestas (*links*) que ligam os elementos de V. Nesta tese, serão usados ainda os seguintes conceitos envolvendo grafos:

- Grafos não dirigidos: se o conjunto E forma pares não-ordenados; ou seja, se existe uma aresta ij ligando o vértice i ao vértice j, não importando a ordem, sendo pois $ij = ji$.

- Grafos com peso: se a aresta ij carrega um valor $w(i, j)$ que pode ser diferente de um ou zero.

- Grafos bipartidos: se o conjunto contendo os vértices V puder ser separado em dois subconjuntos disjuntos de vértices U e T tal que só existem arestas ligando vértices de U com T, não existindo vértices de U com U, nem de T com T.

- Grafos multipartidos ou k-partidos: uma generalização do grafo bipartido, quando o conjunto contendo os vértices V puder ser separado em k subconjuntos disjuntos de vértices $U_1, U_2, ..., U_k$.

- Matriz de adjacências: trata-se de uma matriz $\mathbf{M}$ que representa um grafo $G = (V, E)$, com n vértices, quadrada $n\ x\ n$ e simétrica, onde:

$$m_{ij} = \begin{cases} w(i,j), & \text{se } \{i,j\} \in E \\ 0, & \text{caso contrario.} \end{cases}$$

  sendo que, $m(i, j) = 0$ indica ausência de aresta entre vértices i e j; e $w(i, j)$ representa o peso entre vértices i e j.

- Grau dos vértices: dada uma matriz de adjacências **M** com pesos wij, o grau de um vértice i é definido como somatório dos pesos de suas arestas:

$$d_i = \sum_{j=1}^{n} w_{i,j} \tag{4.2}$$

  ou seja, o somatório das linhas em **M**.

- matriz de graus: é uma matriz quadrada diagonal **D** com o d_i em sua diagonal, ou:

$$\begin{bmatrix} d_1, & \dots & \dots & 0 \\ 0, & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & d_n \end{bmatrix}$$

- Corte de grafos (*cut*): envolve produzir um corte de arestas num grafo G de modo a gerar duas partições (subgrafos), com vértices agrupados em dois conjuntos disjuntos A e $\overline{A}$ (não A). O somatório dos pesos das arestas cortadas define o tamanho do corte, dado pela equação (4.3):

$$W(A, \overline{A}) = \sum_{i \in A, j \in \overline{A}} w_{ij} \tag{4.3}$$

  sendo a generalização para k-partições dada por:

$$cut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i}) \tag{4.4}$$

- Volume de um conjunto de vértices: definido como a soma dos pesos de todas as arestas ligadas a um conjunto de vértices A:

$$vol(A) = \sum_{i \in A, j \in A} w_{ij} \tag{4.5}$$

Conforme visto na Introdução, GAPIN faz uso de dois tipos de grafos: grafos de baixo nível ou de granularidade fina; e grafos de alto nível ou de granularidade grossa.

4.3.1 Grafos de baixo nível

São grafos construídos a partir de um PDBid em nível atômico: o conjunto V é composto por átomos e o conjunto E por arestas cujos pesos $w(i, j) = Ac_{ij}$, na condição

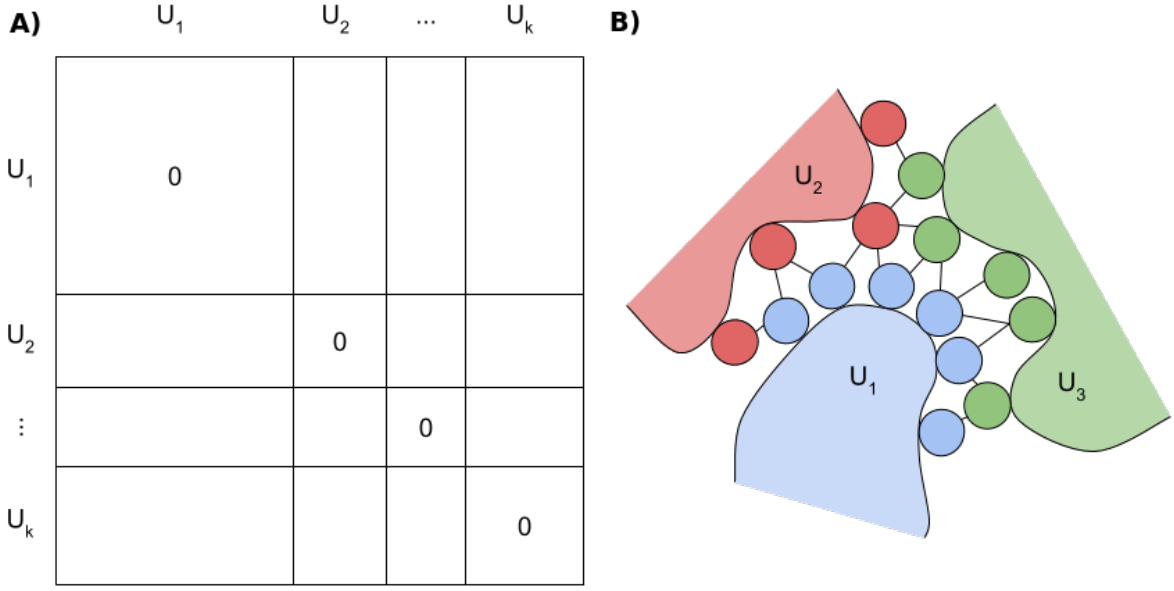


Figura 4.3: A) Representação de uma matriz de adjacências k-partida **M** para k cadeias. U_i representa o conjunto de átomos de cada cadeia i. Não há arestas ($m_{i,j} = 0$) entre átomos intracadeias (entre U_i e U_i). B) Exemplo de um rede envolvendo 3 cadeias U_1, U_2 e U_3, coloridas por cores diferentes, com átomos simbolizados por círculos (nós) e arestas indicando um Ac diferente de zero. Nota-se a presença de contatos ternários ao centro, envolvendo átomos de 3 cadeias diferentes, interligados entre si.

de que $Ac_{ij} \geq Ac_{min}$. O grafo é k-partido, no sentido de só computar Ac entre átomos com rôtulos (*labels*) de cadeias diferentes no PDBid, onde k o número de cadeias.

Internamente, um grafo k-partido advém de uma matriz de adjacências **M** também k-partida, como na figura (4.3A). As células $m_{i,j}$ de **M** são zeradas para todos os átomos i e j intracadeias. Tal construção mantém a matriz quadrada e simétrica, o que se revela útil para os algoritmos de agrupamento em grafo que são utilizados, conforme será visto mais adiante. Note também que a matriz **M** permite contatos não binários entre átomos de diferentes cadeias; ou seja, um átomo da cadeia U_1 pode fazer contatos ao mesmo tempo com átomos de outras cadeias $U_2, U_3, ..., U_k$ (4.3B). Tais mapeamentos não binários podem constituir, sem dúvida, em mais uma inovação do GAPIN.

4.3.2 Grafos de alto nível

São grafos construídos a partir do agrupamento em grafos de baixo nível. O objetivo do agrupamento é inferir não só a existência de comunidades de átomos densamente agrupados ao longo das interfaces intermoleculares, mas também como essas comunidades se estruturam e se intercomunicam (Fortunato [2010]). Isso inclui tam-

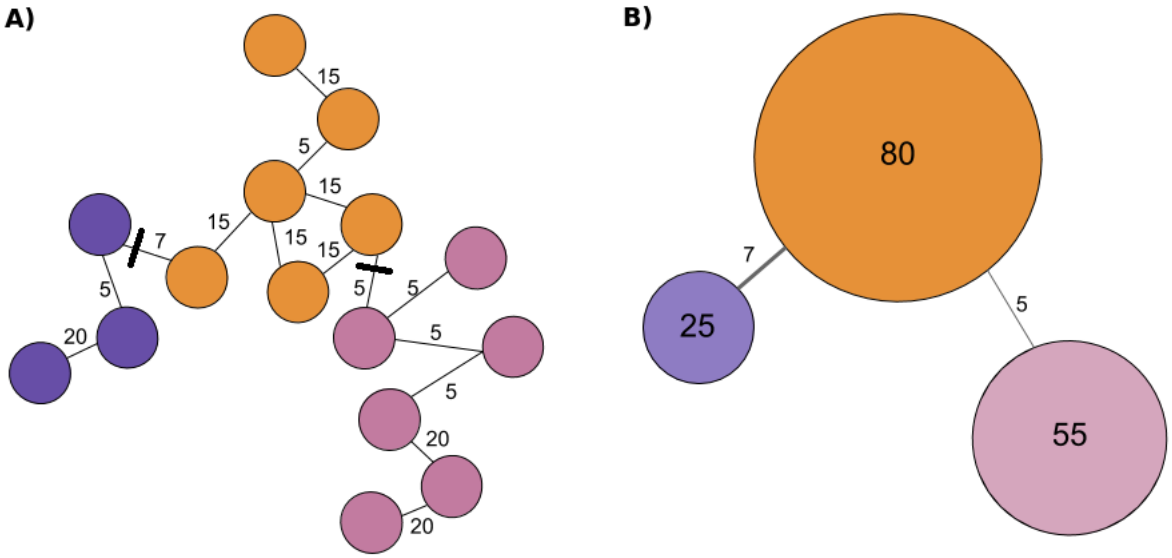


Figura 4.4: Exemplo de grafo de alto nível construído a partir de um grafo de baixo nível. A) Mesmo grafo de baixo nível da figura (4.3B), pondo em evidência 3 grupos pelas 3 cores diferentes. Os traços pretos indicam onde houve o corte (*cut*) das arestas de modo a produzir 3 partições no grafo. Os números próximos às arestas indicam as respectivas áreas de contato (Ac). B) O grafo de alto nível derivado do grafo de baixo nível. Os rótulos nos nós indicam o somatório das áreas de contatos (pesos) da arestas internas aos grupos. Os rótulos nas arestas indicam o volume do corte, ou o somatório das arestas cortadas para produzir as partições. Graficamente, tanto nós quanto arestas são proporcionais aos respectivos valores dos rótulos.

bém a varredura dos agrupamentos (*cluster scanning*), um estudo de como os grafos de alto nível se reorganizam frente diferentes números de partição. Essa varredura foi inspirada em (da Silveira et al. [2009b]). Tais análises resultaram em algumas descobertas inesperadas, referentes à associação entre certas comunidades de átomos com regiões *Hot Spots*, candidatas a compor alvos terapêuticos para novos fármacos (vide Capítulo V - Resultados e Discussões).

Um exemplo hipotético de grafo de alto nível pode ser visto na figura (4.4B). Ela mostra como esse grafo é construído a partir de um grafo de baixo nível (figura 4.4A). Dado um número de grupos requerido, o algoritmo de agrupamento em grafo procura particionar o grafo de forma fazer o melhor corte (*cut*) possível de arestas, quebrando o grafo em subgrafos não conexos, de modo a gerar grupos densos em arestas conforme seus pesos (vide detalhes do algoritmo mais à frente). Grafos de alto nível têm rótulos também nos nós, e eles indicam o somatório dos pesos (áreas de contato) das arestas internas a cada grupo gerado. Logo, os nós neste tipo de grafo têm tamanho. Os rótulos nas arestas simbolizam o volume do corte, ou o somatório das arestas cortadas para gerar as partições. Tais arestas são aquelas que interligam os grupos. Em grafos de alto

nível, tanto nós quanto arestas são graficamente representados no GAPIN proporcionais aos respectivos tamanhos.

4.3.3 Grafos em interfaces cadeia-cadeia e cadeia-ligante

GAPIN oferece dois tipos de interfaces intermoleculares para visualização e análise: cadeia-cadeia e cadeia-ligante. Para ambos são gerados grafos de baixo nível e alto nível. As matrizes de adjacências para grafos de baixo nível em interfaces cadeia-cadeia seguem o exemplo dado na figura (4.3A). As interfaces cadeia-ligante passam pelo mesmo processo, com alguns detalhes importantes. Nos arquivos PDBs, eventuais ligantes (o que inclui água e íons) sempre pertencem a uma cadeia, geralmente as que eles se encontram mais próximos. GAPIN modifica essa associação quando se trata de cadeia-ligante, fazendo com que cada ligante pertença a uma cadeia diferente própria na hora de montar a matriz de adjacências. Isso vale também para cada molécula de água, que são consideradas como se cada uma pertencesse a uma cadeia diferente. Sendo assim, águas podem fazer contatos entre elas mesmas, quando se trata de grafos cadeia-ligante. Em grafos cadeia-cadeia isso não acontece, porque segue o padrão default do PDB de associar os ligantes e águas às cadeias mais próximas.

4.4 Agrupamento em grafos

No processo de geração dos grafos de alto nível, GAPIN utiliza um algoritmo de agrupamento espectral normalizado bastante semelhante aos descritos em (Von Luxburg [2007]), mas fazendo uso de matrizes Laplacianas assimétricas ($\boldsymbol{L_{rw}}$) construídas a partir de matrizes de adjacências de grafos de baixo nível. A $\boldsymbol{L_{rw}}$ tem a seguinte definição:

$$\begin{aligned} \boldsymbol{L} &= \boldsymbol{D} - \boldsymbol{W} \\ \boldsymbol{L_{rw}} &= \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W} \end{aligned} \tag{4.6}$$

onde **D** é uma matriz de graus; **W** é matriz de adjacências com pesos wij; **L**, uma matriz Laplaciana; **I**, matriz identidade.

Pelo teorema de Rayleigh-Ritz citado em (Von Luxburg [2007]), com a decomposição espectral de $\boldsymbol{L_{rw}}$, pode-se usar os autovetores com os menores autovalores para gerar uma aproximação de um corte mínimo especial chamado $Ncut$, em que os cortes são normalizados pelo volume dos grupos, ou:

$$NCut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{vol(A_i)} \tag{4.7}$$

Então, minimizar o *Ncut* é equivalente a fazer partições num grafo G que tenha máximo volume de arestas dentro dos grupos e mínimo fora. Cabe frisar que o problema de encontrar o mínimo *Ncut* global é NP-difícil (*NP-hard*) (Wagner & Wagner [1993]). O uso dos espectros de L_{rw} para minimizar *Ncut* é uma aproximação (para um aprofundamento matemático, vide (Von Luxburg [2007])).

Logo, pode-se resumir o algoritmo utilizado como abaixo:

Algoritmo 1: Agrupamento espectral

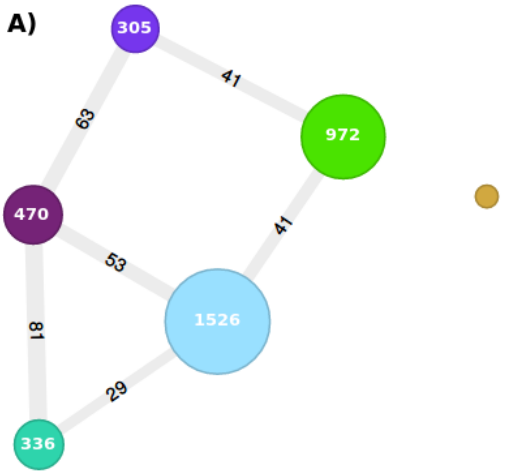
Entrada: Matriz de pesos $\boldsymbol{W} \in \mathbb{R}^{nxn}$, número k de clusteres
1: Calcule a Laplaciana normalizada $\boldsymbol{L_{rw}}$ de $\boldsymbol{W}$
2: Calcule os primeiros k autovetores $u_1, \ldots, u_k$ de $\boldsymbol{L_{rw}}$
3: Faça $\boldsymbol{U} \in \mathbb{R}^{nxk}$ ser a matriz contendo os autovetores $u_1, \ldots, u_k$
4: Rode o algoritmo de agrupamento k-medoid sobre $\boldsymbol{U}$
Saída: Grupos $A_1, \ldots, A_k$.

Percebam que no último passo do **Algoritmo 1** descrito acima usa-se um algoritmo de agrupamento por partição clássico chamado *k-medoid*, encontrado no pacote PAM - *Partitions Arround Medoids* do R (Reynolds et al. [2006]). Luxburg defende que a matrix $\boldsymbol{U}$ com os k primeiros autovetores da decomposição de $\boldsymbol{L_{rw}}$ compreende uma novo espaço de representação dos pontos originais ($\boldsymbol{W}$). Neste novo espaço, os pontos seriam mais facilmente agrupados, sendo possível usar um algoritmo de agrupamento mais simples e rápido, como *k-means* ou *k-medoid*. No GAPIN foi utilizado o último, considerado mais eficiente que o primeiro (Reynolds et al. [2006]).

O algoritmo *k-medoid* é bem simples, e pode ser descrito em alto nível como:

Algoritmo 2: *k-medoids* **(PAM)**

Entrada: Matriz $\boldsymbol{U}$
1: selecione k pontos para serem os *medoids* iniciais
2: repetir:
3: associe todos os pontos aos *k-medoids* mais próximos
4: selecione um ponto p que não seja *medoid*
5: calcule o custo C de trocar o *medoid* m_i por p
6: se $C < 0$ então troque *medoid* m_i por p
7: até não haver mais mudanças nos *medoids*
Saída: Grupos $A_1, \ldots, A_k$.



B)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	$\boldsymbol{Q}\%$
[1,]	305	41	63	0	0	0	**75**
[2,]	41	972	0	0	0	41	**92**
[3,]	63	0	470	81	0	53	**70**
[4,]	0	0	81	336	0	29	**75**
[5,]	0	0	0	0	29	0	**100**
[6,]	0	41	53	29	0	1526	**93**

Figura 4.5: A) Figura de um grafo de alto nível de interfaces cadeia-cadeia apolares com 6 grupos da 1PPF. O rótulo do pequeno nó em bege mais à direita é de 29Å^2. B) Respectiva tabela com a matriz de adjacências. Valores indicam áreas de contato em Å^2, exceção à coluna $\boldsymbol{Q}\%$ que representa a qualidade do nó, em porcentagem.

4.5 Qualidade dos agrupamentos

A qualidade do agrupamento espectral pode ser avaliada seguindo a intuição da minimização de $Ncut$: um bom agrupamento é aquele que maximiza os contatos internamente ao grupo e minimiza externamente. Olhando para uma matriz de adjacência (figura 4.5B) de um grafo de alto nível ((figura 4.5A) é fácil ver que bons grupos são aqueles que concentram os valores nas diagonais.

Assim, dado um grafo de alto nível, uma métrica de qualidade simples seria pegar a razão entre volume das arestas sobre o volume do nó. No Exemplo da figura (4.5B), a qualidade para o agrupamento no nó 1 seria: $305/(41 + 63 + 305) = 0.75$. Ou seja, 75% do volume total envolvendo o nó 1 vem de arestas internas ao nó.

Logo, com essa definição, um agrupamento perfeito com qualidade máxima produziria uma matriz de adjacência diagonal n x n, com n componentes conexos. Curiosamente, essa matriz de adjacência de um grafo de alto nível pode ser vista como uma matriz de confusão (Zaki et al. [2014]), usada para medir o desempenho de algoritmos em algumas técnicas de aprendizado de máquina. Um classificador perfeito também geraria uma tabela de contingência diagonal com todos os valores verdadeiros-positivos ou verdadeiros-negativos nessa diagonal (e nulo para todos os valores restantes). Nesse sentido, a maioria das métricas derivadas da matriz de confusão pode ser válida para as matrizes de adjacências de grafos de alto nível, como precisão, cobertura/revocação, acurácia, F1 *score* etc, com a diferença de que a matriz de adjacência é simétrica. Essa simetria faz com que os valores falso-negativos fiquem iguais aos falsos-positivos, uni-

ficando várias das métricas da matriz de confusão, como cobertura e precisão. Nossa métrica de qualidade, por exemplo, pode ser calculada como uma medida de sensibilidade. Entre os trabalhos futuros também estará um estudo mais aprofundado dessa unificação entre matriz de adjacências em grafos de alto nível e a matriz de confusão.

4.6 Varredura de agrupamentos

Em algoritmos de agrupamento por partições (a exemplo do *k-medoids*), como o número de grupos (k) deve ser especificado *a priori*, a abordagem usual procura um k que maximize a qualidade geral do agrupamento, utilizando métricas tais como coeficiente de silhueta (Rousseeuw [1989]) ou modularidade (Newman [2004]). Nesta tese, seguiu-se um método de varredura (da Silveira et al. [2009b]) de agrupamento, mostrando o perfil das partições no grafo para k crescente, deixando para o usuário a conclusão de qual o melhor número de grupos. Ademais, tal varredura pode trazer informações valiosas sobre a organização estrutural das interfaces intermoleculares, uma vez que o algoritmo de agrupamento está tentando minimizar os $Ncuts$ para cada k. Também permite que o usuário infira qual a propensão de um nó a ser particionado no próximo k.

Na Figura 4.6, é possível ver uma sequencia de k-particionamentos (de 3 a 6) para o grafo cadeia-cadeia somente com interações apolares da nossa conhecida 1PPF. Nós coloridos em marrom-vermelho foram mais resistentes aos particionamentos do que amarelos-brancos. O incremento de k pára quando a qualidade mínima do agrupamento chega a 50% ou quando k atinge 20 partições. Esse limite de 20 foi determinado por questões de performance.

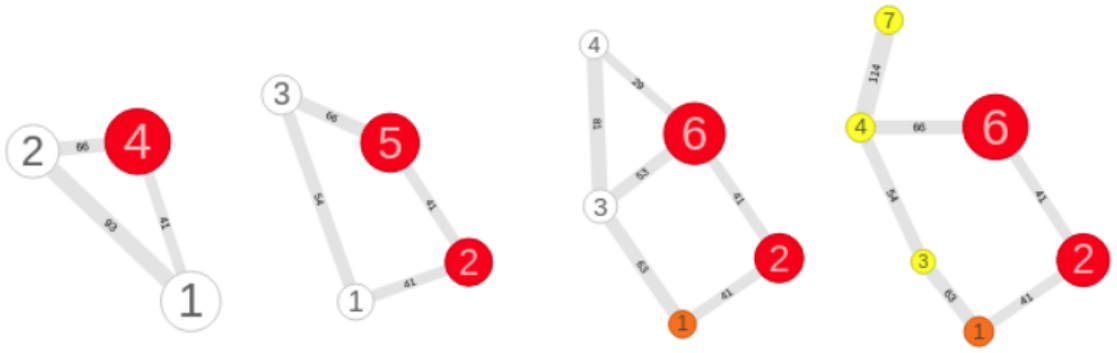


Figura 4.6: Varredura de agrupamentos para interfaces cadeia-cadeia com interações apolares em 1PPF, variando k de 3 a 6. Cores em marrom-vermelho indicam nós mais preservados que em amarelo-branco.

Comparação $3 \Rightarrow 4$*:*

	$V3$	$V4$	$Dist$	$A3$	$A4$	$A4/A3$	$\boldsymbol{Pre_{index}}$
$1-1$	1	1	0.0	1242	1242	1.00	**100**
$2-2$	2	2	2.2	2725	2095	0.00	**0**
$3-3$	3	3	0.6	571	571	1.00	**100**

Comparação $4 \Rightarrow 5$*:*

	$V4$	$V5$	$Dist$	$A4$	$A5$	$A5/A4$	$\boldsymbol{Pre_{index}}$
$1-1$	1	1	0.0	1242	1242	1.00	**100**
$2-2$	2	2	0.7	2095	2026	0.97	**48**
$3-3$	3	3	0.0	571	571	1.00	**100**
$4-4$	4	4	0.0	450	450	1.00	**100**

Tabela 4.1: Tabela exemplificando o cálculo do índice de preservação (Pre_{index}). Na comparação $3 \Rightarrow 4$, sobrepõe-se o grafo de 3 partições contra o de 4. Vê-se pelas áreas $A3$ e $A4$ que as sobreposições de nós $1-1$ e $3-3$ atendem ao critério de preservação ($Dist < 1.4$ e $r > 0.85$). Mas, isso não acontece para $2-2$. Na Comparação $4 \Rightarrow 5$, todos atendem ao critério de preservação. Mas, a sobreposição $2-2$ advém de uma sobreposição não preservada na comparação anterior. Logo, é feita a média: $(0.00 + 0.97)/2 = 0.48$.

4.7 Índice de preservação do nó

Ao longo deste trabalho foi criado um índice de preservação que mede quanto um nó resiste a ser reparticionado. Dada uma sequencia de k-partições sucessivas, para cada nó correspondente entre as iterações i e $i+1$, o índice obtém a média das razões ($r = \frac{A_{i+1}}{A_i}$) se essa relação for maior que um determinado limite (definido como 0.85). Nós correspondentes são aqueles cujos centróides estão a uma distância menor que 1.4 Å. Se a relação for menor que o limite ou a distância for maior que 1.4[3], a proporção será redefinida para zero. A Tabela 4.1 mostra como este índice funciona com 2 interações.

4.8 Tipos de interações

Na versão atual, o GAPIN trabalha apenas com dois tipos de interações atômicas: polar e não polar. Os átomos são classificados de acordo com a Tabela 4.2, adaptada de Alves [2015] e Sobolev et al. [1999].

Cabem alguns comentários. A abordagem histórica para com as categorizações de tipos de interações atômicas em estruturas resolvidas foi binária, classificando todo e

[3]Fará parte de estudos futuros uma investigação sobre a influência destes parâmetros sobre o índice de preservação

Átomo	Polaridade
Carbonos backbone e carbonos alfa	Polar
Carbonos de proteínas em uma ligação covalente a átomos polares (exceto para HIS e TRP)	Polar
Outros carbonos de proteínas	Apolar
CYS.S em SS-bond, MET.S	Apolar
O ou N	Polar
Qualquer carbono não proteína	Apolar
Quaisquer outros átomos	Polar

Tabela 4.2: Classificação apolar x polar de interações atômicas utilizadas no GAPIN.

qualquer átomo de carbono e enxofre como apolar e os demais polares (Lee & Richards [1971]). Sobolev et. al. (1999) talvez tenha sido um dos primeiros a ganhar projeção por uma classificação não-binária mais elaborada (Sobolev et al. [1999]), que se dividia em 8 classes entre: I - hidrofílicos, II - aceptores de hidrogênio, III - doadores de hidrogênio, IV - hidrofóbicos, V - aromáticos, VI - neutros, VII - neutros-doadores e VIII - neutros aceptores. A maioria dessas classes (IV a VIII) tentava diferenciar carbonos conforme o contexto e vizinhança em termos de ligações covalentes.

Nesta tese, optou-se por uma classificação binária (apolar x polar), mas com alguma preocupação em contextualizar alguns carbonos nas interfaces cadeia-cadeia, principalmente aqueles que tinham, como vizinhos covalentes, átomos polares. Desse modo, todos os carbonos do *backbone* foram considerados polares, dado que tanto o carbono C da carbonila quanto o carbono alfa CA estão ligados covalentemente a vizinhos polarizados[4]. Conclui-se ainda que todo o *backbone* em si foi considerado polar. A única exceção em que carbonos ligados a átomos polares foram considerados apolares diz respeito aos carbonos das cadeias laterais da histidina (HIS) e triptofano (TRP), dada a tendência verificada empiricamente de que tais resíduos se acomodam bem em núcleos hidrofóbicos (White & Wimley [1999]).

Nas interfaces cadeia-ligante preferiu-se algo próximo de uma categorização binária clássica, com todos carbonos como apolares, e demais átomos polares.

4.9 Alinhamento de grafos de alto nível

Um dos importantes recursos que acompanha o GAPIN é a possibilidade de alinhar dois grafos de alto nível. Isso pode ser feito manualmente ou automaticamente. Para o primeiro caso, o usuário carrega dois (ou mais) PDBs e opera o alinhamento

[4]A ligação peptídica entre o nitrogênio da amida (ao qual o CA está ligando) e o oxigênio da carbonila (ao qual o C está ligado) compõe um dipolo elétrico.

com o *mouse*, usando o botão de sincronização entre estruturas PDB renderizadas e os grafos de alto nível. Para o segundo caso, após carregar um PDB inicial, o usuário deve selecionar a opção de alinhamento no menu e escolher outro PDB para alinhar.

Algoritmo 3: *Topos − alinhamento de grafos*

Entrada: matrizes de adjacências $\boldsymbol{A_1}$,$\boldsymbol{A_2}$; matrizes de centroides $\boldsymbol{C_1}$,$\boldsymbol{C_2}$

Function ALIGNMENT($\boldsymbol{A_1}$,$\boldsymbol{A_2}$,$\boldsymbol{C_1}$,$\boldsymbol{C_2}$)
    n_1 = número de nós em $\boldsymbol{C_1}$
    n_2 = número de nós em $\boldsymbol{C_2}$
    k = número de nós a ser considerado (subgrafo)
    n = número de vetores singulares direitos
    m = tamanho da lista *best*
    $ListC_1$ = combinações de nós(n_1,k)
    $ListC_2$ = combinações de nós(n_2,k)
    $Best$ = lista vazia de tamanho m
    Para cada i em $ListC_1$ faça:
        $\boldsymbol{V_1}$ = n primeiros vetores singulares direitos de SVD($\boldsymbol{C_1}$,$ListC_1[i]$)
        Para cada j em $ListC_2$ faça:
            $\boldsymbol{V_2}$ = n primeiros vetores singulares direitos de SVD($\boldsymbol{C_2}$,$ListC_2[i]$)
            $\boldsymbol{Rot}$ = alinhe os vetores ($\boldsymbol{V_1}$,$\boldsymbol{V_2}$) por uma matriz de cossenos diretores
            $Score$ = EVALUATE($\boldsymbol{A_1}$,$\boldsymbol{A_2}$,$\boldsymbol{C_1}$,tranformation($\boldsymbol{C_2}$,$\boldsymbol{Rot}$))
            $Best$ = salve os melhores *scores* ($Best$,$Score$,$\boldsymbol{Rot}$)
    return($Best$)

Function EVALUATE($\boldsymbol{A_1}$,$\boldsymbol{A_2}$,$\boldsymbol{C_1}$,$\boldsymbol{C_2}$)
    ListD = lista de nós superimpostos ($\boldsymbol{C_1}$,$\boldsymbol{C_2}$)
    ListA = lista com a razão dos rótulos dos nós superimpostos ($\boldsymbol{A_1}$,$\boldsymbol{A_2}$,ListD)
    ListE= lista com o cosseno das arestas superimpostas ($\boldsymbol{C_1}$,$\boldsymbol{C_2}$,ListD)
    return(soma(ListA)+soma(ListE))

Saída: lista com os n melhores *scores* e matrizes de alinhamentos $\boldsymbol{Rot}$.

Nesta tese, foi criado um algoritmo automático de alinhamento de grafos, chamado inicialmente de *Topos*, em que se leva em consideração tanto a topologia do grafo quanto as posições dos centroides dos nós, como descrito pelo pseudocódigo do **Algoritmo 3**. A ideia central por trás desse algoritmo usa o alinhamento dos n primeiros vetores singulares direitos, que resultam da decomposição SVD - *Singular Value Decomposition* - das matrizes de centroides. SVD é uma fatoração generalizada para qualquer matriz $\boldsymbol{A_{n,m}}$, dada por $\boldsymbol{A} = \boldsymbol{LSR}^T$, em que $\boldsymbol{L}$ é chamada de matriz de vetores singulares esquerdos, $\boldsymbol{S}$ é uma matriz diagonal com os valores singulares, e as colunas de matriz $\boldsymbol{R}$ (linhas de $\boldsymbol{R^T}$) têm os vetores singulares direitos (Zaki et al. [2014]). Como $\boldsymbol{L}$ e $\boldsymbol{R}$ são ortogonais, é possível usar ambos como uma base de vetores.

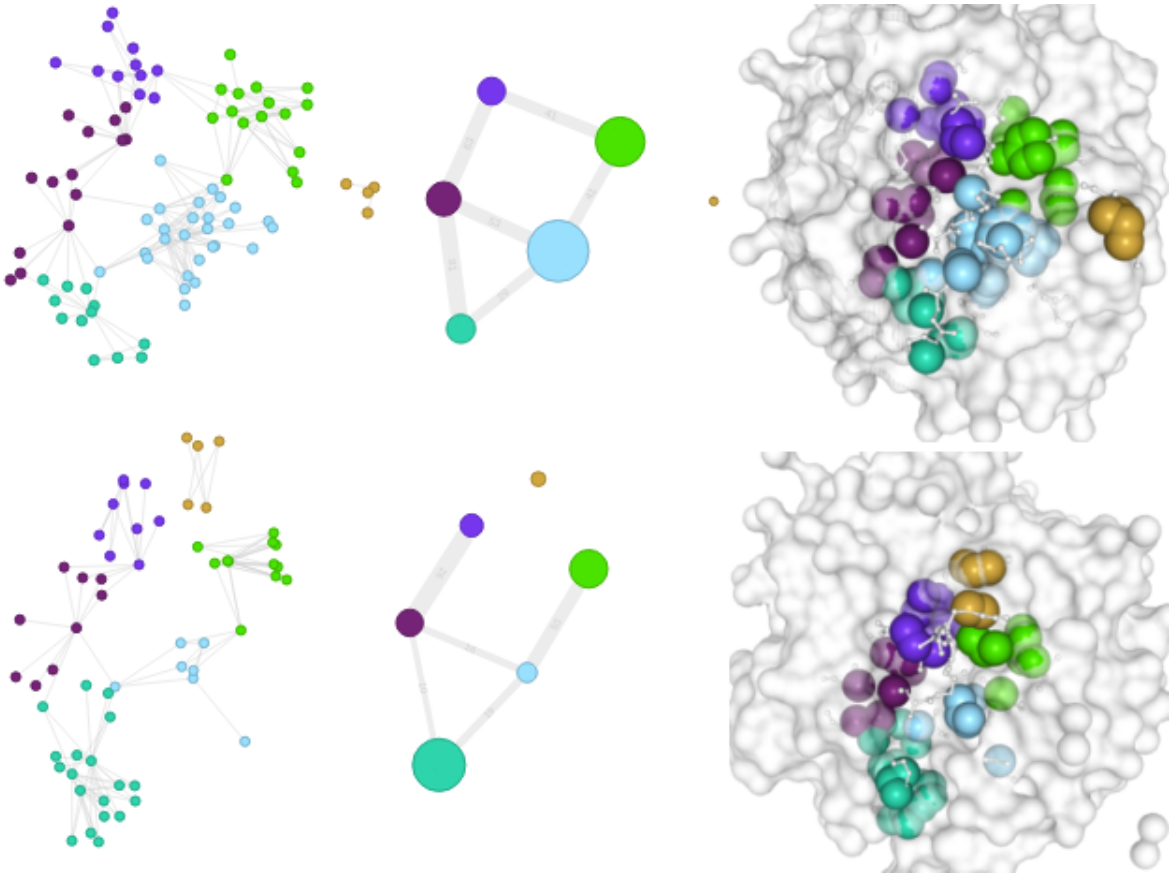


Figura 4.7: Resultado do alinhamento de grafos de alto nível para interfaces cadeia-cadeia apolares entre 1PPF (complexo entre Elastase de Leucócito Humano e Inibidor Ovomucoide de Peru) com 1TEC (complexo entre Subtilisina Termitase da bactéria *T. vulgaris* com Inibidor Eglina C de sanguessunga). À esquerda, grafos de baixo nível; ao centro, grafos de alto nível; à direita, estruturas renderizadas. Nós alinhados têm as mesmas cores, bem como respectivos átomos nas estruturas associados aos nós.

O alinhamento de duas bases de vetores diferentes pode ser feito por uma matriz de cossenos diretores, entre os ângulos dos eixos correspondentes (Kelly [2019]).

EVALUATE é uma função que associa um *score* aos alinhamentos levando em consideração a razão das áreas entre os nós sobrepostos e o cosseno dos ângulos entre as arestas sobrepostas. Quanto maior o *score*, melhor o alinhamento. Cabe destacar que o alinhamento usando *Topos* não é essencial para os resultados aqui apresentados, uma vez que os alinhamentos também poderiam ser feito manualmente. Está sendo descrito nesta tese apenas por ter sido incorporado ao GAPIN para que já pudesse ser testado e avaliado pelos usuários.

Na versão atual do GAPIN, apenas alinhamentos de interfaces cadeia-cadeia são permitidos. Alinhamentos de interface cadeia-ligantes ainda não estão habilitados. Espera-se que alinhamentos envolvendo ligantes e um estudo minucioso do algoritmo

Topos sejam objetos de outros artigos a serem publicados, compondo mais desdobramentos de trabalhos futuros a esta tese.

Um exemplo de alinhamento de grafos de alto nível para interfaces cadeia-cadeia apolares pode ser visto na figura (4.7), entre 1PPF (Elastase-Ovomucoide) e 1TEC (Subtilisina-EglinC).

4.10 Sincronização visual

Um centroide com coordenadas 3D está associado a cada nó, seja nos grafos de baixo nível ou de alto nível. Nos primeiros, os centroides são as coordenadas 3D dos átomos correspondentes. Nos segundos, os centroides são os centros geométricos das coordenadas dos átomos agrupados. Para ambas as visualizações dos grafos, é aplicado um *layout* que leva em conta essas coordenadas, projetadas no plano x-y. Isso permite manipulações sinérgicas e sincronizações visuais entre gráficos e estruturas PDB renderizadas. Refrisa-se que nos grafos de alto nível, os vértices e arestas são desenhados proporcionalmente com os respectivos volumes dos rótulos. Exemplo de sincronismo pode ser visto na figura (4.7).

4.11 GAPIN - Engenharia de Software

O GAPIN foi construído seguindo as práticas de Integração Contínua (Duvall et al. [2007]) e Entrega Contínua (Humble & Farley [2011]), em que toda vez que uma nova versão possui pelo menos um (*build*) verde, é imediatamente implantada no ambiente de produção. O código fonte da GAPIN está atualmente armazenada no *Bitbucket* (Bitbucket.org [2019a]) e utilizando *Bitbucket pipelines* (Bitbucket.org [2019b]) para o processo de entrega contínua e implementação (Figura 4.8).

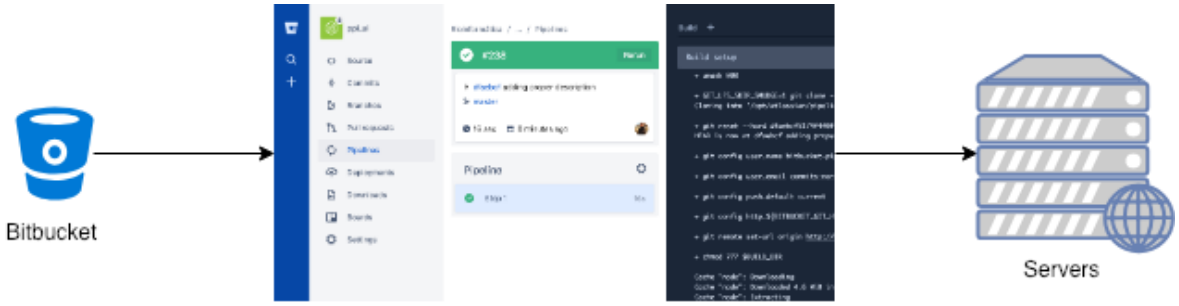


Figura 4.8: Visão geral do projeto seguindo as técnicas de *Continuous Deployment* com *Bitbucket Pipelines* em que uma vez que alguma alteração é enviada ao Bitbucket, um pipeline é iniciado rodando uma bateria de testes a fim de garantir que as alterações não afetem o código em ambiente produtivo e dê um feedback rápido antes do deployment

4.12 GAPIN - Arquitetura

A arquitetura geral do GAPIN pode ser dividida em 2 partes: um lado *Front End* e um lado *Back End*.

4.12.1 *Front end*

O GAPIN foi construído com foco na simplicidade, onde os usuários podem interagir com o sistema de forma amigável, manipular informações e fazer análises facilmente. Tudo isso como uma aplicação WEB rodando em um navegador moderno. O *layout* de estilo foi criado com o *Bootstrap 4* (Otto et al. [2015]) que por padrão é responsivo, permitindo que os usuários acessem o aplicativo a partir de diferentes dispositivos, como *Laptops*, *Tablets*, *Smartphones* e assim por diante.

O *framework* de *front end* usado pelo GAPIN é o *Handlebars* (HandleBars [2019]) e as informações que alimentam o *front end* vêm do *NodeJS*, para a parte de *back end*. A parte interativa foi escrita em *NGL* (Rose & Hildebrand [2015]) para renderizações da biomoléculas, *D3Plus* (d3plus [2019]) para os grafos das redes de contatos e *DataTables* (datatables [2019]) para informações do processamento em lote (fila de *Jobs*) interativas e pesquisáveis.

4.12.2 *Back end*

A arquitetura de *back end* da GAPIN é baseada em um padrão assíncrono com o NodeJS (Casciaro & Mammino [2016]), onde o processo é executado em segundo plano e gera promessas (Gallaba et al. [2015]).

Basicamente, como mostra a figura (4.9), a arquitetura *back end* do GAPIN é dividida em 3 partes:

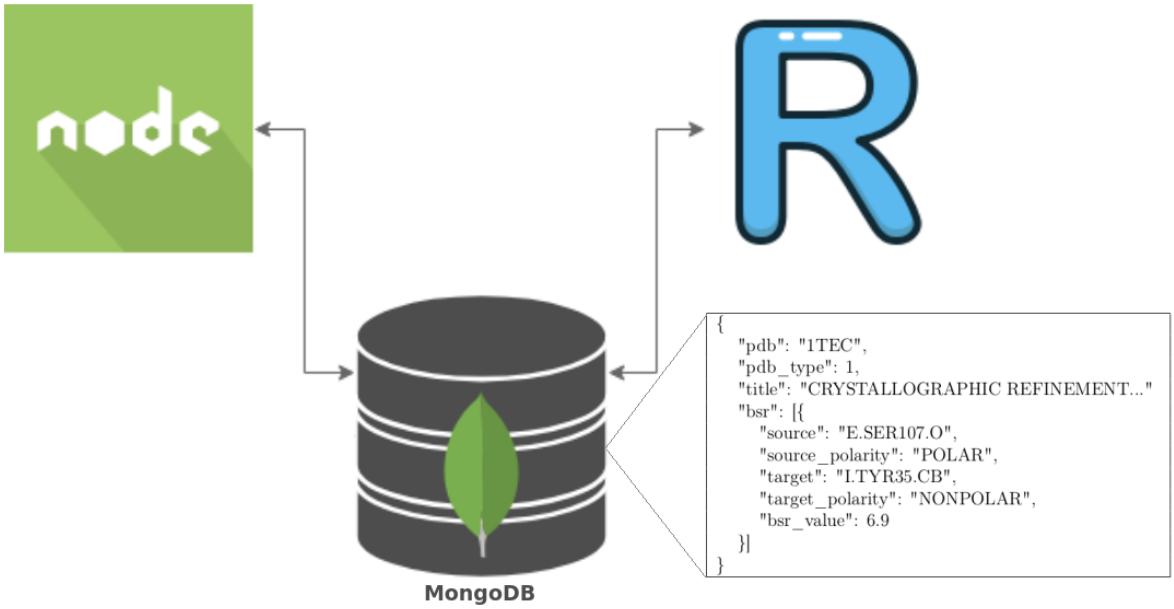


Figura 4.9: Arquitetura *Back end* do GAPIN, com um exemplo de conteúdo (alguns dados da 1TEC) no formato JSON armazenado no MongoDB

1. Aplicativo construído em *NodeJS* (nodejs.org [2019]) e *ExpressJS* (expressjs.org [2019]).

2. Processador escrito em R (R Project [2019])

3. Armazenamento com o MongoDB (MongoDB [2019])

Com base nessa arquitetura, a GAPIN pode tirar proveito do MongoDB que armazena documentos usados para carregar as biomoléculas na parte frontal sem qualquer etapa de processamento, tornando a experiência do usuário boa devido à velocidade de carregar dados, como os átomos que pertencem a rede de contato das interface biomoleculares (figura 4.9).

Capítulo 5

Resultados e Discussões

Um dos grandes objetivos desse trabalho foi habilitar outros pesquisadores a utilizarem o mesmo modelo dos estudos de caso em qualquer outro conjunto de biomoléculas bioativas independente se fazem parte da base de dados do PDB ou se são estruturas customizadas, de forma simples e idempotente, ou seja, que sempre tenha o mesmo resultado para o conjunto de dados selecionado.

Outro ponto importante era que esse habilitador fosse de fácil manuseio e intuitivo e interativo em que o usuário pudesse analisar as estruturas de interfaces no nível atômico, e/ou grupos, bem como uma análise da estrutura da biomolécula.

5.1 GAPIN - Usabilidade

Apresenta-se a seguir, resultados do desenvolvimento do GAPIN, com detalhes da interface computacional e sua usabilidade, com alguns exemplos.

5.1.1 GAPIN - Fluxo de Utilização

Um dos princípios do GAPIN é que sua arquitetura foi pensada em prover um mecanismo colaborativo em que toda vez que um usuário entra com os dados de algum PDBid, o mesmo passa a estar disponível para qualquer outro usuário.

Existe três formas de se interagir com o GAPIN em um primeiro momento sendo eles:

1. Quando não existe o PDBid na base e deseja-se importar o PDBid do site oficial do PDB (https://www.rcsb.org/).

2. Quando não existe o PDBid na base e deseja-se importar um PDBid customizado a partir do computador do usuário.

3. Já existe o PDBid na base do GAPIN e deseja-se iniciar uma análise.

Para o primeiro e segunda casos, basta digitar o código do PDBid desejado em qualquer página do GAPIN na parte superior direita conforme Figura (5.1) e clicar em (*search*). Esse formato permite a entrada de até 5 PDBs separados por virgula por vez. Após clicar em (*"import it right now"*), O usuário será redirecionado imediatamente para a tela de importação do GAPIN conforme (figura 5.2).

Figura 5.1: Página inicial do GAPIN em que o usuário busca por um PDBid que não existe ainda na base de dados

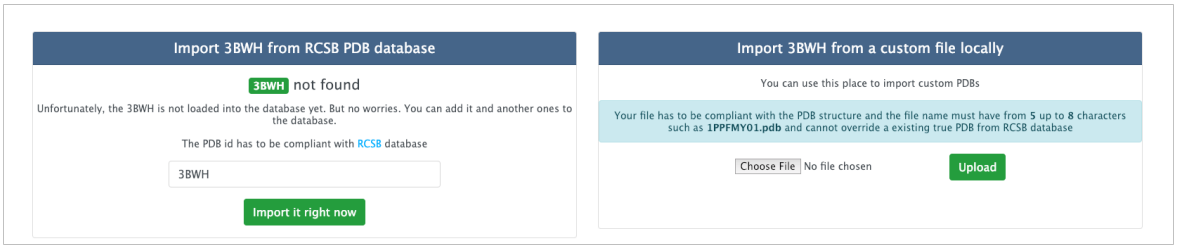


Figura 5.2: Página de importação de PDBid do GAPIN. Do lado direito a opção de importar o PDB pela base do PDB e do lado direito a opção de realizar (*upload*) de um arquivo PDBid customizado.

Caso a opção seja de importar o PDB pela base oficial, o PDBid deve ser exatamente o mesmo oficialmente armazenado nessa base. Caso contrário, se a opção for importar um PDBid customizado, o usuário deve limitar o nome do arquivo de 5 a 8 caracteres e o arquivo não pode passar de 2Mb. Obviamente, o arquivo também deve estar em conformidade com o formato do PDB disponível em: (http://www.wwpdb.org/documentation/file-format).

Independente do processo escolhido, após iniciar o processo de importação (figura 5.3) o usuário pode acompanhar o andamento da importação que por sua vez é executada de forma assíncrona, na página de (*jobs*). Vide figuras (5.3, 5.4 e 5.5)

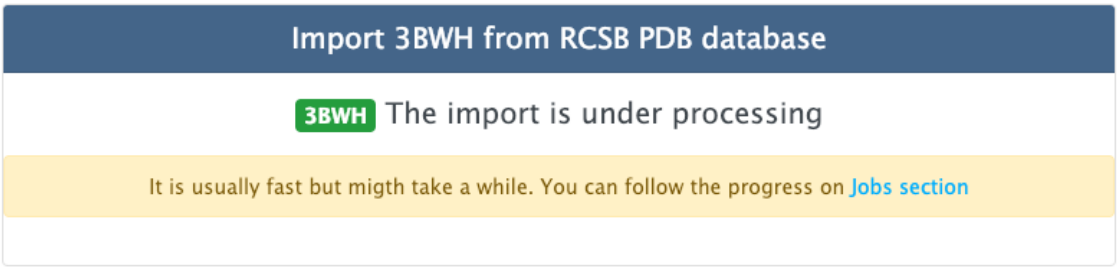


Figura 5.3: Após o processo de importação ser iniciado, o usuário poderá acompanhar o andamento da importação pela tela de (*jobs*)

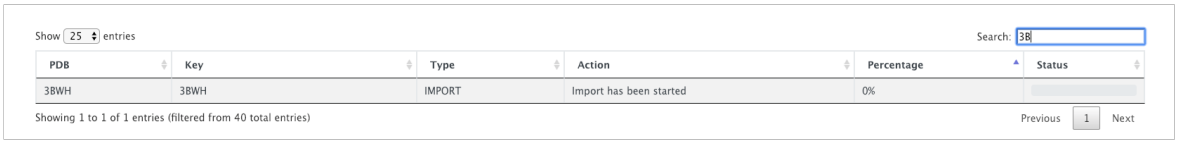


Figura 5.4: Início do processo de importação de qualquer PDB

Figura 5.5: Primeiros processos já finalizados para a etapa de importação.

Após a importação, a coluna que remete ao PDBid ficará habilitada com um *link* que redirecionará o usuário à tela principal da aplicação.

No lado direito (figura 5.6C) os usuários podem interagir com a estrutura da interface, como girar para qualquer lado com o mouse utilizando o botão esquerdo, manipular o *zoom* ampliando ou reduzindo segurando o botão direito do *mouse* e rolando para cima ou para baixo, e reposicionar a estrutura e qualquer parte da camada com o clique do meio do *mouse*.

Do lado inferior esquerdo, GAPIN provê uma visualização em duas dimensões dos contatos identificados a nível atômico. Ao passar o *mouse* sobre qualquer átomo nessa estrutura é possível ver a composição do nome do átomo no seguinte formato: {nome da cadeia}.{nome do resíduo}{identificador único do resíduo}.{nome do elemento} (figura 5.7). Nesse caso temos um carbono Gama que faz parte do resíduo fenilalanina da enzima em 1PPF.

Ao clicar sobre esse elemento, um destaque é dado as conexões de primeiro nível bem como a seleção é refletida na estrutura do lado direito (figura 5.8). Nota-se também

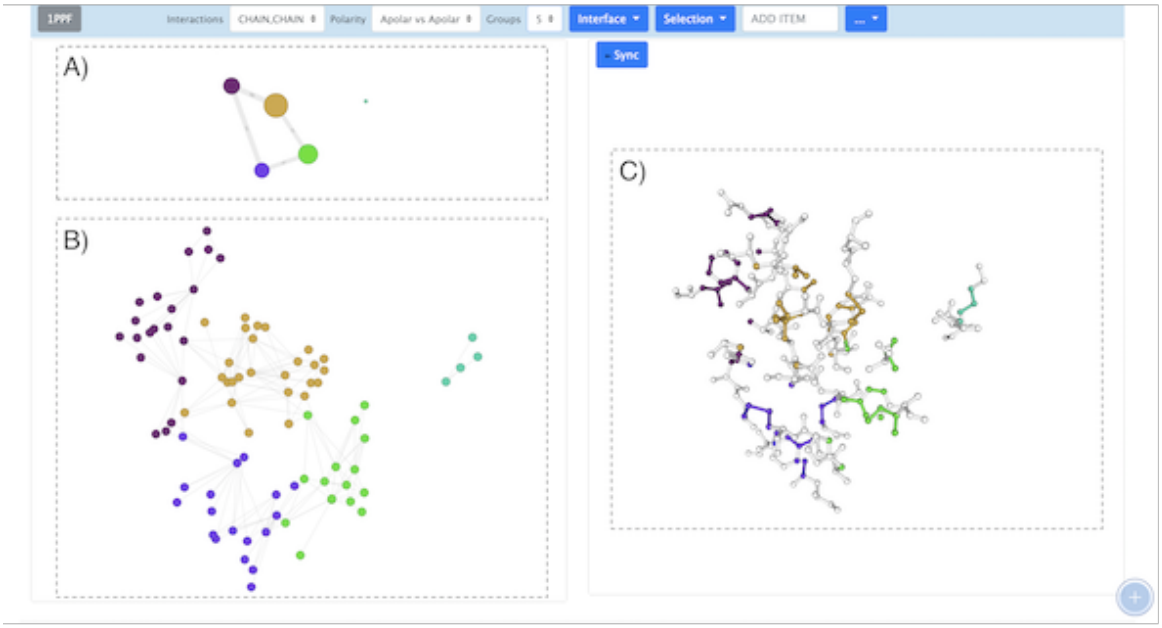


Figura 5.6: Tela principal do GAPIN exibindo a 1PPF. A) Grafo de alto nível o qual mostra os agrupamentos. Nesse caso pode-se ver a divisão em 5 grupos sendo 4 contectados e um desconexo. B) Grafo de baixo nível ou nível atômico o qual cada nó do grafo é um átomo. C) estrutura em 3D interativa da proteína em questão.

que ao selecionar um átomo, uma caixa aparece em destaque com o nome dos átomos que possuem conexões de primeiro nível. Outra informação disponível é o valor da área de contato entre o átomo selecionado e seus vizinhos de primeiro nível em *Angstrons* quadrados (*Link weight*).

Na parte superior esquerda, é possível ver a mesma distribuição das interações do grafo a nível atômico, só que agrupado por número de grupos. Veja na figura (5.6) que a as conexões seguem o mesmo padrão, todavia agrupados pelos átomos que compõem aquele grupo. Da mesma forma que o nível atômico, ao passar o *mouse* sobre um grupo é possível ter acesso a algumas informações adicionais tais como valor acumulado da área de contato de todos os átomos que fazem parte daquele grupo.

Bem como o grafo a nível atômico, o usuário pode selecionar qual grupo gostaria de dar destaque clicando sobre o mesmo e vendo em destaque o que essa seleção significa na estrutura da biomolécula e na sua rede a nível atômico (figura 5.10).

A barra de navegação superior, oferece algumas possibilidades de customização da visualização no GAPIN. Da esquerda para a direita é possível visualizar a biomolécula por interações:

- CHAIN, CHAIN: Interações cadeia-cadeia.

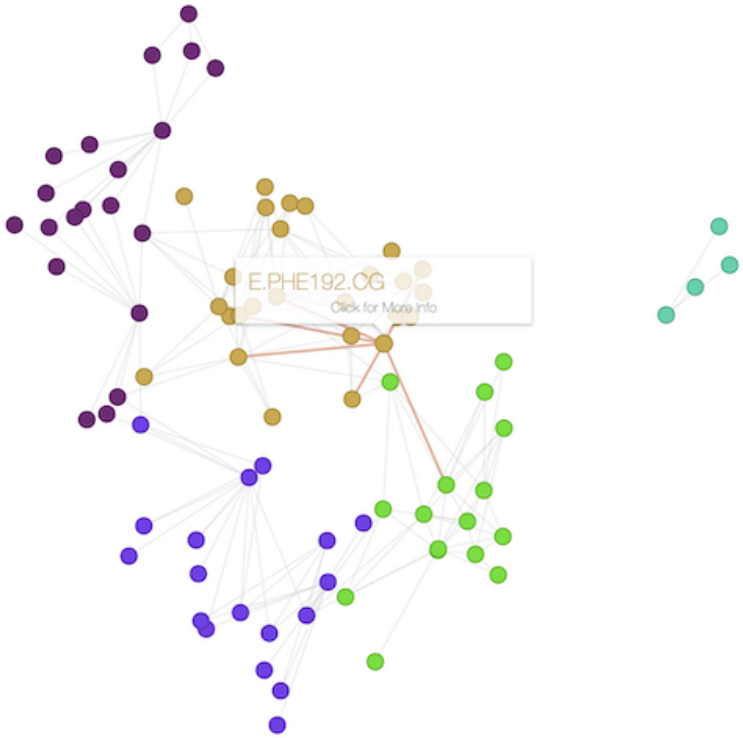


Figura 5.7: *Mouse hover* sobre átomos no grafo da rede de interações.

- ANY, LIG: Interações cadeia-ligante.

Em seguida, existe a opção de visualização por polaridade sendo:

- Apolar, Apolar: Somente interações apolares.
- Polar, Polar: Somente interações polares.
- ALL: Qualquer tipo de interação sendo apolares ou polares

A próxima opção é o número de grupos em que o GAPIN vai subdividir as interfaces. Esse número varia de 1 a 20 sendo que o limite máximo é calculado baseado na qualidade mínima de 50

Todas essas opções refletem tanto na estrutura da biomolécula, quanto nos grafos que representam as redes de contato.

O próximo conjunto de opções refletem a estrutura. O primeiro (*dropdown*) chamado interfaces, force combinações possíveis de interações com a interface do lado direito, sendo elas:

- Representação da interface de contato em *ball+stick* (figura 5.11A) e *spacefill* (figura 5.11B).

Figura 5.8: Interação do usuário com a rede de nível atômica do lado inferior esquerdo refletindo a seleção na estrutura do lado direito

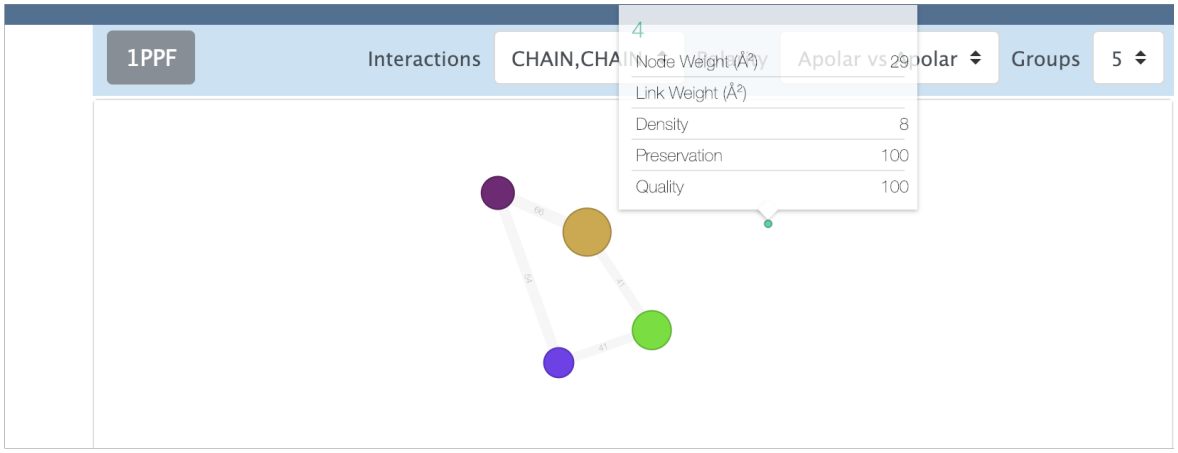


Figura 5.9: Interação do usuário com (*mouse hover*) do grafo em nível de grupos

- Representação em cores por grupos (figura 5.12A), por cadeias (figura 5.12B), por elementos (figura 5.12C), por polaridade (figura 5.12D) e sem definições de cores (figura 5.12E).
- Visualização das interfaces contendo todas as interações sendo CHAIN,CHAIN ou ANY,LIG acrescido da estrutura completa do resíduo em cor branca (figura 5.13A), somente ANY sem os ligantes também com a estrutura completa dos resíduos (figura 5.13B), somente os ligantes com a estrutura completa dos resíduos

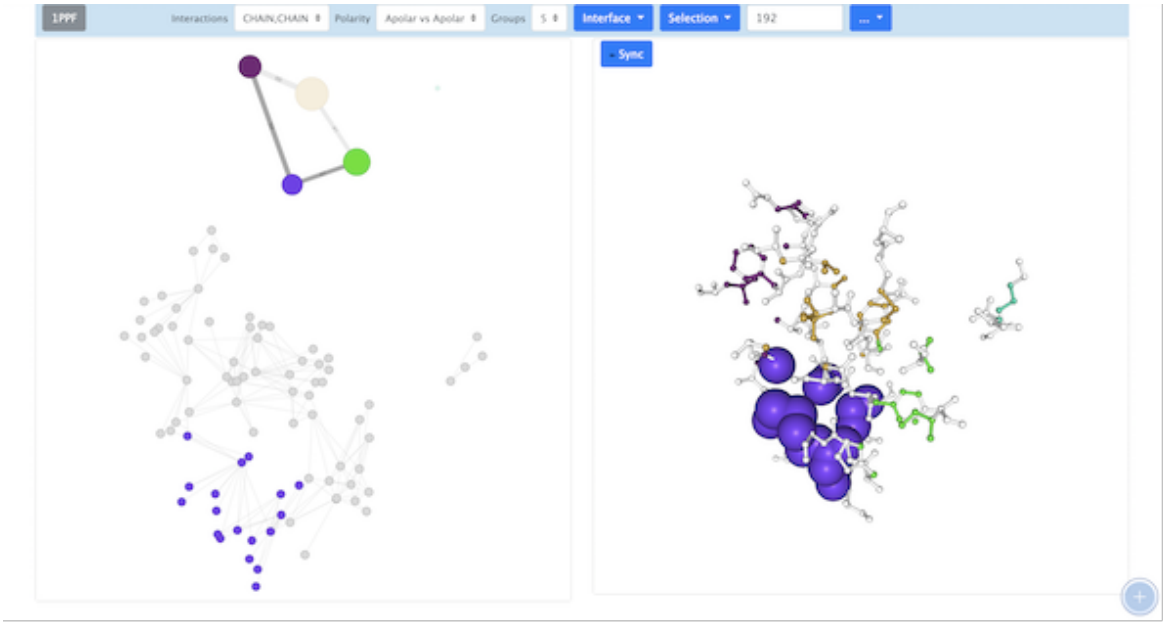


Figura 5.10: Interação do usuário com a seleção de grupos refletindo na estrutura do lado direito e na rede a nível atômico do lado inferior esquerdo

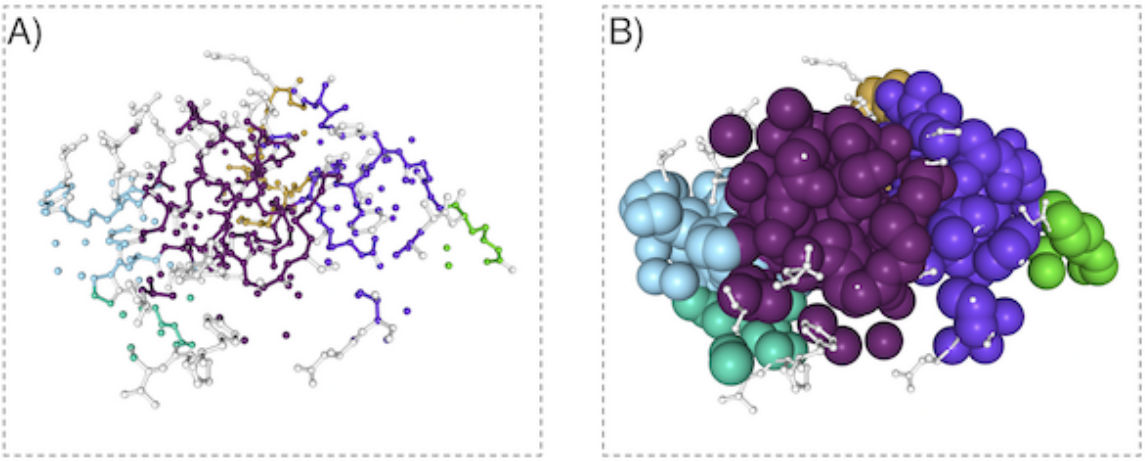


Figura 5.11: Representações em dois formatos distintos da 1PPF. A) Representação em *ball+stick* B) Representação em *spacefill*

(figura 5.13C). Todavia é possível ter as mesmas representações considerando somente os átomos que fazem parte do contato sem a visualização da estrutura completa dos resíduos (figuras 5.13E, F e G).

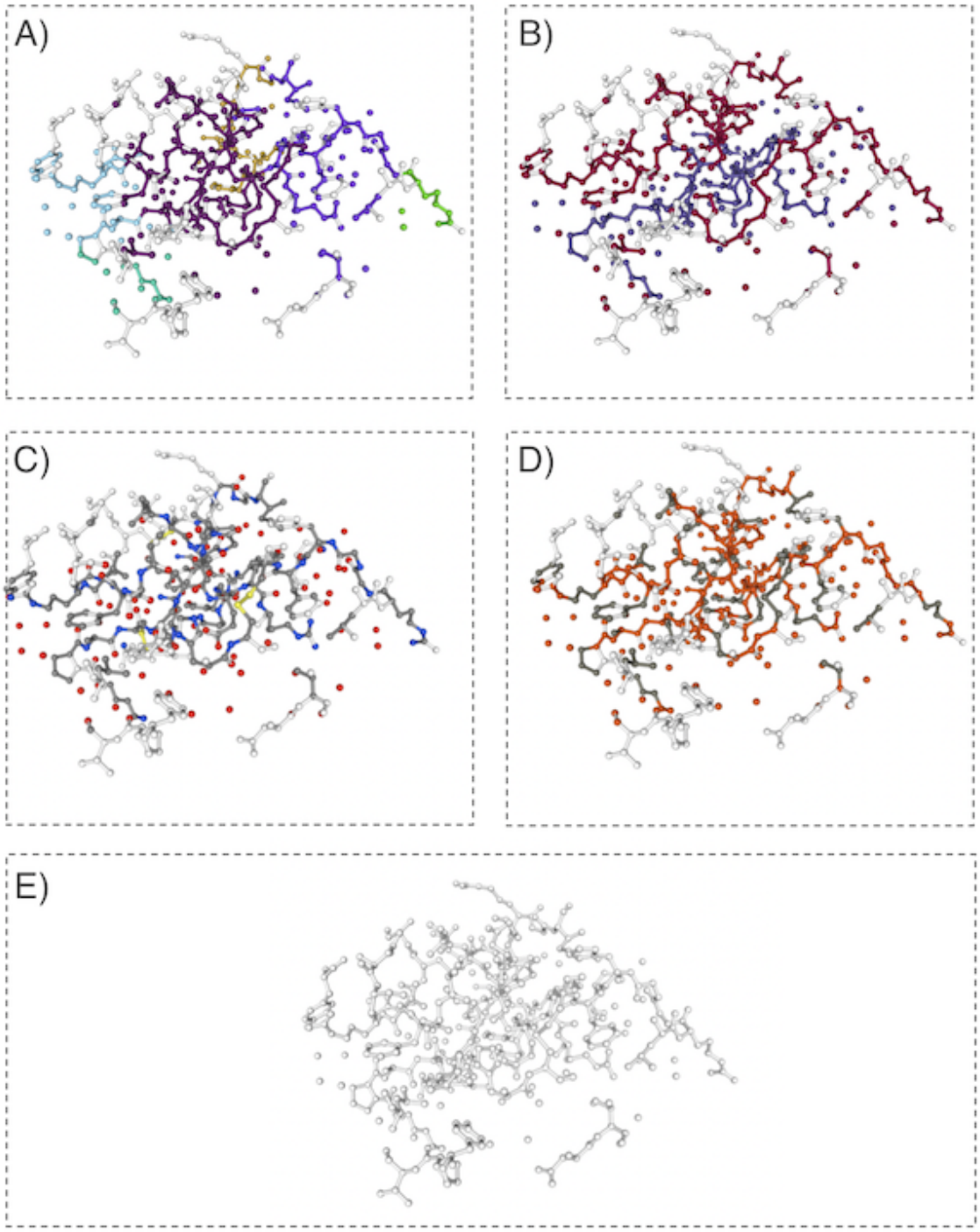


Figura 5.12: Representação das cores no GAPIN. A) Divisão de cores por grupos. B) Divisão de cores por cadeias. C) Divisão de cores por elementos. D) Divisão de cores por polaridade. E) Sem divisão de cores

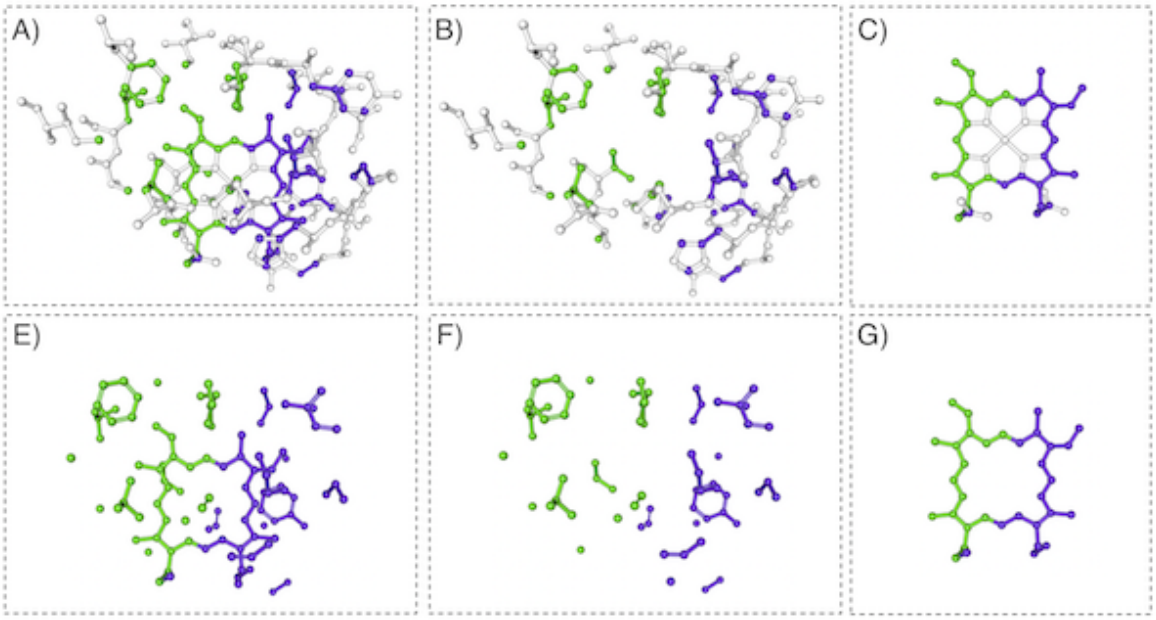


Figura 5.13: Representação das combinações de visualização da biomolécula. A) Somente ANY sem os ligantes também com a estrutura completa dos resíduos B) Somente os ligantes com a estrutura completa dos resíduos C) Somente os átomos que fazem parte do contato D) Somente os átomos que fazem parte do contato sem a visualização da estrutura completa E) Somente os átomos que fazem parte do contato sem os resíduos

- Representação da estrutura em formato *surface* das cadeias (figura 5.14A), das cadeias e ligantes (figura 5.14B), cadeias e águas (figura 5.14C), e somente ligantes (figura 5.14D).

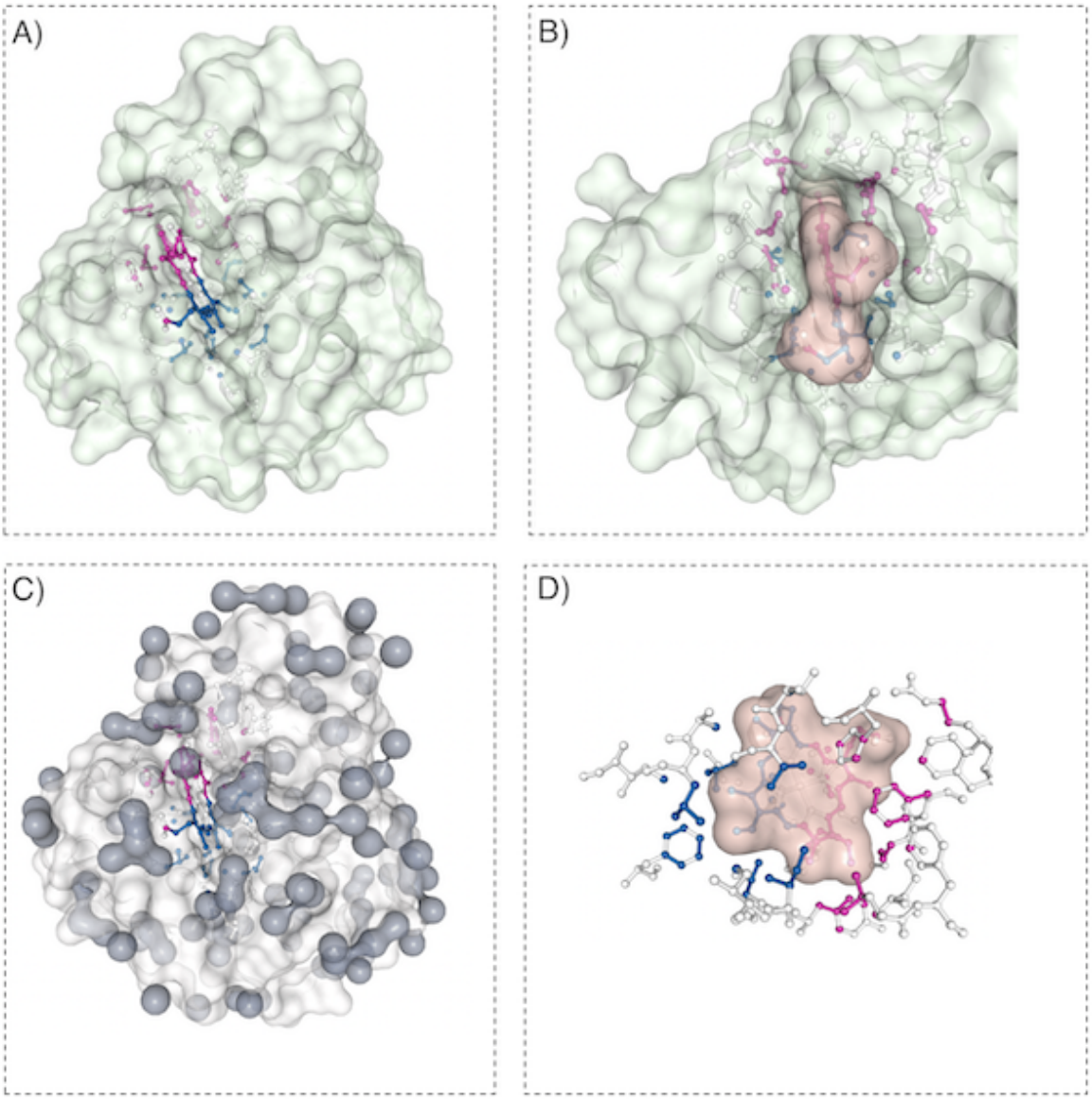


Figura 5.14: A) Representação das combinações de *surface* de uma biomolécula pelo GAPIN. B) Cadeias e ligantes B) Cadeias e águas C) Somente ligantes

Existem inúmeras combinações possíveis de visualizações como por exemplo a (figura 5.15) em que é possível ver todas as interfaces que tangem a biomolécula 1HMD Holmes et al. [1991]. Outra forma de representação pode ser vista com a 1PPF em que é possível dar um destaque a enzima em formato de *surface* e a interface separada por

cores dos grupos (figura 5.16). Outro formato é a mesma 1PPF agora com o inibidor também em formato *surface* (figura 5.17)

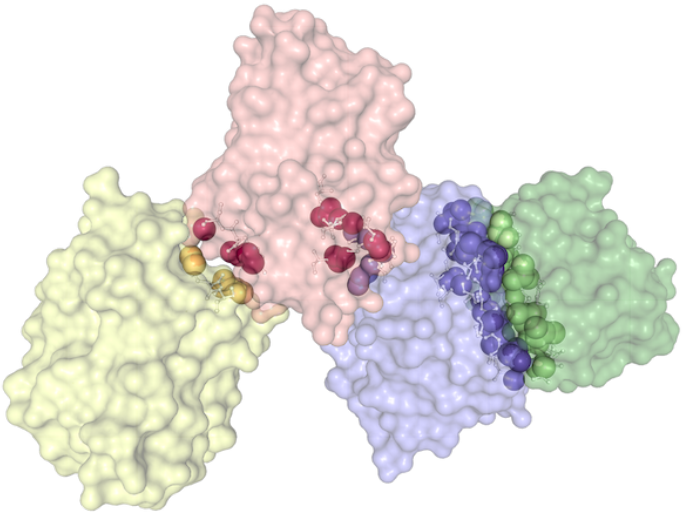


Figura 5.15: Representação das interfaces da 1HMD

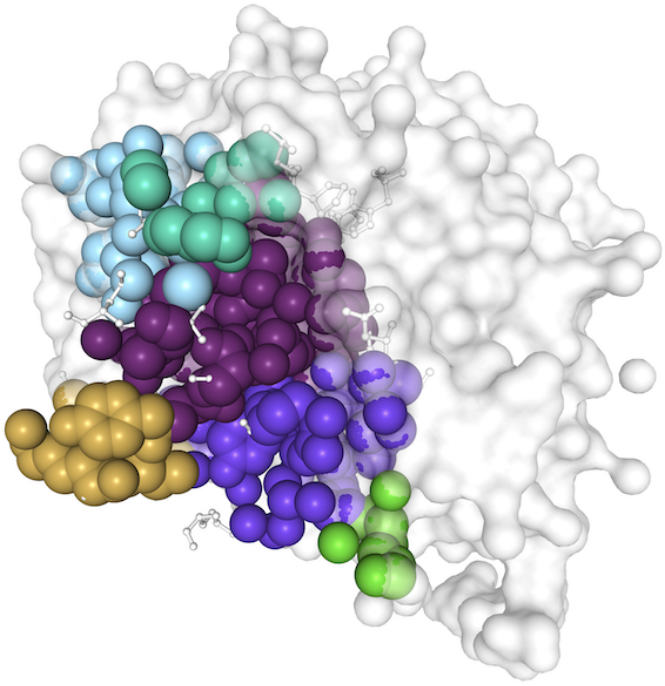


Figura 5.16: Representação proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface com o Inibidor em formato $Ball + Stick$ coloridos por uma distribuição de 5 grupos.

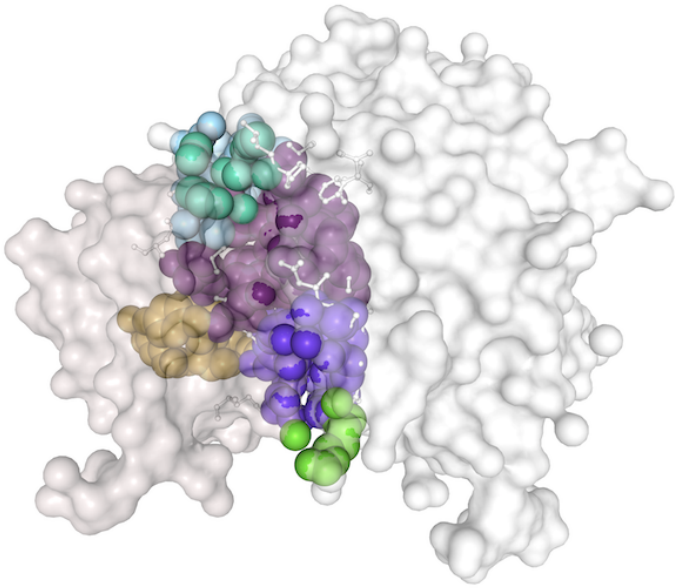


Figura 5.17: Representação proteína 1PPF dando destaque a enzima e inibidor em formato *surface* e os átomos que fazem parte da interface $Ball + Stick$ coloridos por uma distribuição de 5 grupos.

O próximo conjunto de opções de interação com a interface remete ao que tange a seleção nos grafos a esquerda, seja em nível de grupos ou a nível atômico. O conjunto de opções fornece:r

- Representação da interface de contato em *ball + stick* (figura 5.18) ou *spacefill* (figura 5.19)

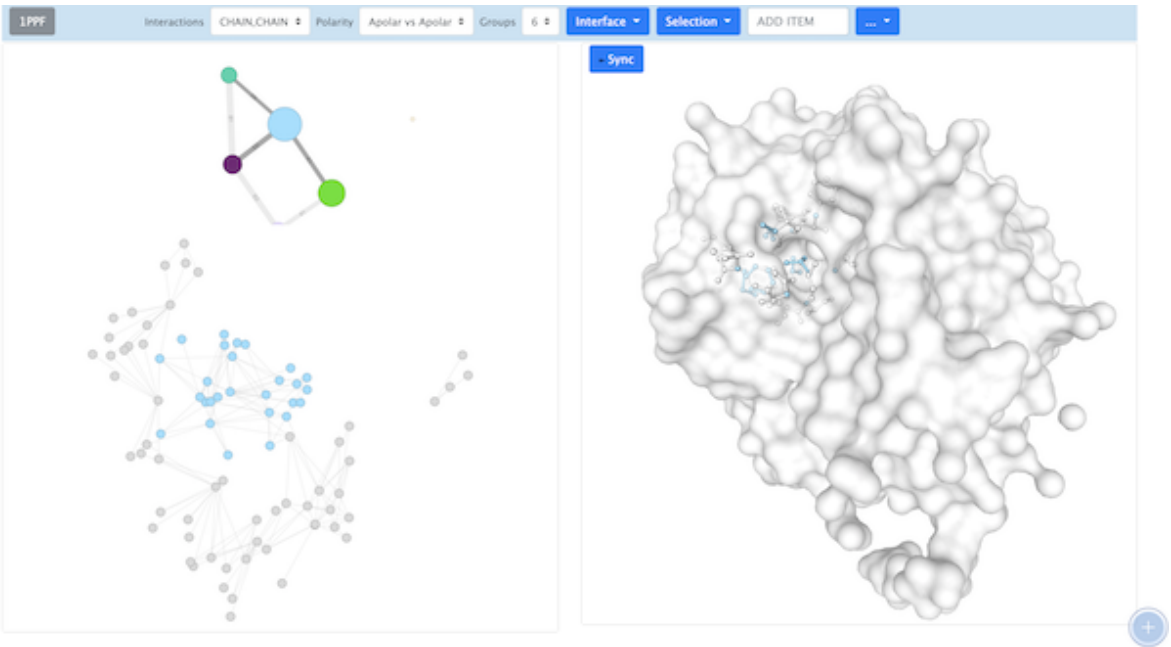


Figura 5.18: Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a Enzima em formato (*surface*) e os átomos que fazem parte da interface em formato (*Ball Stick*) contendo a estrutura completa do resíduo em que o átomo pertence

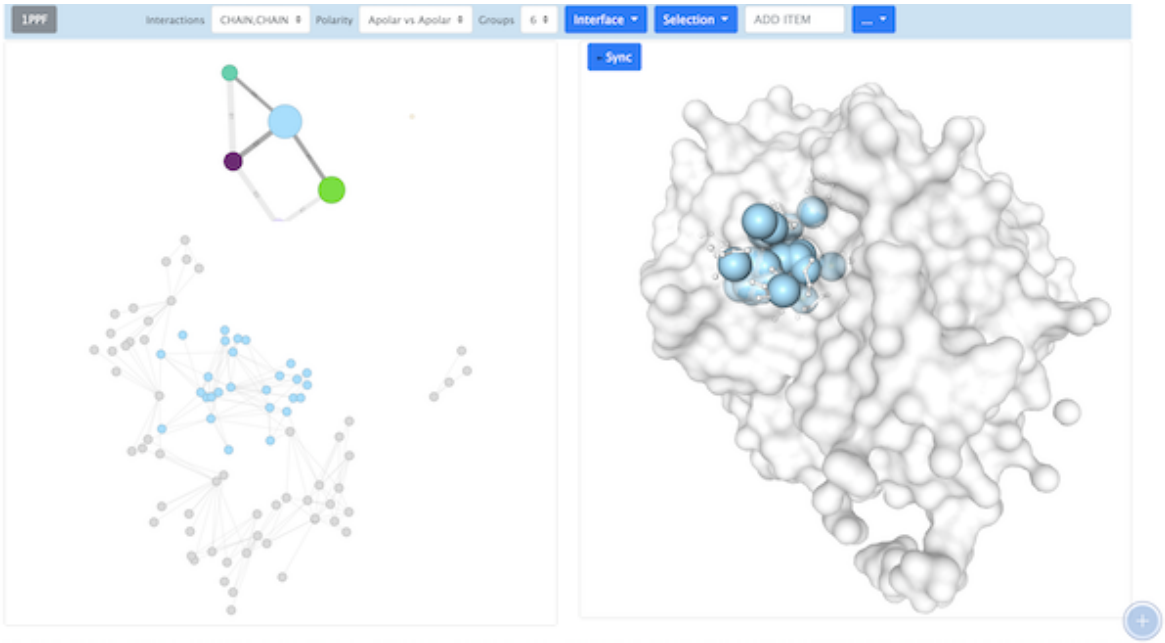


Figura 5.19: Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato *Spacefill* contendo a estrutura completa do resíduo em que o átomo pertence.

- Representação da interface de contato em *ball + stick* (figura 5.20) ou *spacefill* (figura 5.21) sem a representação do resíduo o qual pertencem os átomos da seleção.

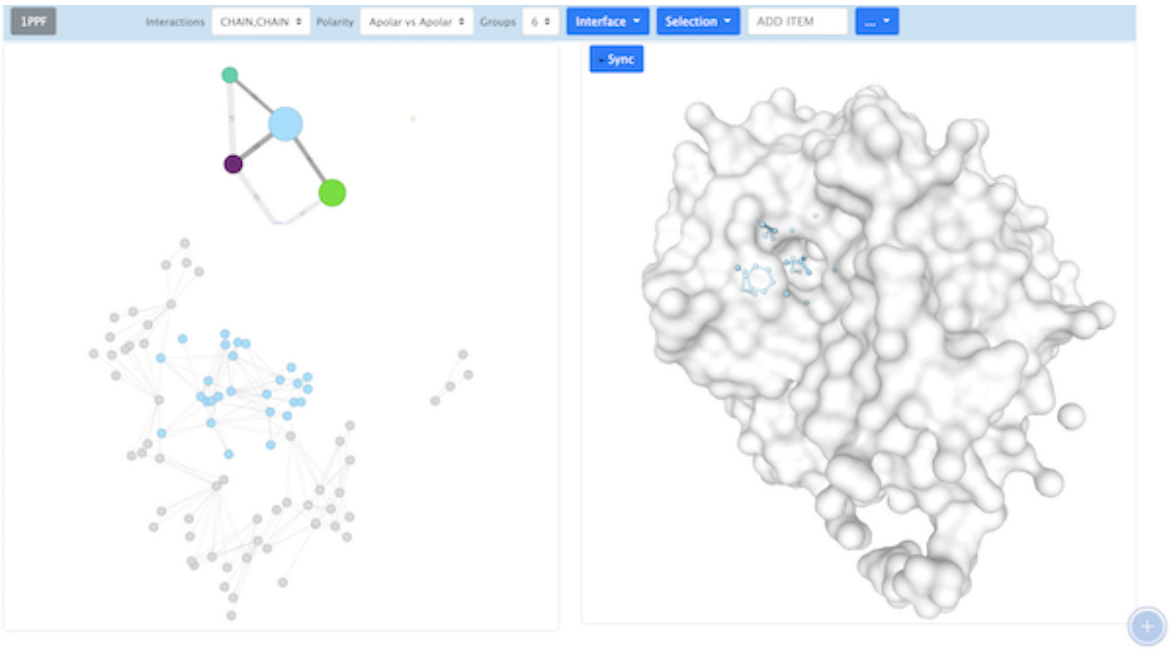


Figura 5.20: Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato *BallStick*)

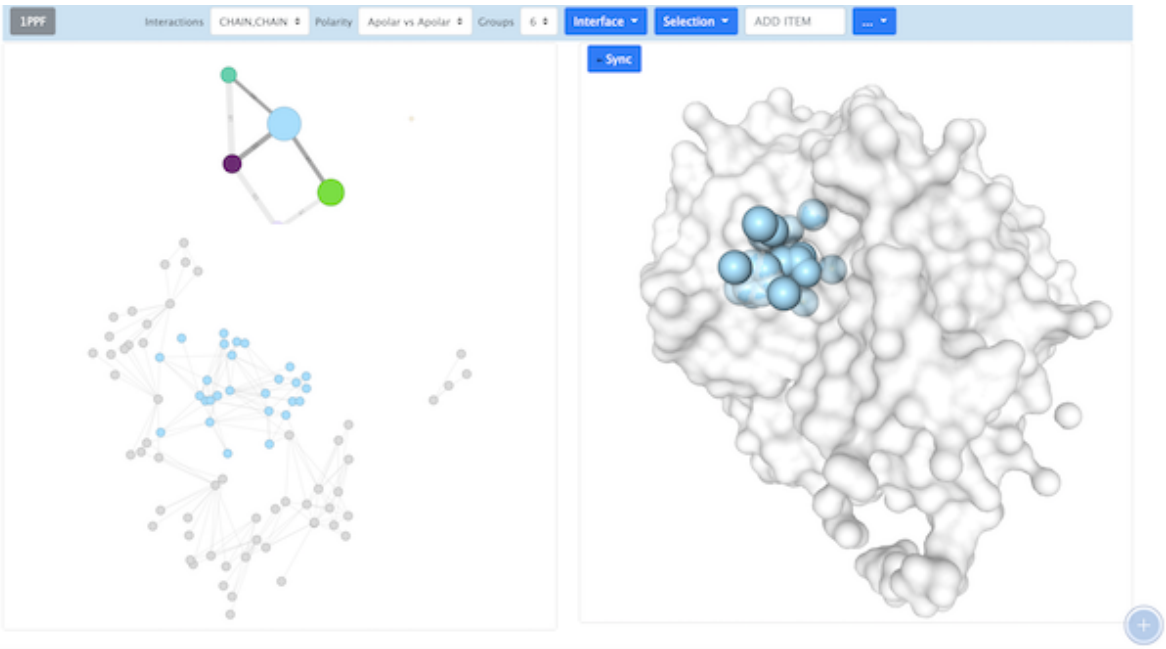


Figura 5.21: Representação somente da interface selecionada do grupo 6 da proteína 1PPF dando destaque a enzima em formato *surface* e os átomos que fazem parte da interface em formato *spacefill*.

Bem como a visualização dos átomos que compõem a interface podem ser caracterizados por cores, o mesmo se aplica a interface selecionada, ou seja, cores por cadeias (figura 5.22A), por grupos (figura 5.22B), por elementos (figura 5.22C), por polaridade (figura 5.22D).

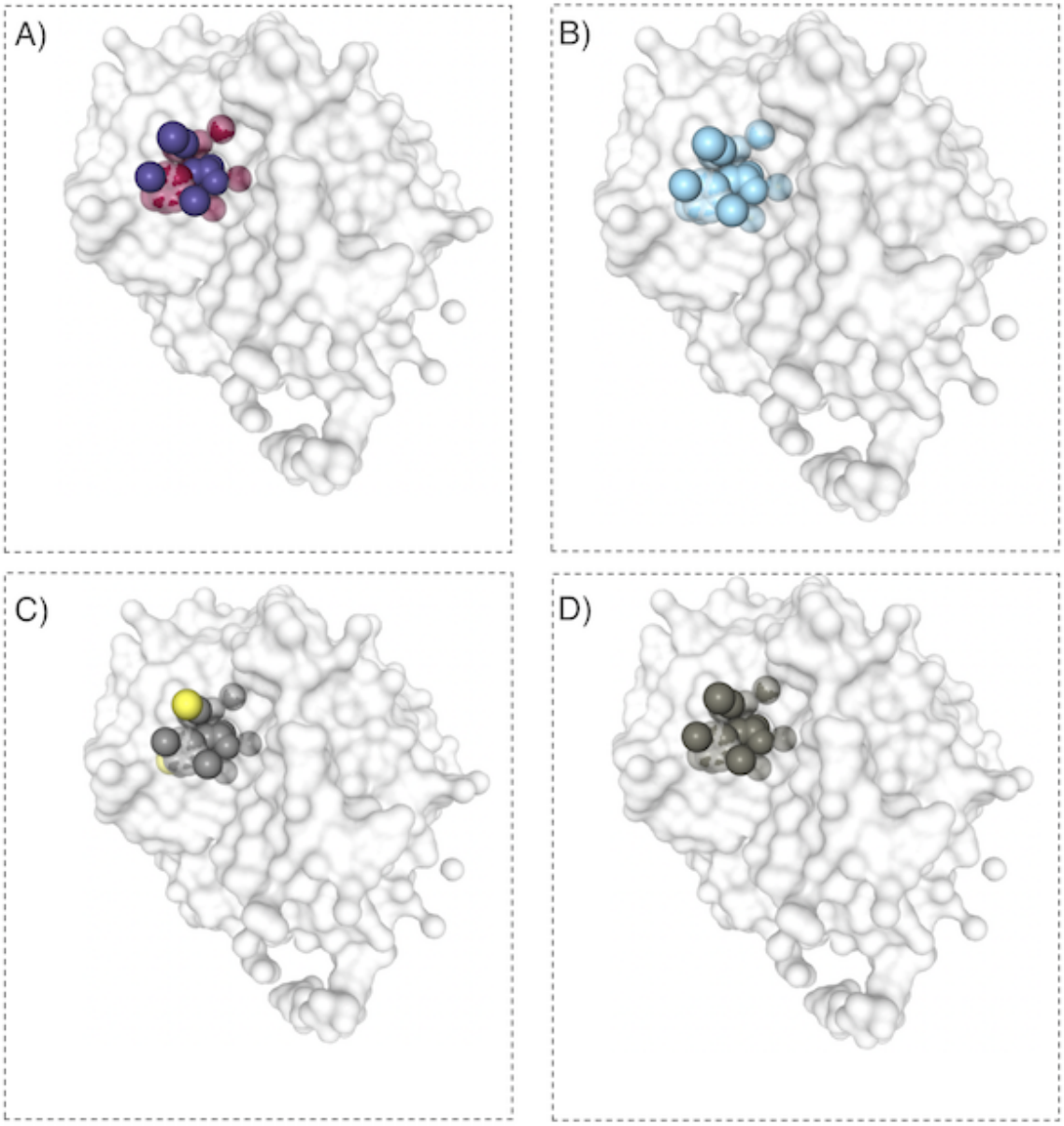


Figura 5.22: Representação das cores da interface selecionada no GAPIN

A última opção disponível desse conjunto é para limpar (*clear*) toda a seleção realizada.

Imediatamente a direita do menu *interfaces* existe uma opção para adicionar individualmente átomos ou conjunto de átomos que tenham correspondência com a

sequência digitada nesse campo para busca de um elemento. O padrão de busca é o formato utilizado pelo GAPIN para nomear contatos, ou seja, se o usuário digitar I.LEU18.CB na tela de análise da 1PPF, GAPIN irá selecionar esse elemento e dar destaque, conforme figura (5.23A). Caso o usuário queira visualizar todo o resíduo, basta digitar I.LEU18 e o a Leucina 18 do inibidor será colocada em evidência conforme figura (5.23B). Pode-se também selecionar apenas o nome de uma das cadeias e todos os elementos que pertencerem a essa cadeia serão selecionados (figura 5.23C). O fato de já existir uma seleção pré-existente não impede a adição de novos elementos conforme figura (5.23D) em que já havia uma seleção em torno do elemento I.GLU19.CB e suas conexões de primeiro nível e fora adicionado o resíduo I.LEU18.

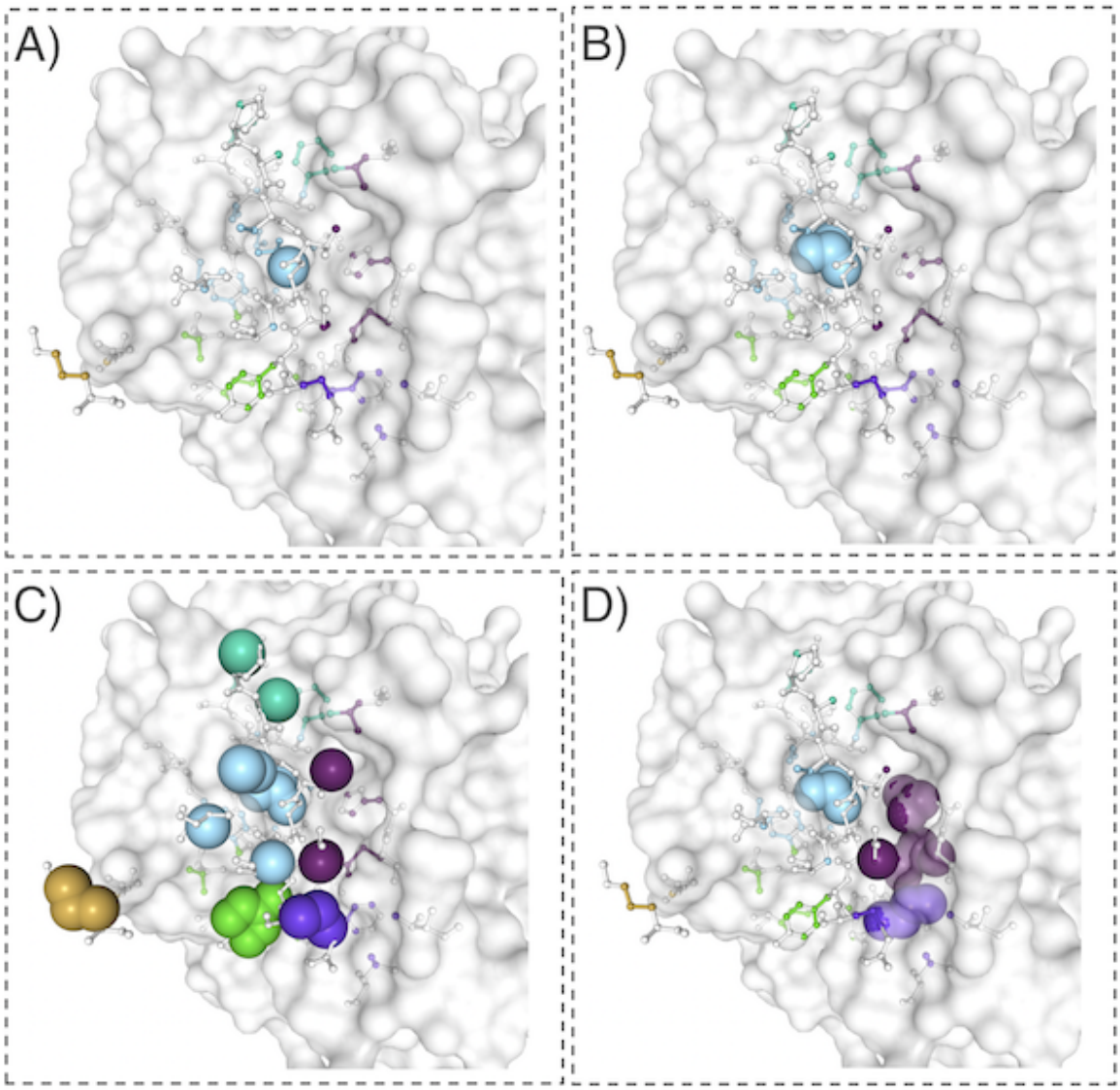


Figura 5.23: Modelo de seleção manual de átomos. A). Caso o usuário queira visualizar todo o resíduo, basta digitar I.LEU18 e o a Leucina 18 do inibidor será colocada em evidência conforme figura B) Pode-se também selecionar apenas o nome de uma das cadeias e todos os elementos que pertencerem a essa cadeia serão selecionados C) O fato de já existir uma seleção pré-existente não impede a adição de novos elementos conform D) em que já havia uma seleção em torno do elemento I.GLU19.CB e suas conexões de primeiro nível e fora adicionado o resíduo I.LEU18.

Inicialmente, o GAPIN já provê um conjunto de cores iniciais as quais podem ser ajustadas caso o usuário assim queira no meu *more options* conforme figura (5.24). Esse recurso permite ao usuário por meio de uma paleta de cores, selecionar as cores desejadas para representar os elementos de um grupo e a cor da *surface* da estrutura. Na medida que o usuário for utilizando o mouse para mudar as cores, as mesmas

imediatamente serão aplicadas tanto no grafo quanto na estrutura.

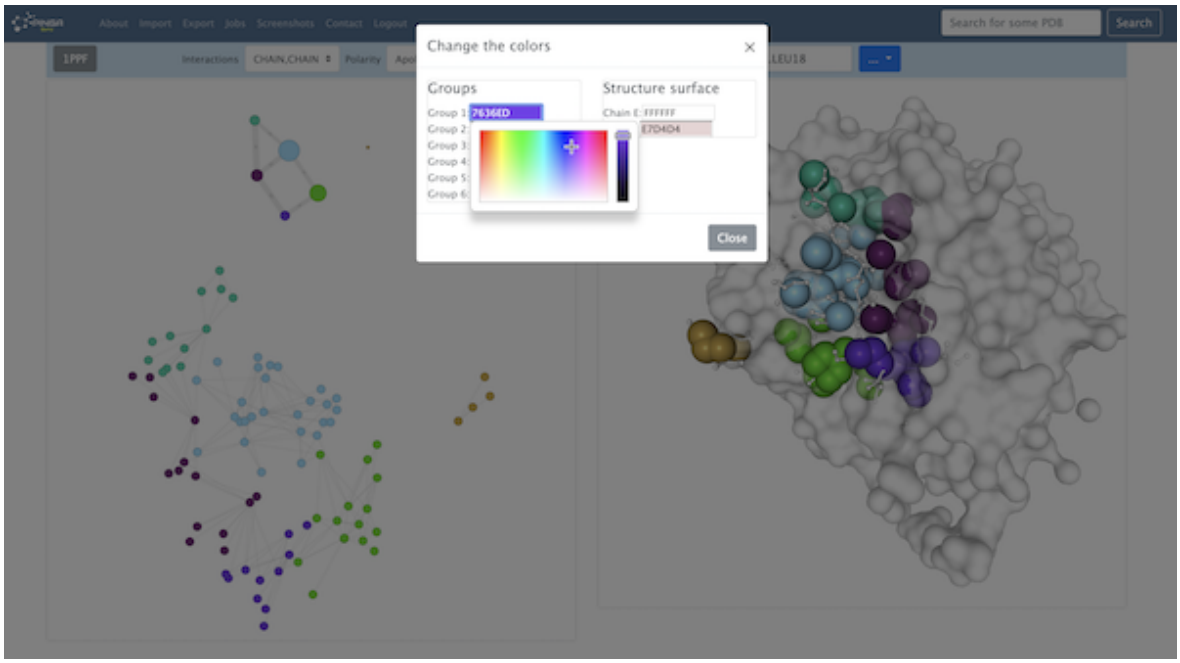


Figura 5.24: Opção de seleção de cores pelo GAPIN para o grafo de contatos e a estrutura.

Além disso, o menu *more options* provê a opção de visualização por cadeias, excluindo as cadeias não selecionadas. Outro recurso utilitário do GAPIN é a opção de tirar *screenshots* dos grafos e da estrutura, ambos em alta resolução.

GAPIN também provê a possibilidade de adicionar outas biomoléculas na mesma página conforme figuras (5.25) e (5.26).

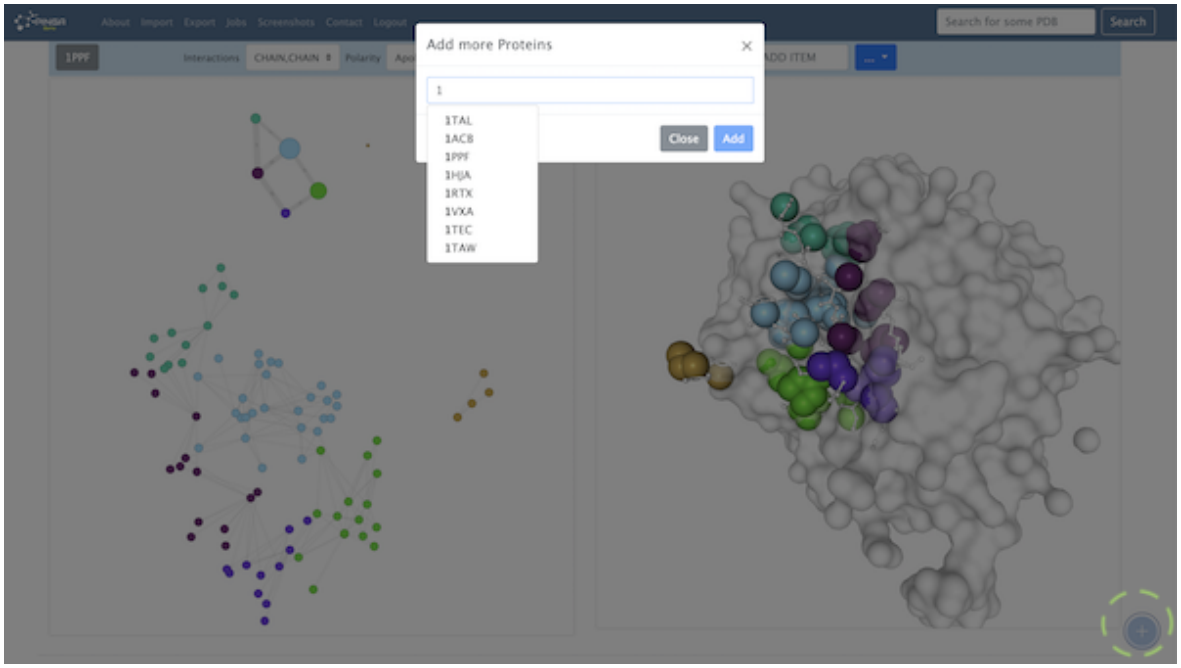


Figura 5.25: Fluxo de adicionar uma nova biomolécula no GAPIN para visualização.

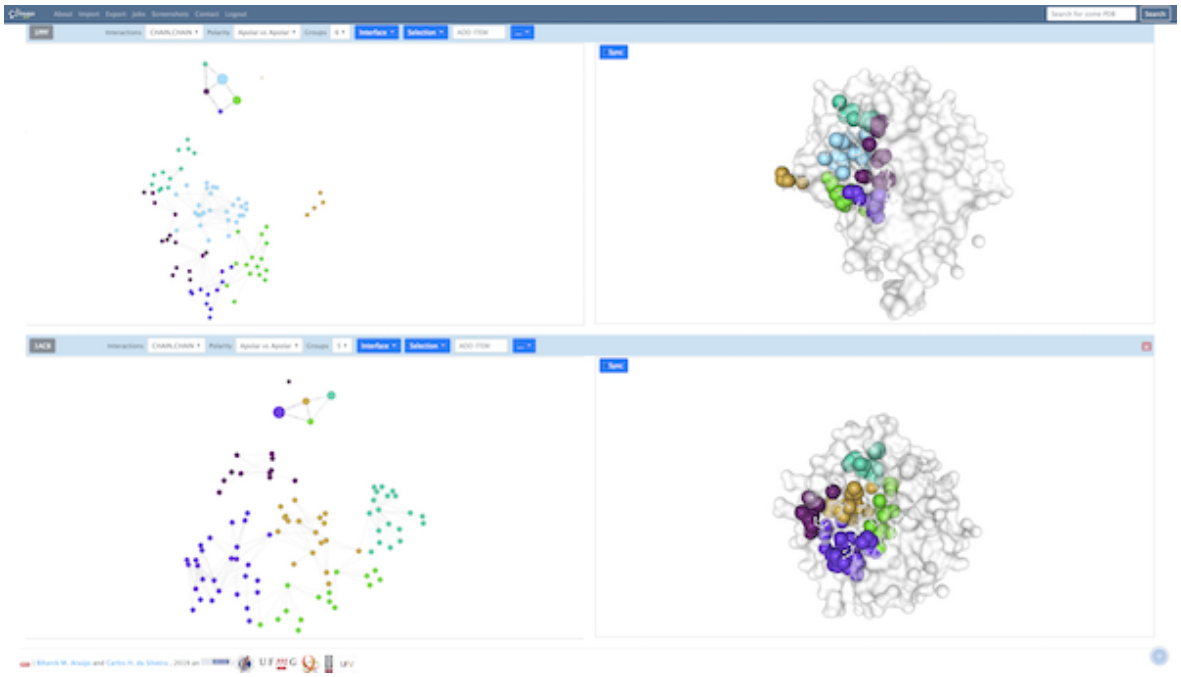


Figura 5.26: Tela contendo duas biomoléculas para análise

Durante a etapa de análise da estrutura, alguns movimentos podem tirar o sincronismo entre os grafos e a mesma. Todavia, existe uma ação no sistema para sincronizar o grafo em relação a posição atual da estrutura a fim de gerar correspondência entre as formas. Para tal, basta o usuário clicar no botão *sync* (figura 5.27A), que os grafos

serão reorganizados em relação a estrutura (figura 5.27B).

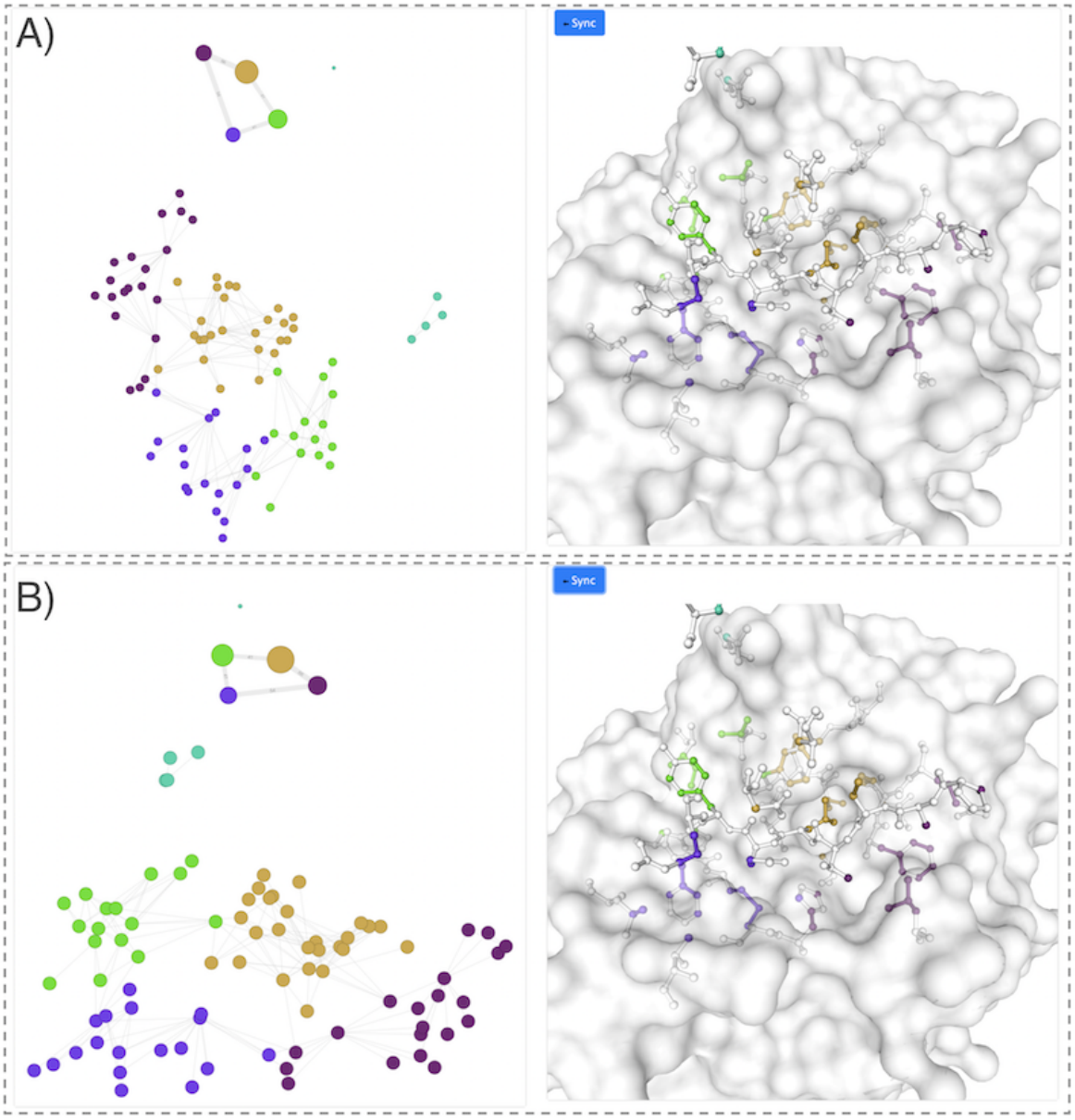


Figura 5.27: Processo de sincronização dos grafos em relação a estrutura. A) Pré sincronização. B) Pós sincronização

Por meio do menu (*more options*) é possível também ter acesso a outros dois conjuntos de funcionalidades que se destacam no GAPIN, *SPOTS* e Alinhamentos.

Os grafos de alto nível que se localizam na parte superior esquerda também são utilizados para o cálculo de *SPOTS* os quais auxiliam na identificação grupos de resíduos que podem ter uma contribuição relevante para a ligação da energia livre e serem candidatos a novos alvos drogáveis. O modelo proposto pelo GAPIN mostra uma evolução na distribuição de grupos de uma biomolécula, ou seja, o quão resistente

é um nó em um grupo. Veja na figura (5.28) que quanto mais quente ou próximo do vermelho mostra que o grupo é mais resistente de quebrar na medida que aumenta-se a quantidade de grupos, e quando mais frio ou próximo do branco, mais fácil é de quebrar. Observe que o grupo 4 se mantêm a partir de 4 grupos até 7 grupos.

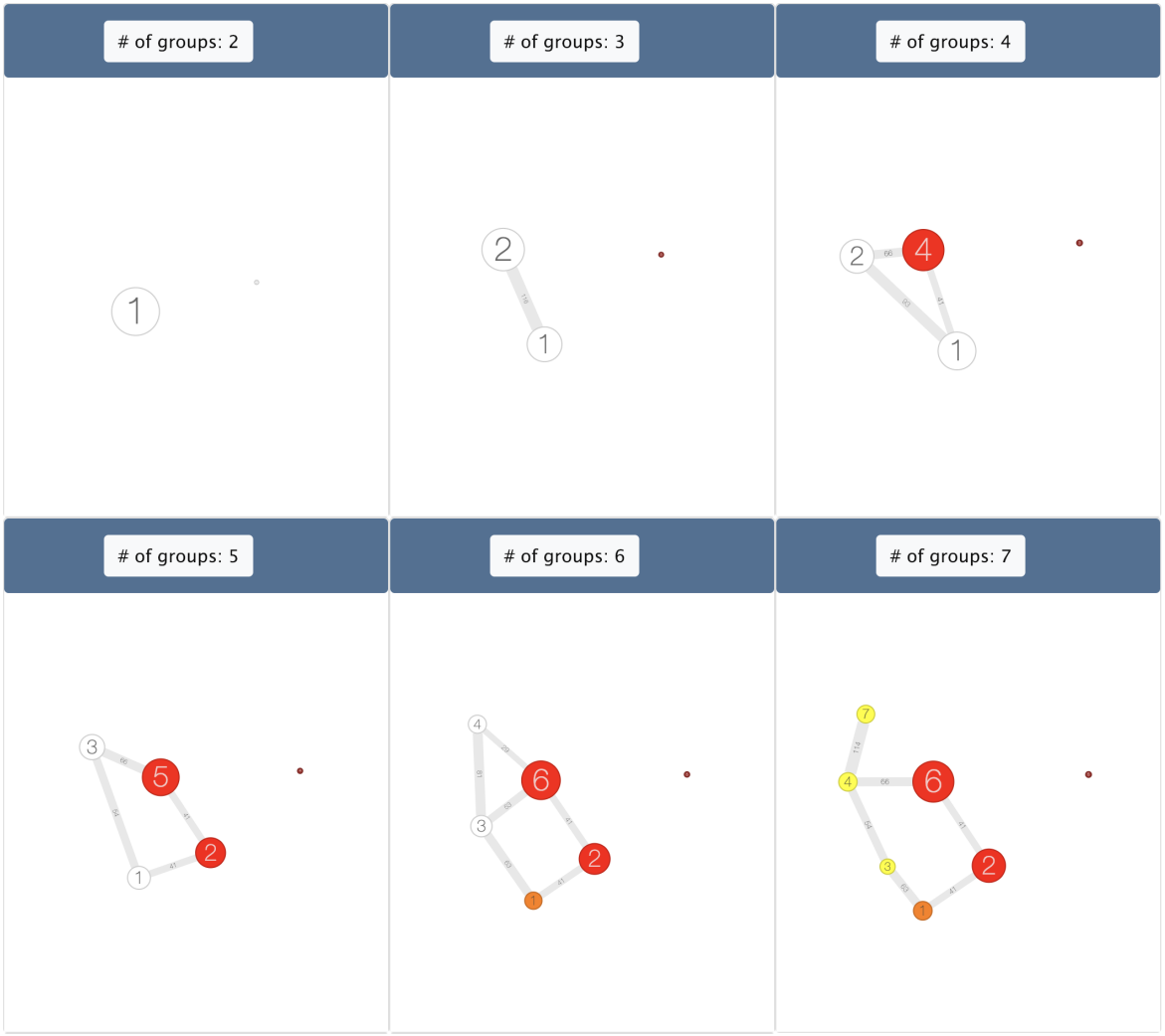


Figura 5.28: Visualização dos *SPOTS* ao longo do aumento da distribuição do número de grupos da 1PPF. Da esquerda pra direita de cima para baixo vemos a distribuição de grupos da 1PPF Apolar,Apolar Chain vs Chain começando 2 grupos até 7.

A outra opção disponível no menu (*more options*) como já dito anteriormente é a de alinhamentos. Essa opção habilita o usuário a realizar alinhamentos entre as estruturas disponíveis no GAPIN. Uma vez calculadas as estruturas, as mesmas ficam disponíveis para qualquer outro usuário do sistema (figura 5.29).

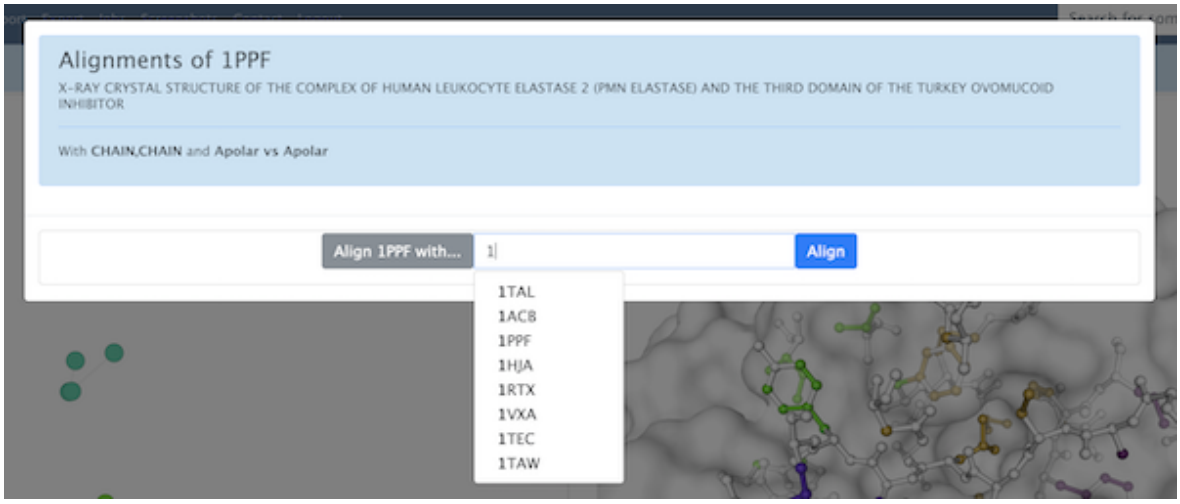


Figura 5.29: Tela de seleção de alinhamento com outra biomolécula

Uma vez calculado os alinhamentos, o usuário é redirecionado para uma tela a qual pode ver um padrão de para cada alinhamento da estrutura chamada de estrutura pai (figura 5.30). Nesse caso, a 1PPF em azul claro é a origem do alinhamento, seguido de 5 possibilidade das outra estrutura, nesse caso a 1TEC. Note que para esse alinhamento foi identificado outro conjunto seguindo o mesmo padrão da 1PPF como origem na terceira linha primeira coluna e posteriormente mais 5 sugestões de alinhamentos para a 1TEC. O que define a ordem de aparição é a qualidade do alinhamento, quanto mais próximo de 1 melhor.

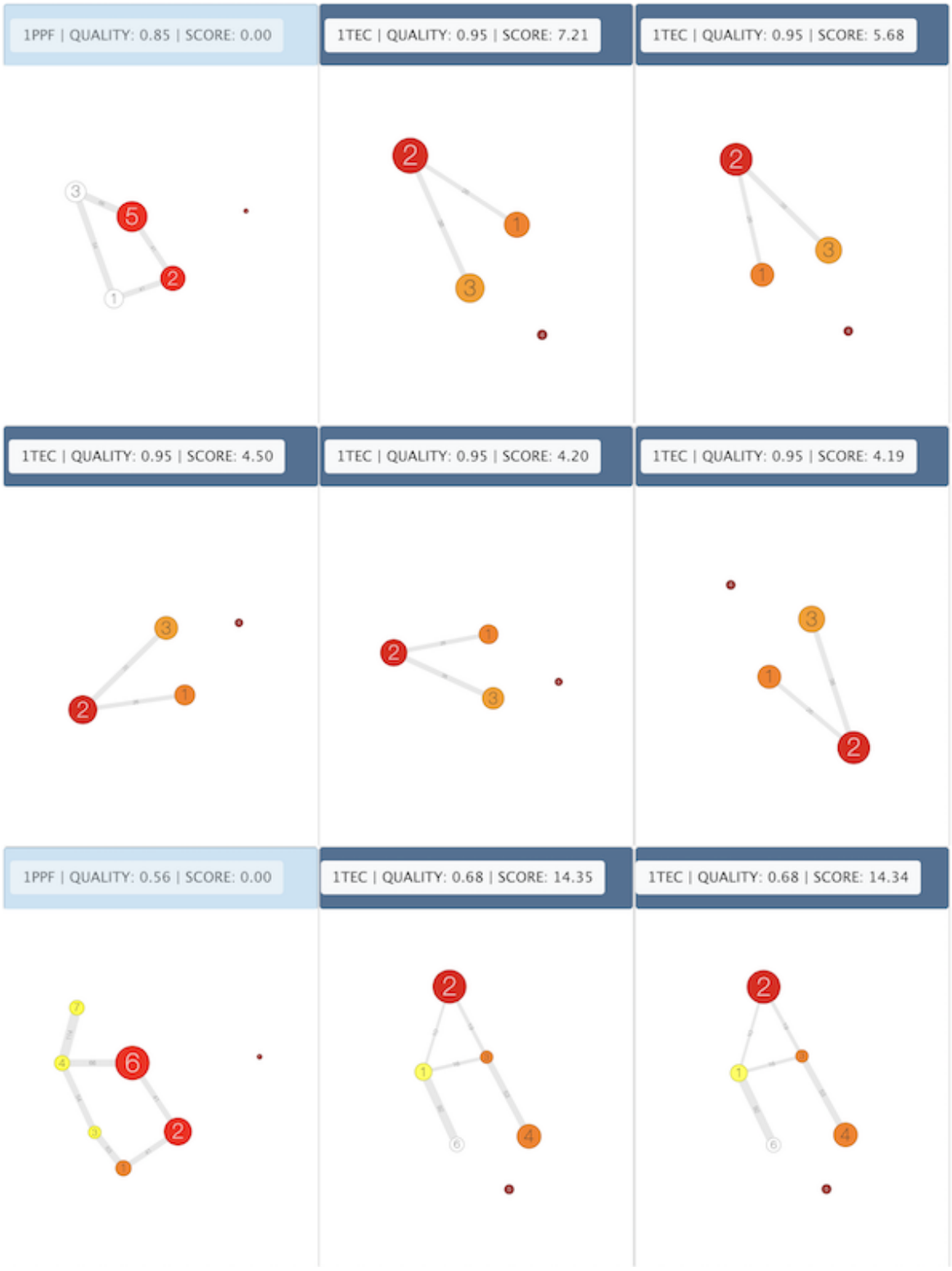


Figura 5.30: Tela de alinhamentos da 1PPF com 1TEC.

Caso o usuário queira ver como qualquer um desses alinhamentos se comporta, basta clicar sobre o botão na parte superior de cada caixa. Após essa ação, o usuário será redirecionado para a tela principal do sistema em que poderá realizar a análise das duas biomoléculas alinhadas (figura 5.31).

Figura 5.31: Tela de visualização dos alinhamentos selecionados pelo usuário entre a 1PPF e 1TEC

A fim de compartilhar com a comunidade científica, GAPIN também provê um mecanismo de exportação dos dados de interfaces calculadas por meio de um arquivo CSV. Nesse arquivo o usuário vai encontrar um formato de uma matriz de adjacência formada pelos contatos encontrados pela metodologia BARS.

5.2 Experimentos

5.2.1 Áreas de Contato

Superfícies desempenham um papel fundamental quando se trata de estudar as interações intermoleculares. Pode-se assumir que o grau (afinidade) de interação entre duas moléculas depende de alguma forma também da extensão e complementação das superfícies de contato entre elas. Quando maior a superfície de contato compartilhada, maiores as possibilidades de interações, ainda que essa relação possa ser não-linear e dependente de outros fatores físico-químicos (Lee & Richards [1971]).

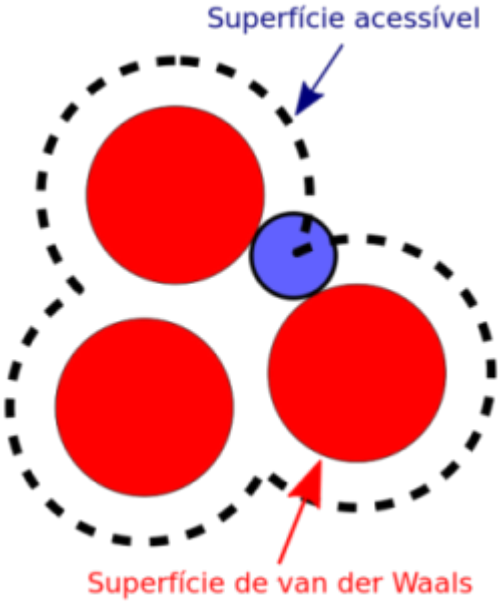


Figura 5.32: Representação 2D da ASA - *Accessible Surface Area* tal qual definidos pioneiramente por Lee & Richards (Lee & Richards [1971]). Uma sonda (*probe*) é rolada entorno dos átomos, cujo volume é definido pelos respectivos raios de van der Waals, computando uma unidade de área nas regiões de contato.(figura adaptada de Keith Callenberg, Wikipedia commons, CC BY-SA 3.0).

Mas, como melhor quantificar essas superfícies? No início da década de 1970, Lee & Richards propuseram um método que tornou-se clássico: ASA ou SASA - *Accessible Surface Area ou Solvent-Accessible Surface Area* (Lee & Richards [1971]). Propunha medir a superfície atômica acessível ao solvente através de uma sonda esférica (geralmente, uma molécula de água) que rolasse por toda a superfície de van der Waals dos átomos de uma proteína (ou qualquer outra biomolécula). Onde a sonda tocasse uma parte de um átomo em seu rolamento seria computado uma unidade de área, consolidada ao final por átomo, resíduo, cadeia etc. Afinal, se está acessível ao solvente, está (em tese) disponível para interações com qualquer outra biomolécula (figura 5.32).

Paralelamente, os avanços nas técnicas de cristalografia de raios X disponibilizavam um número crescente de biomoléculas com estruturas 3D resolvidas (Bernstein et al. [1977]). Por essa época também, começaram aparecer estudos teóricos e empíricos que encararam o desafio de tentar extrair parâmetros termodinâmicos e físico-químicos com base somente nas informações das estruturas resolvidas (Murphy & Freire [1992]). Por exemplo, foi possível estabelecer correlações lineares estatisticamente significativas entre ASA e variação da capacidade calorífica a pressão constante (ΔC_p), variação de entalpia (ΔH), variação de entropia (ΔS) e variação de energia livre de Gibbs[1] (ΔG)

[1]Doravante, sempre que nos referirmos à energia livre, ela será de Gibbs

Parâmetro	**Experimental**	**Computacional**
ΔC_p $(kJ\ K^{-1}\ mol^{-1})$	-1.1 ± 0.1	-1.4
ΔH° $(kJ\ mol^{-1})$	-2.5 ± 1.0	2.3
ΔS° $(J\ K^{-1}\ mol^{-1})$	195 ± 4.0	190
ΔG° $(kJ\ mol^{-1})$	-60 ± 0.5	-54

Tabela 5.1: Comparação de parâmetros termodinâmicos para Elastase Pancreática de Porco com o Inibidor Ovomucoide de Peru. Parte experimental feita com ITC - *Isothermical Titration Calorimety*, 25° C. Parte computacional feita a partir de ASA polar e apolar da estrutura PDBid 3EST superimposta à 1PPF. Foi feito assim porque não havia estrutura resolvida do complexo da Elastase de Porco com Inibidor Ovomucoide. Para detalhes metodológicos vide (Baker & Murphy [1997]).

para fenômenos como: solubilidade de compostos em água (Hermann [1972]), interações proteína-proteína (Janin & Chothia [1990]), hidrofobicidade de aminoácidos (Chothia [1974]), enovelamento de proteínas (Baldwin [1986]), interações proteína-ligantes (Eisenhaber [1999]). Estimou-se valores de energia livre entre 10 a 30 (com média de 18) $cal\ mol^{-1}$ para cada $Å^2$ de carbono hidrofóbico exposto ao solvente (Eisenhaber [1996]).

Baker & Murphy (1997) resolveram testar a confiabilidade dessas estimativas de parâmetros termodinâmicos baseados em ASA, com dados experimentais calorimétricos envolvendo a ligação (*binding*) da Elastase Pancreática de Porco com o Inibidor Ovomucoide de Peru (Baker & Murphy [1997]). Resultados comparativos podem ser vistos na tabela 5.1. Percebe-se que os resultados computacionais com ASA aproximaram-se bem dos resultados experimentais para a maior parte dos parâmetros, exceção talvez ao ΔH. Mas, como sua contribuição ao ΔG final é baixa ($< 5\%$), tal discrepância tem pouco impacto na energética consolidada do *binding*. Permite também concluir que a interação entre essa enzima e seu inibidor é entropicamente dirigida, tendo na entropia de solvatação o fator dominante. Isso seria o esperado se ambas cadeias comportassem-se como corpos rígidos, num modelo chave-fechadura, sem grandes mudanças conformacionais das cadeias durante o processo de *binding*. Baker & Murphy (1997) reforçam essa conclusão, ressaltando a confiabilidade do uso de ASA para inferir parâmetros termodinâmicos somente para complexos que se agregaram como corpos rígidos, sem mudanças conformacionais induzidas ou alosterismos (Baker & Murphy [1997]).

Conforme já dito antes, GAPIN faz uso de uma metodologia diferente do ASA para inferir áreas envolvidas em interações, denominada BARS. Ela computa de forma analítica (por uma equação) a área de contato não acessível ao solvente (Ac) dado dois átomos isolados. Traz como vantagem não somente uma maior performance no cálculo das áreas, mas também a possibilidade de definir os átomos participantes e sua

conectividade na construção das redes de contatos, oferecendo uma análise mais rica e detalhada na forma como as interfaces intermoleculares se organizam e se estruturam (vide Capítulo III - Materiais e Métodos).

Uma das consequência da metodologia BARS é que *Ac* será zero sempre que permitir a interveniência de uma ou mais moléculas de água entre dois átomos. Isso pode ser interpretado como produção de uma potencial cavidade, desconfigurando a possibilidade de contato direto entre tais átomos. Por outro lado, se a distância entre eles encurta de tal forma a não mais permitir essa interveniência, pode-se interpretar como uma expulsão de águas ou dessolvatação, com efeitos entrópicos positivos[2] na termodinâmica do sistema.

Logo, poderia também a metodologia BARS ter correlação com ASA em complexos intermoleculares? Poderia ser usada para estimar parâmetros termodinâmicos de tais complexos, tal como ASA? Essas foram duas importantes perguntas que floresceram no decorrer desta tese.

Para responder a primeira pergunta, foi montada uma base de dados local a partir do *Affinity Database 2.0* (Kastritis et al. [2011]), que é formado por um conjunto não-redundante de cerca de 180 complexos proteína-proteína, heterogêneo em funções, compreendendo: proteínas-G e receptores extracelulares, antígenos-anticorpos, enzimas-inibidores e enzimas-substratos. Inclui também, quando disponíveis, anotações físico-químicas e termodinâmicas, tais como: temperatura, pH, constantes de dissociação (k_d), ΔG de ligação, I-RMSD e outros.

Merece destaque o I-RMSD - *Interface Root Mean Square Deviation*, definido como a distância RMSD entre os carbonos alfas das estruturas isoladas e em complexo. RMSD[3] é uma medida do quão sobrepostos estão um conjunto de pontos coordenados; quanto mais sobrepostos, menor o RMSD. Assim, o I-RMSD é sensível em detectar se a formação dos complexos se deu como corpos rígidos ou se houve rearranjos conformacionais. Baixos I-RMSD indicam poucos rearranjos (Hwang et al. [2010]).

A base local do *Affinity Database 2.0* compreendeu 68 PDBids, consequência da filtragem dos complexos com I-RMSD$<$ 1.5Å, valor sugerido em (Kastritis et al. [2011]) para excluir grandes mudanças conformacionais. Sobre essa base calculou-se o ΔASA total[4] das interfaces dos complexos bem como as BARS áreas, o que gerou a correlação vista na figura (5.33). As definições de apolar e polar foram as do GAPIN. Os valores de correlação de Person (paramétrica) e Spearman (não-paramétrica) foram, respectivamente: apolar x apolar (0.79,0.79), polar x polar (0.76,0.74), todos x todos (0.83,

[2]Se os átomos envolvidos forem apolares.

[3]$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}{n}}$

[4]A partir do freeASA (Mitternacht [2016]) presente no pacote vanddraabe (Esposito [2017]) do R

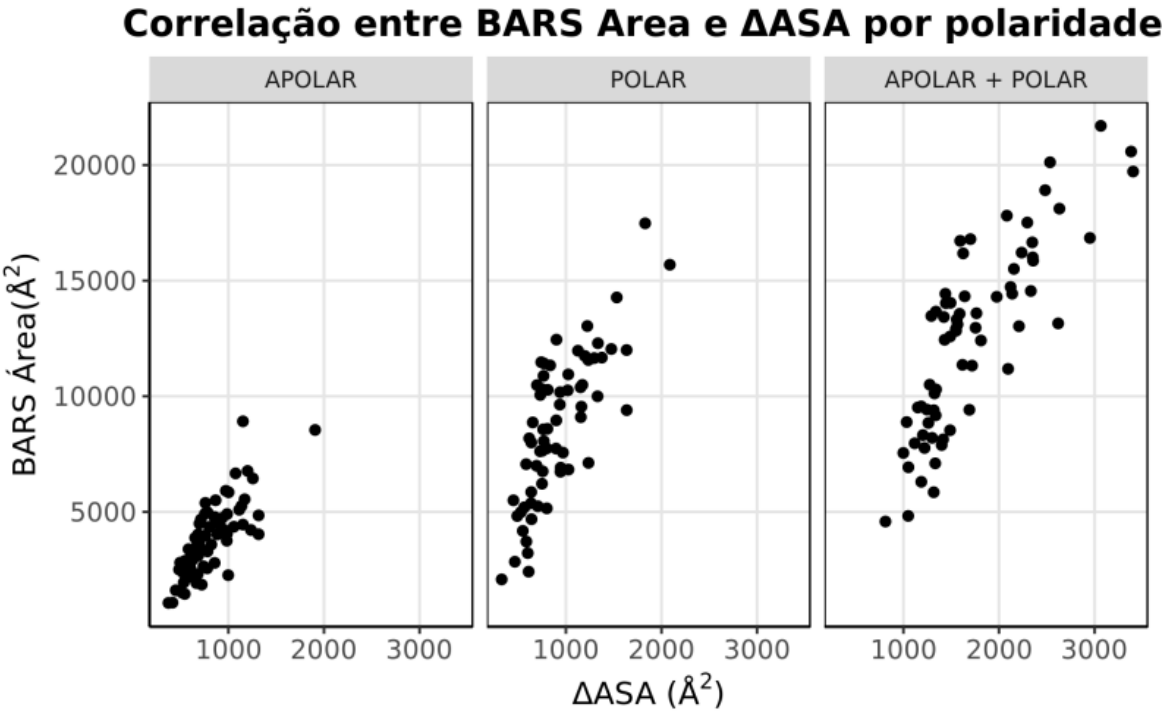


Figura 5.33: Gráficos de correlação BARS x ΔASA envolvendo 68 complexos filtrados do *Affinity Database 2.0* com I-RMSD< 1.5.

0.84). Percebe-se visualmente e pelos valores que parece haver uma boa correlação entre ΔASA e as BARS áreas, para as áreas totais das interfaces cadeia-cadeias, sendo um pouco menor nas interações polares. Isso indica que os valores de áreas aferidos pela metodologia BARS podem ter implicações termodinâmicas.

A segunda pergunta não teve como ser respondida nesta tese, e integrará estudos futuros a serem feitos como mais um de seus desdobramentos. Sua resposta exige um planejamento mais rigoroso dos experimentos de modo a compor estatísticas confiáveis que não enviesem as conclusões. Mas, os resultados das correlações indicam que pode ser possível usar BARS para estimar parâmetros termodinâmicos envolvidos na formação dos complexos intermoleculares.

5.2.2 Spots

Um segundo experimento conduzido nesta tese envolveu uma análise dos *Spots*, os nós dos grafos de alto nível. Foi usado o nome *Spots* por ser um termo consagrado na literatura para eventuais regiões drogáveis das interfaces cadeia-cadeia (Clackson & Wells [1995]). Bons *Spots* seriam aqueles mais energeticamente evidentes nessas interfaces.

Mutações pontuais trocando determinados resíduos por alanina (Morrison & Weiss [2001]) têm sido o procedimento padrão para discriminar aqueles que possam concentrar a maior parte da energia livre de ligação ou *binding*(Xia et al. [2010]). Trocar um resíduo com cadeia lateral maior por algum outro com cadeia lateral menor, minimiza os efeitos físico-químicos e topológicos do resíduo alvo que possam estar contribuindo para a energética do *binding*. Quanto maior a contribuição, maior será a diferença energética da mutação. Dentre os resíduos com menores cadeias laterais estão a alanina e glicina. A escolha da alanina advém do fato de sua troca eliminar a maior parte da cadeia lateral (preservando apenas o carbono beta) sem conferir ao ponto mutacionado uma flexibilidade de torção excessiva, como seria o caso se a mutação fosse por glicina. Tal técnica também é chamada de varredura por alaninas ou *alanine scanning* (Morrison & Weiss [2001]).

A literatura classifica os tipos de *Spots* conforme o impacto energético da substituição por alanina. Se a alteração provocar um $\Delta\Delta G$ de ligação de pelo menos 2.0 $kcal.mol^{-1}$, então o resíduo alvo é chamado de *Hot Spot*. Se menos que isso, um *Warm Spot*. Se maior que 4.0 $kcal.mol^{-1}$, um *Red-Hot Spot* (Li et al. [2004]).

Quando os primeiros grafos de alto nível com seus nós constituídos por agrupamentos densos de contatos atômicos surgiram, inevitavelmente veio a pergunta: teriam os grupos alguma relação com os *Spots*?

Para responder a essa pergunta, foram utilizados os dados de treinamento que embasaram o preditor de *Spots* chamado APIS - *A combined model based on Protrusion Index and Solvent accessibility* (Xia et al. [2010]), compreendendo 15 complexos proteína-proteína, contendo 62 resíduos *Hot Spots* e 92 *non-Hot Spots*. Esses dados estavam disponíveis no material suplementar[5] do referido artigo. Usou-se a base de treinamento por haver convencimento de que os autores tiveram o cuidado de deixá-la minimamente enviesada.

Mapiou-se a presença dos *Spots* (resíduos mutacionados) de acordo com sua localização nos nós dos grafos de alto nível, de modo a aferir se haveria correlação entre essa presença e o volume[6] das arestas que dão tamanho aos nós Para isso, como o número de partições (k) pode variar em função da varredura de agrupamentos (*cluster scanning*), precisou-se encontrar uma forma de fixar k para cada um dos 15 complexos. O critério adotado foi usar um k que implicasse em uma qualidade de agrupamento ($Q\%$) mínima de 85% (vide Capítulo III - Materiais e Métodos). Esse valor limite de $Q\%$ também foi definido de forma empírica. Estudar a influência da sua variação ficou

[5] 12859_2009_3631_MOESM1_ESM.DOC *Additional file 1*

[6] Lembrando que esse volume é o somatório das áreas de contatos Ac contidas nas arestas internas aos nós.

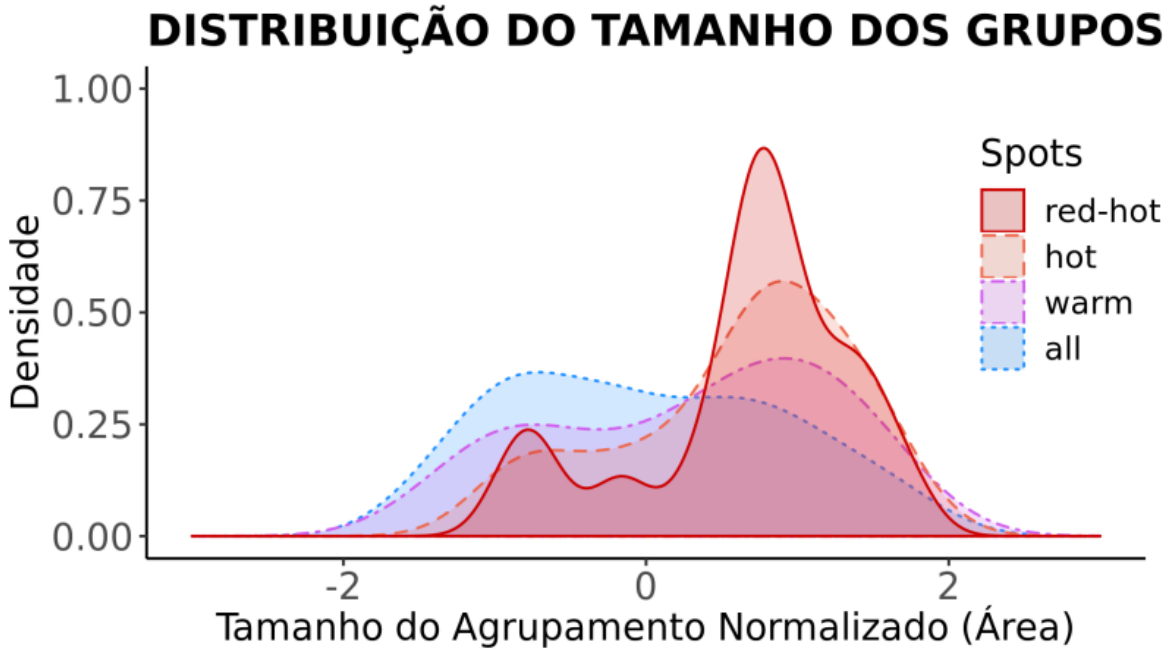


Figura 5.34: Gráfico da distribuição do tamanho dos grupos envolvendo 15 complexos da base de treinamento do programa de predição de *Spots* APIS (Xia et al. [2010]). O tamanho do agrupamento foi normalizado por *z-score* para cada complexo. Constata-se que quanto mais energético é um *Spot*, maior é o tamanho relativo dos grupos em que ele está.

para os trabalhos futuros a essa tese. Mas, 85% parece um nível aceitável de qualidade para os agrupamentos, numa escala de 0 a 100.

Cabe destacar mais alguns detalhes. Primeiro, para fazer com que os tamanhos dos grupos fossem comparáveis entre si, foi necessário normalizá-los para cada um dos 15 complexos selecionados. Para isso foi aplicado um *z-score* aos *Ac* originais, de modo que, assumindo uma distribuição próxima da normalidade para os tamanhos[7] a média dos tamanhos dos nós para cada complexo fosse ajustada para zero e os desvios padrões para um. Logo, quando mais positivo for o *z-score*, maior o tamanho relativo do nó; quanto mais negativo, menor o tamanho. Segundo, como GAPIN trabalha no nível atômico e os *Spots* são definidos no nível de resíduos, uma mutação por alanina pode afetar mais de um nó do grafo de alto nível, dado que a partição pode separar átomos de um mesmo resíduo em diferentes nós. Nessa condição, pode-se interpretar esse resíduo como tendo influência em diferentes agrupamentos; ou ainda, que ele se apresenta como um resíduo de maior centralidade do ponto de vista da rede de contatos, comportando-se como *link*, ponte ou transição entre diferentes agrupamentos, e que sua ausência

[7]Desvios dessa condição poderiam enviesar a normalização, mas como seria um efeito compartilhado por todos os complexos, é razoável pressupor que a conclusão geral seria pouco afetada. Fará parte de estudos futuros uma análise comparativa com outras métricas de normalização.

Teste	Tipo	Red Hot	Hot	Warm Hot
Kolmogorov-Smirnov	não-paramétrico	0.012	0.019	0.32
T de Student	paramétrico	0.0033	0.0021	0.11
Mann–Whitney–Wilcoxon	não-paramétrico	0.0066	0.0038	0.10

Tabela 5.2: Testes estatísticos para as distribuições de tamanho dos grupos (*p-values*) referentes aos dados da figura (5.34)

poderia provocar maior segmentação da interface. Foi deixado para trabalhos futuros uma avaliação mais profunda dos parâmetros topológicos das redes de contatos, o inclui o papel da centralidade dos nós.

O resultado dessa análise pode ser visto na figura (5.34). Vê-se um claro viés no sentido de que quanto mais energético é um *Spot*, relativamente maior ele é. As análises estatísticas confirmam essa primeira impressão visual. A tabela (5.2) apresenta 3 testes estatísticos distintos que estimam a probabilidade dos dados (ou parâmetros) amostrais decorrerem de uma mesma distribuição comum. Todas sugerem um *p-value* estatisticamente significativo (no nível de confiança de 0.95) para a hipótese alternativa de diferenciação das distribuições para *Red-Hot Spots* e *Hot Spots*, mas não para *Warm Spots*, em relação à distribuição geral dos tamanhos. Isso indica que grupos com maiores volumes de *Ac* em suas arestas internas têm maiores chances de abrigar *Spots* mais energéticos que os menores.

Trata-se de um resultado surpreendente e de certa maneira inesperado. Uma explicação possível é que, de alguma forma, o agrupamento espectral baseado na minimização do *Ncut* está capturando subredes mais intrincadas e densas em áreas de contatos. Isso estaria em linha com o que defende alguns autores para as propriedades das regiões *Hot Spots*, como aquelas com maior complementaridade topológica e densidade de interações (Moreira et al. [2007], Xia et al. [2010]), ainda que os bolsões de ligação (*binding pockets*) em PPIs tendam a ser alvos drogáveis relativamente rasos e pequenos, embora em alguns casos possam ser mais alongados, como nos epitopos em complexos antígeno-anticorpo (Arkin et al. [2014]).

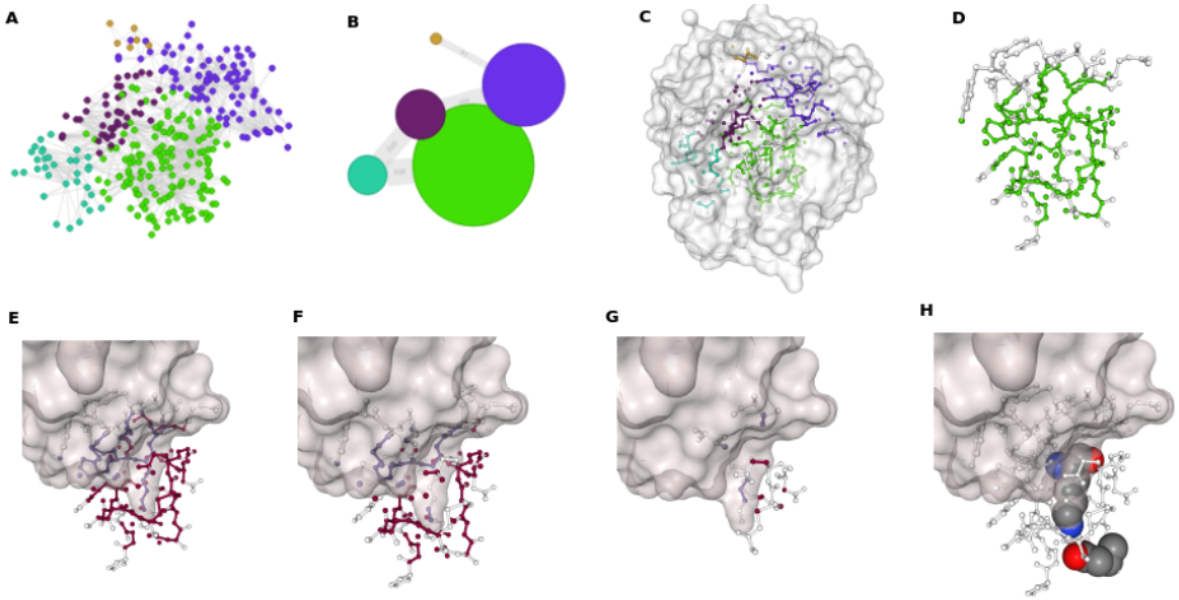


Figura 5.35: Análise de *Spots* da 2PTC, uma Tripsina Bovina em complexo com seu principal inibidor BPTI - *Bovine Pancreatic Trypsin Inhibitor*. (A) Grafo de baixo nível com 5 partições, coloridas de forma a discriminá-las, compreendendo todos os tipos de contatos (apolares e polares). (B) Grafo de alto nível. O alvo da mutação na cadeia BPTI (LYS15) pela alanina está no maior grupo, em verde claro. (C) Grafo de baixo nível no contexto geral da tripsina em $surface$ Connolly. (D) No grafo de alto nível, se o usuário clicar no nó verde claro , esse grupo pode ser isolado dos demais. (E) Grupo destacado colorido por cadeia (Tripsina em vermelho, BPTI em azul) e com representação de $surface$ do inibidor. É possível ver melhor agora que este grupo capturou centro ativo da tripsina, o que inclui o bolsão de especificidade. (F) Foco nos átomos com interações polares. (G) Foco em átomos com interações apolares. Nota-se que o bolsão é essencialmente polar. **(H)** Dois importantes resíduos da interface formando uma ponte salina: ASP189 (Tripsina) e LYS15 (BTPI), destacando a propensão das tripsinas em receber resíduos carregados positivamente no bolsão catalítico.

Seja como for, esse resultado põe GAPIN numa condição favorável à pesquisa de novos fármacos em PPIs pela análise cuidadosa dos *Spots*, especialmente os de maior tamanho. Um exemplo pode ser visto na figura (5.35), para o caso da 2PTC - Beta Tripsina de boi em complexo com seu mais clássico inibidor chamado BPTI - *Bovine Pancreatic Trypsin Inhibitor*, considerando todos os tipos de interações (apolares e polares). Da base de 15 complexos utilizados da base de treinamento APIS, 2PTC é a que apresenta maior $\Delta\Delta$G de ligação ($10 kcal.mol^{-1}$). Vê-se que o maior agrupamento, em verde claro, capturou as interações do centro ativo, mais especificamente do bolsão que dá maior seletividade à tripsina para clivar ligações peptídicas após resíduos carregados positivamente (em pH neutro), como lisinas e argininas. Com GAPIN, é possível isolar esse agrupamento maior dos demais, verificar que ele é dominado mais

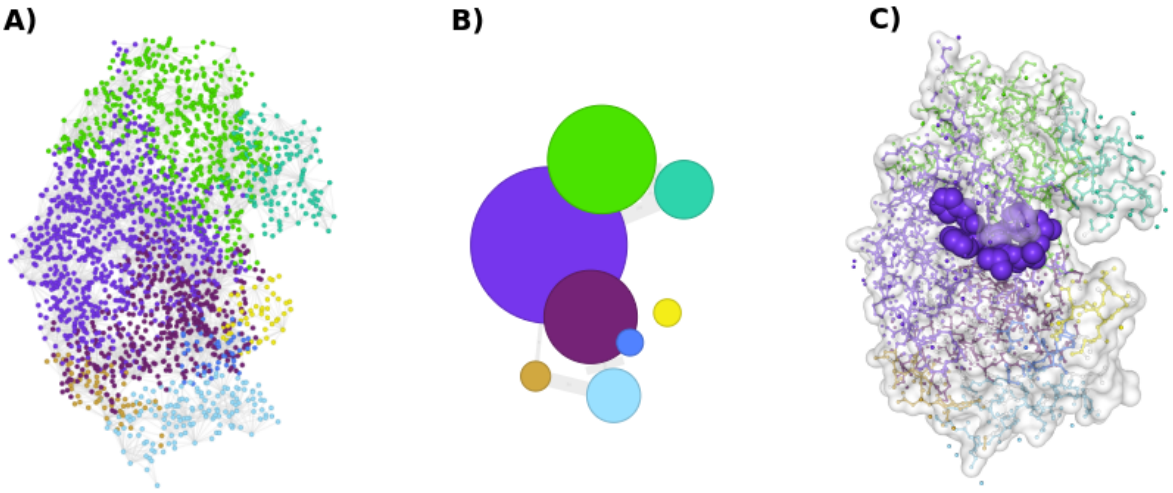


Figura 5.36: Complexo cadeia-ligante (ANY-LIG) do PDBid 3Q70 - Aspártico Protease do vírus da AIDS (HIV) e o inibidor péptido-mimético Ritanovir, particionado em 8 grupos. A) Grafo de baixo nível colorido conforme grupos. B) Grafo de alto nível, com o nó de maior tamanho em roxo. C) Estrutura sincronizada aos grafos, com a protease renderizada em *surface* e *ball+stick*, estando o inibidor Ritanovir em *spacefill*. O grupo ao qual o Ritanovir pertence é o maior (em roxo).

por interações polares que apolares, dar destaque a certos átomos ou resíduos, como a ponte salina entre a LYS15 do BPTI e o ASP189 da tripsina, o principal par iônico responsável pela referida especificidade dessa enzima. Um pesquisador poderia usar todas essas informações para lhe auxiliar na prospecção ou desenho de candidatos à fármacos inibidores de tripsina ou enzimas correlatas.

Enzimas são, de longe, o principal alvo terapêutico em estudos de triagem virtual (*virtual screening*) de candidatos a fármacos, sendo as proteases a segunda subclasse mais usada, atrás das quinases (Ripphausen et al. [2010]). Maior parte dos fármacos que atuam em proteases irá intervir diretamente no sítio ativo, inibindo ou modulando sua atividade. Muitos dos antivirais em uso no mercado são inibidores de proteases (K Patick & E Potts [1998]), a exemplo do Ritonavir (Randolph & DeGoey [2005]), presente nos coquetéis anti-HIV no tratamento da AIDS. Na figura (5.36) é possível ver o Ritanovir inibindo uma Aspártico Protease do HIV. Novamente, o maior grupo em roxo capturou o centro ativo, onde encontra-se ancorado o ligante Ritanovir, destacado em *spacefill*. Este exemplo do Ritanovir nos faz perguntar se a correlação verificada nas interfaces cadeia-cadeia entre tamanho do grupo e *Spots* energéticos também não se mostraria válida nas interfaces cadeia-ligante. Não houve como montar um experimento para responder essa pergunta, que também ficará para trabalhos futuros.

	ENZIMA			INIBIDOR		
PDB	**NOME ESPÉCIE**	***CLAN FAMILY***	***CLASS FOLD***	**NOME ESPÉCIE**	***CLAN FAMILY***	***CLASS FOLD***
1ACB	α-Chimotripsina *Bos taurus*	PA S1	*All-beta* *trypsin-like*	Eglina C *H. medicinalis*	IG I13	*Alpha and Beta* *CI-2 family*
1TEC	Thermitase *T. vulgaris*	SB S8	*Alpha and Beta* *subtilisin-like*	Eglina C *H. medicinalis*	IG I13	*Alpha and Beta* *CI-2 family*
1CSE	Subtilisina Carlsberg *B. subtilis*	SB S8	*Alpha and Beta* *subtilisin-like*	Eglina C *H. medicinalis*	IG I13	*Alpha and Beta* *CI-2 family*
1MEE	Mesentericopep. *B. mesentericus*	SB S8	*Alpha and Beta* *subtilisin-like*	Eglina C *H. medicinalis*	IG I13	*Alpha and Beta* *CI-2 family*
1SBN	Subtilisina BPN *B. amyloliquefaciens*	SB S8	*Alpha and Beta* *subtilisin-like*	Eglina C L45R *H. medicinalis*	IG I13	*Alpha and Beta* *CI-2 family*
1PPF	Elastase Leucócito *Homo sapiens*	PA S1	*All-beta* *trypsin-like*	Ovomucoide Peru *M. gallopavo*	IA I01	*Small Protein* *Kazal-type*
1CHO	α-Chimotripsina *Bos taurus*	PA S1	*All-beta* *trypsin-like*	Ovomucoide Peru *M. gallopavo*	IA I01	*Small Protein* *Kazal-type*
3SGB	SGT *S. griseus*	PA S1	*All-beta* *trypsin-like*	Ovomucoide Peru *M. gallopavo*	IA I01	*Small Protein* *Kazal-type*
1R0R	Subtilisina Carlsberg *B. subtilis*	SB S8	*Alpha and Beta* *subtilisin-like*	Ovomucoide Peru *M. gallopavo*	IA I01	*Small Protein* *Kazal-type*

Tabela 5.3: Complexos serino-peptidase e inibidores, conforme (Gonçalves-Almeida et al. [2011]). *Clan* refere-se às estruturas terciárias relacionadas e *family* às sequências relacionadas, da classificação de peptidases do MEROPS (Rawlings et al. [2018]). *Class* e *Fold* são classificações estruturais do SCOP (Murzin et al. [1995])).

5.2.3 Alinhamentos

Já vimos como GAPIN pode ser versátil nas análises de interfaces intermoleculares, tanto cadeia-cadeia quanto cadeia-ligantes, e como os agrupamentos que resultam nos grafos de alto nível podem oferecer informações valiosas sobre possíveis alvos terapêuticos para candidatos a fármacos. Mas, GAPIN conta ainda com um outro recurso igualmente valioso: a possibilidade de estudos comparativos, através do alinhamento dos grafos de alto nível nas interfaces cadeia-cadeia, usando diferentes PDBids.

Para exemplificar a potencialidade desse recurso, foi montado um experimento envolvendo alinhamentos cadeia-cadeia apolar de diferentes complexos serino-peptidase e respectivos inibidores. Em artigo publicado na *Bioinformatics* por nosso grupo de pesquisa (Gonçalves-Almeida et al. [2011]), foi apresentada a ferramenta *HydroPaCe - Hydrophobic Patch Centroids*. Com ela revelou-se evidências consistentes da existência de padrões hidrofóbicos nas interfaces entre serino-peptidases e seus inibidores. Mostrou-se também como tais padrões podem ajudar a explicar o fenômeno de inibição cruzada, quando enzimas e inibidores se comportam de forma promíscua, interagindo entre si. Exemplos clássicos são algumas serino-peptidases do tipo tripsina e tipo subtilisina. Não obstante apresentarem estruturas 3D divergentes (classes *all-beta* contra *alpha and beta*, segundo SCOP (Murzin et al. [1995])) e identidade de sequencia tão baixa quanto 20%, enzimas tipo-tripsina e tipo-subtilisina podem ser inibidas de forma

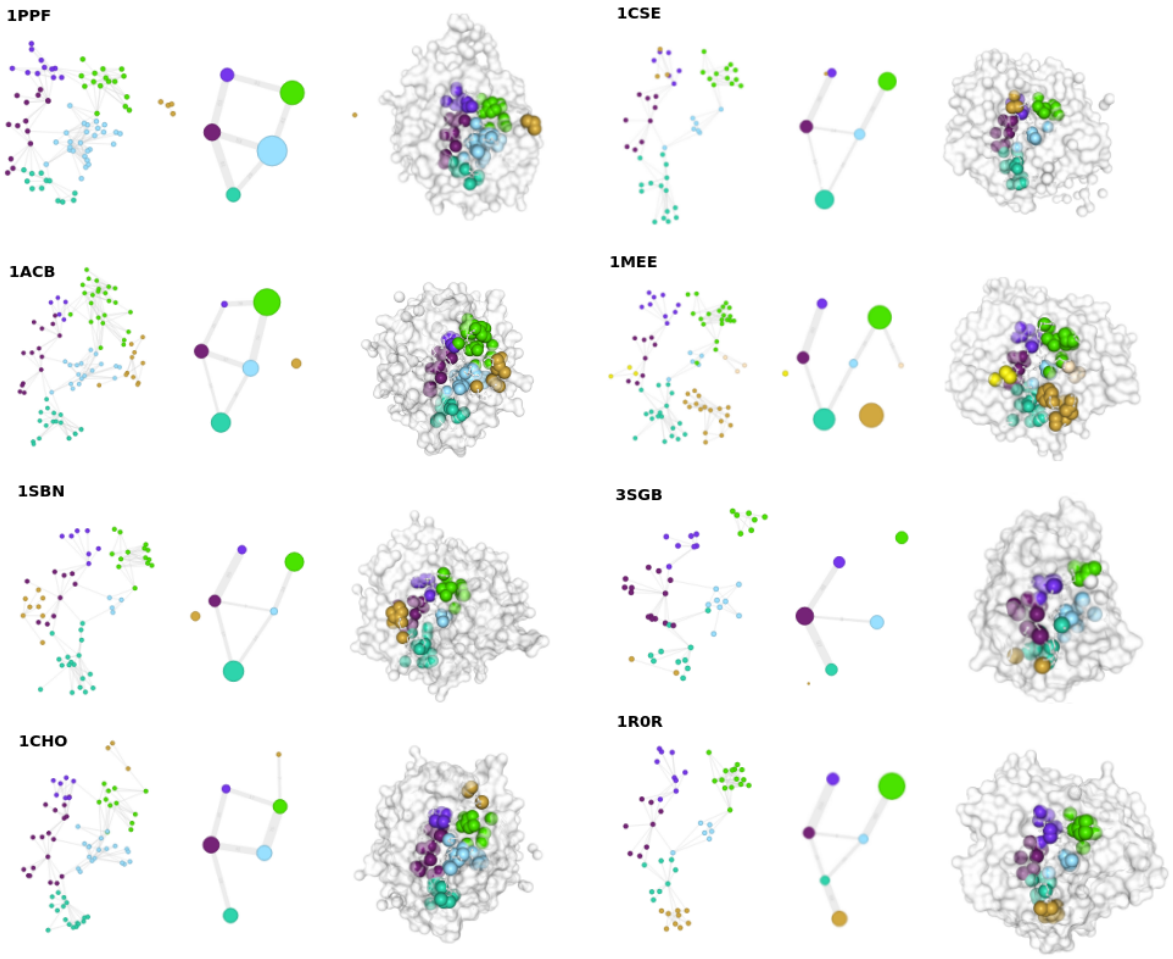


Figura 5.37: Alinhamento da 1PPF com 8 complexos da tabela (5.3). O alinhamento da 1PPF com 1TEC foi exibido na figura (4.7). O alinhamento foi feito pelo algoritmo *Topos*, com ajustes manuais para melhor visualização. Isso inclui também o ajuste manual de cores entre nós correspondentes, mas tudo feito através do GAPIN. Um editor de imagens foi usado apenas para compor a figura final.

cruzada pelos inibidores Ovomucoide de Peru ou Eglina C de Sanguessuga, ambos também com estruturas diferentes (classes *Small Protein* e *alpha and beta*, segundo SCOP) (Gonçalves-Almeida et al. [2011]). Nessas interfaces, os autores indicaram a existência de regiões hidrofóbicas conservadas e sobrepostas usando um conjunto não-redundante de 9 complexos, conforme tabela (5.3).

Os alinhamentos de grafos de alto nível podem ser feitos tanto manualmente quanto automaticamente (usando o algoritmo *Topos*) conforme detalhado no Capítulo IV - Materiais e Métodos, e exemplificado no alinhamento apolar da 1PPF (Elastase de Leucócito Humano e Inibidor Ovomucoide de Peru) com 1TEC (Subtilisina Termitase da bactéria *T. vulgaris* e Inibidor Eglina C de Sanguessuga), vide figura (4.7). A figura (5.37) expande esse alinhamento apolar cadeia-cadeia para os demais complexos

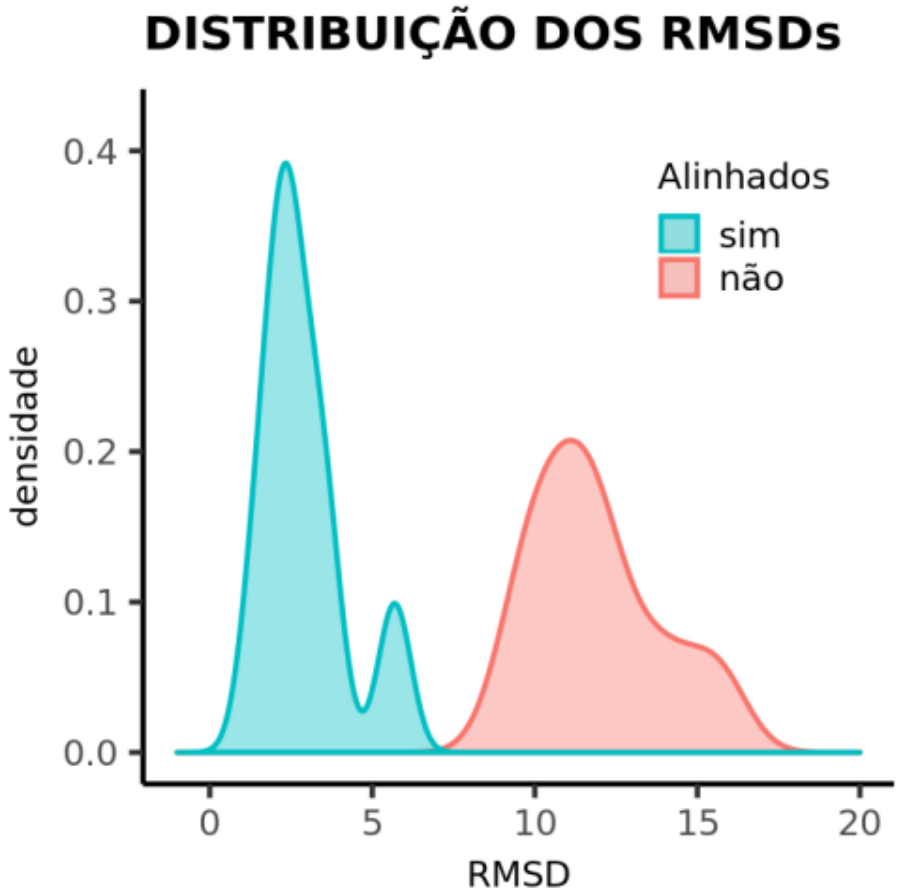


Figura 5.38: Gráfico da distribuição de RMSDs calculados sobre as coordenadas dos centróides dos grafos de alto nível da figura (5.37).

da tabela (5.3). Todos os alinhamentos neste experimento foram feitos de forma automática contra 1PPF, com ajustes manuais para melhor visualização. É possível ver as semelhanças dos grafos de alto nível e a sobreposição notável de nós, o que vai de encontro aos resultados do *HydroPaCe* em (Gonçalves-Almeida et al. [2011]), mas com o diferencial que GAPIN pode operar tais alinhamentos de forma genérica, para uma grande variedade de interfaces cadeia-cadeia do PDB.

A figura (5.38) mostra a distribuição RMSD como forma de avaliar numericamente a qualidade dos alinhamentos. Conforme visto para o I-RMSD, ele mede a raiz quadrada dos desvios médios quadráticos. No geral, quanto menor essa métrica, maior a sobreposição e melhor o alinhamento. O RMSD dessa figura foi calculado tendo como coordenadas os centroides dos nós dos grafos de alto nível. O RMSD médio para grafos alinhados foi de 2.9Å contra 11.9Å para não alinhados, uma redução que parece significativa para estruturas tão diferentes.

Pode-se ver que, a despeito das significativas diferenças estruturais tanto na parte

enzima quanto inibidor, um padrão hidrofóbico se mantém em todos os grafos da figura (5.37). E como é razoável pressupor que boa parte das interações protease-inibidor operam segundo um modelo chave-fechadura, tendo (conforme já discutido) na entropia de solvatação o efeito termodinâmico dominante, então a conservação hidrofóbica ajudaria a explicar as inibições cruzadas e a promiscuidade verificada em diferentes complexos desses tipos de enzimas (Waldner et al. [2018]). Seria uma espécie de assinatura apolar, que conferiria a todas essas peptidases um padrão de *binding* comum. Todas que compartilhassem a assinatura teriam condições termodinâmicas favoráveis à inibição ou ligação cruzada. Detalhes topológicos, interações polares e eletrostáticas entrariam na contabilidade da afinidade e eficiência geral da inibição/ligação (Waldner et al. [2018]). Proteases, a despeito das especificidades, têm que manter algum nível de promiscuidade com seus ligantes, dado que devem hidrolisar uma grande variedade de substratos[8]. Não por menos têm sido um desafio para a indústria farmacêutica encontrar inibidores de proteases com maior especificidade se o alvo terapêutico é o sítio catalítico (Drąg & Salvesen [2010]). Essa promiscuidade, se mais elevada em proteases, não parece ser exclusividade delas, como atestam os estudos de (Gao & Skolnick [2010]). Juntando tudo, é possível que a complementaridade hidrofóbica seja uma condição necessária para a formação estável de interfaces cadeia-cadeia, ainda que não suficiente (Waldner et al. [2018]). Essa foi a principal hipótese que levou ao trabalho do *HydroPace* (Gonçalves-Almeida et al. [2011]) e que foi discutida em maior profundidade em (Alves [2015]). Nesta tese, mostrou-se que GAPIN tem condições de aprofundá-la ainda mais, generalizando essa discussão para um amplo espectro de PPIs.

[8]Exceção, talvez, àquelas que atuam nas cascatas de sinalização, que tendem a ter um grau maior de especificidade (Paetzel et al. [2002])

Capítulo 6

Conclusões e Perspectivas

Apresentou-se aqui GAPIN - ***G****rouped and* ***A****ligned* ***P****rotein* ***I****nterface* ***N****etworks*, uma ferramenta 100% WEB focada na visualização e análise das interfaces intermoleculares em formato PDB. Para tanto, GAPIN oferece dois tipos básicos de abstrações visuais: de um lado, grafos representando as redes de contatos intermoleculares; de outro, estruturas atômicas renderizadas em 3D. Ambos lados mantém operações sincronizadas entre si, oferecendo ao usuário um duplo efeito visual que resulta num sinergismo de informações e operações muito maior do que seria possível com cada abstração isolada.

Conforme visto ao longo desta tese, GAPIN reúne uma série de características e inovações que, juntas, a tornam única frente outras ferramentas similares. Uma primeira a destacar é sua generalidade. GAPIN é capaz de oferecer um sistema de visualização e análise para quaisquer tipos de interfaces, tanto cadeia-cadeia quanto cadeia-ligante, sejam entre proteínas-proteínas (PPI), que detém maior quantidade de dados, sendo as mais bem estudadas na literatura, ou quaisquer combinações entre proteínas, ácidos nucleicos, carboidratos, lipídios, pequenas moléculas (ligantes), íons ou até mesmo água. Isso é possível porque GAPIN define interfaces em nível atômico, o que faz com que todas as interações intermoleculares sejam vistas como conjuntos de átomos e não de monômeros, facilitando a generalização para análises a qualquer tipo de biomolécula com estrutura resolvida no PDB.

Uma segunda, diz respeito à granularidade das visualizações. Inspirado nas ideias de um microscópio computacional (Dror et al. [2012]), GAPIN foi planejado para ser uma ferramenta multinível ou multiescala. Nesse sentido, trabalha com dois níveis de grafos, na representação das redes de contatos das interfaces: grafos de baixo nível (granularidade fina) e grafos de alto nível (granularidade grossa). O primeiro é montado tendo como vértices os átomos e as arestas ligam vértices (átomos) de cadeias diferentes

que estejam em contato, formando grafos bipartidos ou k-partidos. As arestas têm como pesos as áreas de contatos (Ac) tal que $Ac > Ac_{min}$. O segundo, nasce do agrupamento espectral feito sobre os grafos de baixo nível. Nesses grafos de alto nível, os vértices agora são grupos ou comunidades de átomos e respectivas arestas com $Ac > Ac_{min}$. Tal dupla granularidade oferece ao usuário de GAPIN ricas possibilidades de análises, que juntas com o sinergismo entre redes e estruturas renderizadas, permitem isolar e estudar regiões específicas das interfaces conforme interesse.

Foi mostrado que a maneira como se definiu áreas de contatos em GAPIN constituiu uma inovação com vários desdobramentos bem-vindos. O uso clássico na literatura quando se fala em áreas atômicas tem sido o ASA (ou SASA) - *Accessible Surface Area ou Solvent-Accessible Surface Area*, proposto por (Lee & Richards [1971]) e aprimorado por outros (Ali et al. [2014]). Embora ASA ofereça as áreas expostas ao solventes disponíveis para contatos, ela não mapeia quem faz contato com quem. GAPIN aprimorou a metodologia chamada BARS[1], criada e desenvolvida inicialmente no trabalho de (Alves [2015]). BARS calcula a área não-exposta ao solvente em dois átomos isolados de cadeias diferentes. E faz isso analiticamente com grande eficiência (em $O(1)$) através de uma equação (eq. 4.1). BARS permite ainda definir não somente quais são os átomos que pertencem às interfaces ($Ac > Ac_{min}$) como também diz que átomo tem área de contato com qual outro, construindo a rede de contatos. E como BARS computa a área não-acessível ao solvente, um $Ac = 0$ implica interveniência de uma ou mais moléculas de água. Do contrário, um $Ac > 0$ implica em expulsão de moléculas de água ou dessovatação, levando a metodologia BARS a ser passível de ter uma interpretação termodinâmica. Mostrou-se que BARS oferece tudo isso mantendo boa correlação com ASA (Person e Spearman entre 0.74 e 0.84, respectivamente), conforme indicado na figura (5.33) em experimento montado a partir de uma base (*Affinity Database 2.0*) de varredura de mutações por alanina (Kastritis et al. [2011]).

Outra inovação de GAPIN envolve a varredura de agrupamentos. A norma da literatura tem sido encontrar um número ótimo de grupos k, dada uma métrica como coeficiente de silhueta (Rousseeuw [1989]) ou modularidade (Newman [2004]). Nesta tese pretendeu-se deixar variar k, de modo que o usuário pudesse acompanhar possíveis efeitos e padrões nessa variação. Associado a isso, criou-se uma métrica de qualidade $Q\%$ que mede o quão resistente à partição é um vértice nos grafos de alto nível.

Durante o desenvolvimento e testes com essa variação de grupos, foi ficando cada vez mais evidente a possibilidade de correlação entre as partições e os *Spots*, termo consagrado na literatura para regiões nas interfaces cadeia-cadeia passíveis de se

[1] iniciais dos autores implicados na sua criação e desenvolvimento

apresentarem como alvos drográveis (Clackson & Wells [1995]). Quanto mais energético é o *Spots*, melhor alvo ele será para um projeto de pesquisa de novos fármacos em PPIs. A técnica mais usada para aferir essa energética é a varredura por mutações de alanina (Morrison & Weiss [2001]), verificando os efeitos sobre a variação da energia livre (ΔG) de *binding* ao trocar determinado resíduo por ela.

De forma inesperada, encontrou-se nesta tese uma correlação: quanto maior o tamanho de uma partição, maiores as chances dela abrigar um *Spot* energético. Tal resultado adveio de um experimento envolvendo dados de teste do preditor de *Spots* APIS (Xia et al. [2010]), contendo 15 complexos do PDB, com anotações sobre os $\Delta\Delta G$ de *binding* para 154 resíduos mutacionados para alanina. Ao cruzar esses dados com o tamanho dos grupos gerados por GAPIN, emergiu a intrigante figura (5.34). É visualmente claro a relação entre o tamanho e a categoria energética do *Spot*, reforçada por testes estatísticos da tabela (5.2). Parece que o agrupamento espectral conseguiu particionar as redes de contatos conforme a complexidade das regiões das interfaces, o que estaria alinhado a outras pesquisas da literatura sobre as propriedades dos *Hot Spots* (Moreira et al. [2007], Xia et al. [2010]). A novidade aqui foi constatar que as metodologias usadas em GAPIN conseguem discriminar e apresentar, de pronto, as regiões mais propensas a hospedar *Spots* qualificados como alvos terapêuticos, com uma grande variedade de opções de manipulações visuais e analises, tanto nos grafos quanto nas estruturas renderizadas, conforme indicado nas figuras (5.35) e (5.36).

Por fim, foi acrescentado em GAPIN um recurso considerado bem valioso e promissor: a possibilidade de alinhar grafos de alto nível para efeito de estudos comparativos entre diferentes interfaces intermoleculares. GAPIN permite que tal alinhamento seja feito tanto manual quanto automático. A forma automática faz uso de um algoritmo inovador chamado *Topos*. Para por em evidência a potencialidade deste recurso, foi montado um experimento para testar a efetividade dos alinhamentos. Para tanto, utilizou-se dados de outro trabalho publicado (Gonçalves-Almeida et al. [2011]) por nosso grupo de pesquisa em Bioinformática Estrutural, envolvendo 9 complexos diferentes entre serino-peptidases e inibidores. Tal base contem tanto proteases quanto inibidores de classes estruturalmente diferentes, segundo SCOP (Murzin et al. [1995]). A despeito dessas grandes diferenças nas estruturas terciárias, elas se apresentam promíscuas, ligando-se de forma cruzada em muitos casos. Como isso seria possível? No intuito de ir mais a fundo no entendimento desse fenômeno, foi feito um alinhamento dos respectivos grafos de alto nível, considerando apenas interações apolares, dos 9 complexos, tendo uma delas (1PPF) como referência. Os resultados foram apresentados nas figuras (4.7) e (5.37), onde foi possível perceber um claro padrão de alinhamento entre todos os grafos, com surpreendente sobreposições de nós. Tal padrão sugere a

existência de uma assinatura hidrofóbica permeando as interfaces cadeia-cadeia em todos os complexos, o que ajudaria a explicar a promiscuidade inerente a essa classe de enzimas (Waldner et al. [2018]).

Espera-se que neste trabalho tenha sido possível demonstrar a versatilidade visual e a potencialidade analítica da ferramenta GAPIN, para estudos de uma grande variedade de interfaces intermoleculares, com efetivo poder de auxiliar pesquisadores de diversas especialidades a entenderem melhor as propriedades topológicas e físico-químicas de potenciais alvos terapêuticos na busca por fármacos inovadores.

6.1 Perspectivas

Ao longo do texto desta tese, procurou-se ir destacando algumas limitações e desafios, que se revertem em possibilidades concretas de trabalhos futuros, em múltiplas linhas de pesquisa e desenvolvimento, do curto ao longo prazo, tais como:

- Extrair parâmetros topológicos da rede de contatos, tais como: *centralities*, *network diameter*, *clustering coefficient*, *shortest path lenghts* etc, tendo em mente que se trata de grafos bi ou k-partidos.

- Aprofundar estudos sobre a unificação entre matrizes de adjacências em grafos de alto nível e matrizes de confusão.

- Encontrar uma racionalização que tenha sentido físico-químico e termodinâmico para a escolha da área de contato mínima (Ac_{min}) entre dois átomos. No momento, isso está definido de forma empírica em 5Å^2.

- Investigar sobre a influência dos parâmetros $Dist$ e r de nós sobrepostos sobre os seus índices de preservação.

- Estudar a influência da variação da métrica de qualidade $Q\%$ e outras formas de normalização de tamanhos de nós sobre os resultados dos experimentos com *Spots*.

- Verificar se a correlação encontrada nas interfaces cadeia-cadeia entre o tamanho do grupo e *Spots* energéticos também não se mostraria válida nas interfaces cadeia-ligante.

- Estimar e validar parâmetros termodinâmicos tirados diretamente a partir da metodologia BARS.

- Estender alinhamentos das interfaces cadeia-cadeia para interfaces cadeia-ligante
- Promover estudo minucioso do algoritmo de alinhamento *Topos*, com uma avaliação detalhada dos seus pontos fortes e fracos, além de uma comparação numérica e estatística com outros métodos.
- Incorporação de redes de interações intermoleculares no nível interactoma, nos moldes do STRING (Szklarczyk et al. [2017]), Interactome3D (Mosca et al. [2013]) e STITCH (Szklarczyk et al. [2016]), mesmo quando não houver dados estruturais.
- Firmar-se como uma multiplataforma de visualização e análises integradas em bioinformática estrutural e genômica, mais próxima da óptica de um microscópio computacional.
- Permitir que o usuário descubra as perguntas que ainda não foram feitas, para que sejam feitas novas perguntas.

Referências Bibliográficas

Ali, S.; Hassan, M.; Islam, A. & Ahmad, F. (2014). A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Current Protein & Peptide Science*, 15(5):456--476.

Alves, N. R. (2015). Mapeamento de correspondências hidrofóbicas em complexos serino peptidases e inibidores proteicos através da varredura de agrupamento espectral. Tese (Doutorado em Bioinformática) - Universidade Federal de Minas Gerais (UFMG), Belo Horizonte.

Arkin, M. R.; Tang, Y. & Wells, J. A. (2014). Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality. *Chemistry and Biology*, 21(9):1102--1114.

Arkin, M. R. & Wells, J. A. (2004). Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature reviews Drug discovery*, 3(4):301.

Arrowsmith, J. (2012). A decade of change. *Nature reviews. Drug discovery*, 11(1):17--8.

Assenov, Y.; Ramírez, F.; Schelhorn, S. E. E.; Lengauer, T. & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282--284.

Baker, B. M. & Murphy, K. P. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *Journal of molecular biology*, 268(2):557--69.

Baldwin, R. L. (1986). Temperature Dependence of the Hydrophobic Interaction in Protein Folding. *Proceedings of the National Academy of Sciences*, 83:8069--8072.

Barabási, A.-L. (2007). Network medicine–from obesity to the "diseasome". *The New England journal of medicine*, 357(4):404--7.

Barabasi, A.-L. (2016). *Network Science*. Cambridge University Press.

Bell, G.; Gray, J. & Szalay, A. (2006). Petascale computational systems. *Computer*, 39(1):110--112.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235--242.

Bernard, M. C. (1865). *Introduction à l'étude de la médecine expérimentale*. J. B. BAILLIÈRE et FILS, Paris.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(2):535--542.

Bitbucket.org (2019a). Bitbucket. http://www.bitbucket.org. Acessado em:12.01.2019.

Bitbucket.org (2019b). Bitbucket Pipelines. http://bitbucket.org/product/features/pipelines. Acessado em:12.01.2019.

Brini, E.; Fennell, C. J.; Fernandez-Serra, M.; Hribar-Lee, B.; Lukšič, M. & Dill, K. A. (2017). How Water's Properties Are Encoded in Its Molecular Structure and Energies. *Chemical Reviews*, 117(19):12385--12414.

Cafarelli, T. M.; Desbuleux, A.; Wang, Y.; Choi, S. G.; De Ridder, D. & Vidal, M. (2017). Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, 44:201--210.

Casciaro, M. & Mammino, L. (2016). *Node. js Design Patterns*. Packt Publishing Ltd.

Chakrabarty, B. & Parekh, N. (2016). NAPS: Network analysis of protein structures. *Nucleic Acids Research*, 44(W1):W375--W382.

Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640--647.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248(5446):338--339.

Clackson, T. & Wells, J. A. (1995). A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science*, 267(January):383--386.

Connolly, M. L. (1983). Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548--558.

Council, N. R. et al. (2005). *Network science*. National Academies Press.

Cukuroglu, E.; Engin, H. B.; Gursoy, A. & Keskin, O. (2014). Hot spots in protein-protein interfaces: Towards drug discovery. *Progress in Biophysics and Molecular Biology*, 116(2-3):165--173.

d3plus (2019). d3plus. http://www.d3plus.org. Acessado em:12.01.2019.

da Silveira, C. H.; Pires, D. E.; Minardi, R. C.; Ribeiro, C.; Veloso, C. J.; Lopes, J. C.; Meira Jr, W.; Neshich, G.; Ramos, C. H.; Habesch, R. et al. (2009a). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):727--743.

da Silveira, C. H.; Pires, D. E. V.; Minardi, R. C.; Ribeiro, C.; Veloso, C. J. M.; Lopes, J. C. D.; Meira, W.; Neshich, G.; Ramos, C. H. I.; Habesch, R. & Santoro, M. M. (2009b). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, 74(3):727--43.

datatables (2019). Datatables. http://www.datatables.net. Acessado em: 12.01.2019.

DeLano, W. L. (2002). The pymol molecular graphics system. *http://www. pymol. org*.

Doncheva, N. T.; Assenov, Y.; Domingues, F. S. & Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4):670--685.

Doncheva, N. T.; Klein, K.; Domingues, F. S. & Albrecht, M. (2011). Analyzing and visualizing residue networks of protein structures. *Trends in Biochemical Sciences*, 36(4):179--182.

Drąg, M. & Salvesen, G. (2010). Emerging principles in protease - based drug. *Nature Reviews Drug Discovery*, 9(9):690--701.

Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H. & Shaw, D. E. (2012). Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics*, 41(1):429--452.

Duvall, P. M.; Matyas, S. & Glover, A. (2007). *Continuous integration: improving software quality and reducing risk*. Pearson Education.

Eisenhaber, F. (1996). Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures. *Protein Science*, 5(8):1676--1686.

Eisenhaber, F. (1999). Hydrophobic regions on protein surfaces. *Perspectives in Drug Discovery and Design*, 17:27--42.

Esposito, E. (2017). Vanddraabe: Identification and statistical analysis of structurally conserved waters via r. Em *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 253. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA.

expressjs.org (2019). Expressjs. http://expressjs.org. Acessado em:12.01.2019.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75--174.

Gallaba, K.; Mesbah, A. & Beschastnikh, I. (2015). Don't call us, we'll call you: Characterizing callbacks in javascript. Em *Empirical Software Engineering and Measurement (ESEM), 2015 ACM/IEEE International Symposium on*, pp. 1--10. IEEE.

Gao, M. & Skolnick, J. (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107(52):22517--22522.

Gatherer, D. (2010). So what do we really mean when we say that systems biology is holistic? *BMC systems biology*, 4(1):22.

Gonçalves-Almeida, V. M.; Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Meira, W. & Santoro, M. M. (2011). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342--349.

Green, S.; Şerban, M.; Scholl, R.; Jones, N.; Brigandt, I. & Bechtel, W. (2017). Network analyses in systems biology: new strategies for dealing with biological complexity. *Synthese*, pp. 1--27.

HandleBars (2019). HandleBars. http://www.handlebarsjs.com. Acessado em: 12.01.2019.

Hartman, G. D.; Egbertson, M. S.; Halczenko, W.; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D. & Lynch, R. J. (1992). Non-peptide

fibrinogen receptor antagonists. 1. discovery and design of exosite inhibitors. *Journal of medicinal chemistry*, 35(24):4640--4642.

Hecker, M.; Lambeck, S.; Toepfer, S.; van Someren, E. & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, 96(1):86--103.

Hermann, R. B. (1972). Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *Journal of Physical Chemistry*, 76(19):2754--2759.

Holmes, M. A.; Le Trong, I.; Turley, S.; Sieker, L. C. & Stenkamp, R. E. (1991). Structures of deoxy and oxy hemerythrin at 2.0 å resolution. *Journal of molecular biology*, 218(3):583--593.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682.

Hruska, J. (2016). Oracle is finally killing the Java browser plug-in. https://www.extremetech.com/internet/222121-oracle-is-finally-killing-the-java-browser-plug-in, acessado em:21.01.2019.

Humble, J. & Farley, D. (2011). *Continuous delivery: reliable software releases through build, test, and deployment automation*. Addison-Wesley Boston.

Huttlin, E. L.; Bruckner, R. J.; Paulo, J. A.; Cannon, J. R.; Ting, L.; Baltier, K.; Colby, G.; Gebreab, F.; Gygi, M. P.; Parzen, H.; Szpyt, J.; Tam, S.; Zarraga, G.; Pontano-Vaites, L.; Swarup, S.; White, A. E.; Schweppe, D. K.; Rad, R.; Erickson, B. K.; Obar, R. A.; Guruharsha, K. G.; Li, K.; Artavanis-Tsakonas, S.; Gygi, S. P. & Wade Harper, J. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505--509.

Hwang, H.; Vreven, T.; Janin, J. & Weng, Z. (2010). Protein-protein docking benchmark version 4.0. *Proteins: Structure, Function and Bioinformatics*, 78(15):3111--3114.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *Journal of Biological Chemistry*, 265(27):16027--16030.

Jeanquartier, F.; Jean-Quartier, C. & Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. *BMC Bioinformatics*, 16(1):1--16.

K Patick, A. & E Potts, K. (1998). Protease Inhibitors as Antiviral Agents. *Clinical Microbiology Reviews*, 11(4):614--627.

Kamburov, A.; Stelzl, U.; Lehrach, H. & Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Research*, 41(D1):793--800.

Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y. & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353--D361.

Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. a.; Bonvin, A. M. J. J. & Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Protein Science*, 20(3):482--491.

Kayikci, M.; Venkatakrishnan, A. J.; Scott-Brown, J.; Ravarani, C. N.; Flock, T. & Babu, M. M. (2018). Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas. *Nature Structural and Molecular Biology*, 25(2):185--194.

Kelly, P. (2019). Mechanics Lecture Notes Part III: Foundations of Continuum Mechanics. http://homepages.engineering.auckland.ac.nz/∼pkel015/SolidMechanicsBooks. Acessado em: 12.01.2019.

Kim, P. M.; Lu, L. J.; Xia, Y. & Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938--1941.

Kitano, H. (2002). Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560):1662--4.

Krissinel, E. (2011). Macromolecular complexes in crystals and solutions. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):376--385.

Krissinel, E. & Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, 372(3):774--797.

Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L. & Thornton, J. M. (1997). PDBsum: A Web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences*, 22(12):488--490.

Laskowski, R. A. & Swindells, M. B. (2011). Ligplot+: Multiple ligand-protein interaction diagrams for drug discovery. *Journal of chemical information and modeling*, 51 10:2778–86.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3).

Li, X.; Keskin, O.; Ma, B.; Nussinov, R. & Liang, J. (2004). Protein-protein interactions: Hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: Implications for docking. *Journal of Molecular Biology*, 344(3):781--795.

Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardozza, A. P.; Santonico, E.; Castagnoli, L. & Cesareni, G. (2012). MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 40(D1):857--861.

Marcum, J. A. (2008). Does systems biology represent a Kuhnian paradigm shift? *New Phytologist*, 179:587--589.

McGillivray, P.; Clarke, D.; Meyerson, W.; Zhang, J.; Lee, D.; Gu, M.; Kumar, S.; Zhou, H. & Gerstein, M. (2018). Network Analysis as a Grand Unifier in Biomedical Data Science. *Annual Review of Biomedical Data Science*, 1(1):153--180.

Mendeley (2019). Mendeley. https://www.mendeley.com. Acessado em:13.05.2019.

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, 5(189):1--10.

MongoDB (2019). MongoDB. http://www.mongodb.com. acessado em:12.01.2019.

Moreira, I. S.; Fernandes, P. A. & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68:803--812.

Morrison, K. L. & Weiss, G. A. (2001). Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*, 5:302--307.

Mosca, R.; Céol, A. & Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods*, 10(1):47--53.

MRC-LMB (2019). Medical Research Council (MRC) Laboratory of Molecular Biology (LMB). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Laboratory_of_Molecular_Biology. Acessado em:13.05.2019.

Murphy, K. P. & Freire, E. (1992). Thermodynamics of structural stability and cooperative folding behavior in proteins. *Advances in Protein Chemistry*, 43:313--361.

Murzin, A. G.; Brenner, S. E.; Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536--540.

Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B*, 38(2):321--330.

Nielsen, J. & Loranger, H. (2006). *Prioritizing Web Usability*. Voices That Matter. Pearson Education. ISBN 9780132798150.

Noble, D. (2010). Biophysics and systems biology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1914):1125--1139.

nodejs.org (2019). NodeJS. http://nodejs.org. Acessado em:12.01.2019.

O 'donoghue, S. I.; Baldi, B. F.; Clark, S. J.; Darling, A. E.; Hogan, J. M.; Kaur, S.; Maier-Hein, L.; Mccarthy, D. J.; Moore, W. J.; Stenau, E.; Swedlow, J. R.; Vuong, J. & Procter, J. B. (2018). Visualization of Biomedical Data. *Annual Review of Biomedical Data Science Visualization*, 1:275--304.

Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N. H.; Chavali, G.; Chen, C.; Del-Toro, N.; Duesbury, M.; Dumousseau, M.; Galeota, E.; Hinz, U.; Iannuccelli, M.; Jagannathan, S.; Jimenez, R.; Khadake, J.; Lagreid, A.; Licata, L.; Lovering, R. C.; Meldal, B.; Melidoni, A. N.; Milagros, M.; Peluso, D.; Perfetto, L.; Porras, P.; Raghunath, A.; Ricard-Blum, S.; Roechert, B.; Stutz, A.; Tognolli, M.; Van Roey, K.; Cesareni, G. & Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):358--363.

Otto, M.; Thornton, J. et al. (2015). Bootstrap· the world's most popular mobile-first and responsive front-end framework.'. *Getbootstrap. com*.

Paetzel, M.; Karla, A.; Strynadk, N. C. J. & Dalbey, R. E. (2002). Signal peptidases. *Chemical Reviews*, 102(12):4549--4579.

Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M. & Shaw, D. E. (2019). Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, 116(10):4244--4249.

Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605--12.

Pieper, U.; Webb, B. M.; Dong, G. Q.; Schneidman-Duhovny, D.; Fan, H.; Kim, S. J.; Khuri, N.; Spill, Y. G.; Weinkam, P.; Hammel, M.; Tainer, J. A.; Nilges, M. & Sali, A. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 42(D1):1--11.

R Project (2019). R-project. http://r-project.org. Acessado em: 12.01.2019.

Randolph, J. & DeGoey, D. (2005). Peptidomimetic Inhibitors of HIV Protease. *Current Topics in Medicinal Chemistry*, 4(10):1079--1095.

Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N. & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551--1555.

Rawlings, N. D.; Barrett, A. J.; Thomas, P. D.; Huang, X.; Bateman, A. & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, 46(D1):D624--D632.

Reynolds, A. P.; Richards, G.; De La Iglesia, B. & Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475--504.

Ripphausen, P.; Nisius, B.; Peltason, L. & Bajorath, J. (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. *Journal of medicinal chemistry*, 53(24):8461--7.

Rose, A. S. & Hildebrand, P. W. (2015). Ngl viewer: a web application for molecular visualization. *Nucleic acids research*, 43(W1):W576--W579.

Rousseeuw, J. (1989). A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational Application Math.*

Shannon, P.; Markiel, A.; Owen Ozier, .; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13):2498--2504.

Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L. S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y. H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Ben Schafer, U.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C. & Young, C. (2014). Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, (January):41--53.

Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. & Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341--346.

Smith, R. (2017). NVIDIA Volta Unveiled: GV100 GPU and Tesla V100 Accelerator Announced. http://www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced. Acessado em: 26.04.2019.

Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics (Oxford, England)*, 15(4):327--332.

Szklarczyk, D.; Morris, J. H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N. T.; Roth, A.; Bork, P.; Jensen, L. J. & Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362--D368.

Szklarczyk, D.; Santos, A.; Von Mering, C.; Jensen, L. J.; Bork, P. & Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1):D380--D384.

Vidal, M.; Cusick, M. E. & Barabási, A. L. (2011). Interactome networks and human disease. *Cell*, 144(6):986--998.

Von Bertalanffy, L. (1950). The theory of open systems in physics and biology. *Science*, 111(1):23--29.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395--416.

Wagner, D. & Wagner, F. (1993). Between Min Cut and Graph Bisection. Em Borzyszkowski, A. M. & Sokołowski, S., editores, Mathematical Foundations of Computer Science 1993, pp. 744--750, Berlin, Heidelberg. Springer Berlin Heidelberg.

Waldner, B. J.; Kraml, J.; Kahler, U.; Spinn, A.; Schauperl, M.; Podewitz, M.; Fuchs, J. E.; Cruciani, G. & Liedl, K. R. (2018). Electrostatic recognition in substrate binding to serine proteases. *Journal of Molecular Recognition*, (April):1--12.

Wallace, A. C.; Laskowski, R. A. & Thornton, J. M. (1995). LIGPLOT : a program to generate schematic diagrams of protein-ligand interactions Clean up structure. *Protein Engineering*, 8(2):127--134.

White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure*, 28(1):319--365.

Wong, B.; O'Donoghue, S. I.; Gavin, A.-c.; Gehlenborg, N.; Goodsell, D. S.; Heriche, J.-K.; Nielsen, C. B.; North, C.; Olson, A. J.; Procter, J. B.; Shattuck, D. W. & Walter, T. (2010). Visualizing biological data — now and in the future. *Nature Methods*, 7(3s):S2--S4.

Xia, J.-F.; Zhao, X.-M.; Song, J. & Huang, D.-S. (2010). Apis: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC bioinformatics*, 11(1):174.

Yan, J.; Risacher, S. L.; Shen, L. & Saykin, A. J. (2017). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6):1370--1381.

Zaki, M. J.; Meira Jr, W. & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.