PFSTATS: SISTEMA PARA ESTUDO DE FAMÍLIAS DE PROTEÍNAS ATRAVÉS DE DETECÇÃO DE RESÍDUOS CONSERVADOS E DECOMPOSIÇÃO DE REDES DE COEVOLUÇÃO

NELI JOSÉ DA FONSECA JÚNIOR

PFSTATS: SISTEMA PARA ESTUDO DE FAMÍLIAS DE PROTEÍNAS ATRAVÉS DE DETECÇÃO DE RESÍDUOS CONSERVADOS E DECOMPOSIÇÃO DE REDES DE COEVOLUÇÃO

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

Orientador: Lucas Bleicher

Belo Horizonte

Maio de 2016

da Fonseca Júnior, Neli José

F676p PFSTATS: Sistema para estudo de famílias de proteínas através de detecção de resíduos conservados e decomposição de redes de coevolução / Neli José da Fonseca Júnior. — Belo Horizonte, 2016
xxvi, 90 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Lucas Bleicher

1. Bioinformática — Teses. 2. Redes — Teses. I. Orientador. II. Título.

CDU [004.4+519.1](575)

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

PFSTATS: Sistema para estudo de famílias de proteínas através de detecção de resíduos conservados e decomposição de redes de coevolução

Néli José da Fonseca Júnior
Orientadora: Professor Doutor Lucas Bleicher

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais – UFMG, como parte dos requisitos necessários para a obtenção do título de Mestre em Bioinformática.

Examinada por:

_____
Prof. Doutor Lucas Bleicher – UFMG

_____
Profa. Doutora Gisele Lobo Pappa – UFMG

_____
Prof. Doutor José Miguel Ortega – UFMG

_____
Profa. Doutora Laila Alves Nahum - Fiocruz

Belo Horizonte
Setembro de 2016

Dedicuum cest laborae a quelquis personatum que ajudorat a facirelo.

Agradecimentos

Agradeço em primeiro lugar aos meus pais, que sempre me incentivam e se orgulham a cada passo meu. E aos meus irmãos, Thalyta, Matheus e Maria Fernanda.

Também quero agradecer ao meu orientador, Prof. Lucas Bleicher, sem o qual este trabalho não seria possível. Aos colegas e colaboradores Lucas Carrijo, Marcelo Querino e Dhiego Souto. E a todos os membros do nosso grupo de pesquisa.

Aos meus companheiros de república: Keth, Sergio, Dione, Rama, Vitor, Jorge, Robério, Ana, Itálo, Syd, Júlio, Lorena, Pedro, Marília e Nádia, que acompanharam todo o percurso, desde o começo, ajudando nos estudos, nas escolhas e nas dificuldades.

Aos professores Tiago Garcia, Joubert Lima, Fernando Sica e Eduardo Luz, principais responsáveis pelo meu gosto pela academia.

A minha namorada, Thainá, por toda ajuda com conselhos e ensaios.

As repúblicas Pulgatório, Nau Sem Rumo, Sparta e todos os amigos de Ouro Preto.

"Não são as espécies mais fortes que sobrevivem nem as mais inteligentes,
e sim as mais suscetíveis a mudanças."
(Charles Darwin)

Resumo

Análises por conservação e correlação de aminoácidos podem fornecer informações importantes acerca da estrutura e função de famílias de proteínas. Além disso, resultados experimentais sugerem que o enovelamento de proteínas pode ser alcançado com menos caracteres do que os 20 aminoácidos de ocorrência natural. Nosso grupo propôs recentemente um método para obter determinantes de sub-classes funcionais em famílias de proteínas chamado Decomposição de Redes de Coevolução de Resíduos (DRCN). O DRCN consiste de um método baseado em sequência para análises de famílias de proteínas representadas por alinhamentos múltiplos de sequências. Apresentamos um software para análises de famílias de proteínas através de DRCN, estudos de conservação de resíduos, aplicações de redução de alfabeto e busca automática por anotações. Os algoritmos foram agrupados de modo a ter uma aplicação robusta e intuitiva para o estudo de proteínas homólogas. As análises por DRCN necessitam de um único arquivo de entrada obrigatório, um alinhamento múltiplo de sequências (AMS), apesar de que um arquivo no formato PDB também pode ser utilizado para visualização de resultados na estrutura. A qualidade do AMS é o principal fator para obter melhores resultados utilizando esta metodologia, logo, o sistema disponibiliza uma etapa de filtragem de sequências a fim de maximizar a representatividade do AMS através da remoção de fragmentos, sequências mal alinhadas e redundância. Foram estudados quatro domínios de famílias de proteínas: lisozimas de tipo C/alfalactoalbuminas, fosfolipases A2, proteínas reguladoras de nitrogênio PII e o domínio de ligação de DNA dos receptores nucleares IV; três diferentes abordagens de AMS extraídas do PFAM e 19 alfabetos de aminoácidos reduzidos disponíveis na literatura. Nestes estudos, foram encontradas informações sobre sítios catalíticos e de ligação em todas as quatro famílias, além de dados relacionados a estruturas secundárias, núcleo hidrofóbico e sítio de dimerização. Ao observar as arestas de anti-correlação, foi encontrado um ou mais resíduos que separavam duas ou mais subclasses, este é o caso do C122 nas fosfolipases A2. Este nó formou um hub de correlações negativas conectando resíduos de cada uma das outras comunidades identificadas. Sua presença ocorre em 217 sequências,

sendo todas de *Oikopleura dioica*. A utilização de alfabetos reduzidos nas análises por DRCN mostraram aumentar o tamanho das comunidades encontradas, além de manter hipóteses consistentes para seu significado biológico. Porém, em casos como o dos receptores nucleares, o uso de um alfabeto reduzido pode ocultar uma comunidade que compartilha posições em comum com outra.

Palavras-chave: Sistemas Complexos, Conservação de Aminoácidos, Reduções de Alfabeto, Redes de Coevolução.

Abstract

Structural and functional insights about protein families can be obtained by amino acids conservation and correlation analysis. Futhermore, experimental research has suggested that protein folding can be achieved with fewer characters than the 20 naturally occurring amino acids. Our group has recently proposed a method to obtain functional sub-class determinants in protein families, called Decomposition of Residue Coevolution Networks (DRCN). DRCN is a sequence based method for analysis of protein families represented by multiple sequence alignments. We present a software for protein family analysis using DRCN, conservation analysis, alphabet reductions and automatic annotation search. The algorithms were grouped in order to have a robust and intuitive application to the analysis of homologous proteins. The DRCN analysis consists of a unique required input file, a multiple sequence alignment (MSA), besides that a PDB file can be also used to visualize the results in the structure. The MSA quality is a crucial factor to achieve better results with the methodology, therefore, a filtering step is available to maximize its representativeness by removing fragments, poorly aligned sequences and redundancy. We have studied four protein family domains: lysozyme C/Alpha-lactoalbumin, phospholipases A2, nitrogen regulatory protein PII and the DNA binding domain of the nuclear receptors IV; three MSAs aproaches extracted from PFAM and 19 amino acids reducted alphabets from literature. We have found insights about catalyctic and binding sites in all of then, there's also information related to secondary structure, the hydrophobic putative channel and dimer site. By looking for the anti-correlated edges, we could find a residue or a group of residues that separates two or more sub-classes. That's the case of the C122 in the phospholipase A2, this node form an anti-correlated hub that connects every community. Its presence occurs in 217 sequences, all from *Oikopleura dioica*, and all without the phospholipase catalyctic activity. The uses of reduced alphabet in DRCN analysis usually increase the number of residues in each community and in the most cases maintaining a consistent hypothesis for their biological role. But in cases as this nuclear receptors IV study, the uses of a reduced alphabet can hide clusters that share common positions with another

community.

Keywords: Amino Acids Conservation, Complex Systems, Coevolution Networks.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A capacidade de processar e armazenar dados computacionalmente obteve um crescimento astronômico nas últimas décadas. Diante disso, surgiu a possibilidade de armazenar e disponibilizar todo tipo de conhecimento em grandes bancos de dados. Tal avanço computacional vem sendo utilizado em larga escala por diversos áreas do conhecimento, porém o uso de métodos e ferramentas computacionais se tornaram tão frequente nas pesquisas relacionadas a biologia molecular e bioquímica, que um novo campo foi criado, a bioinformática. A bioinformática evoluiu para uma área multidisciplinar que integra o desenvolvimento em tecnologia da informação à biotecnologia e ciências biológicas, utilizando ferramentas computacionais para armazenamento, gerenciamento, mineração e visualização de dados biológicos [O'Leary-Driscoll, 2015].

A descoberta de que a sequência primária de aminoácidos de uma proteína pode determinar tanto função biológica quanto sua estrutura tridimensional [Anfinsen & Corley, 1969; Anfinsen, 1972] foi fundamental para o surgimento da biologia computacional. Logo, estas sequências passaram a serem utilizadas para predição de estrutura e propriedades funcionais. Além disso, a observação de que as sequências biológicas evoluem a taxas mensuráveis e relativamente constantes permitiu a análise da evolução molecular ao nível da sequência [Zuckerkandl & Pauling, 1962, 1965].

A obtenção da sequência de aminoácidos de uma proteína é uma tarefa trivial se comparada à dificuldade de se obter sua estrutura tridimensional ou qualquer outra informação funcional experimental. Consequentemente, houve um aumento exponencial do número de sequências de proteínas armazenadas em bancos de dados públicos, com uma ordem de magnitude maior do que o número de proteínas cuja estrutura é conhecida ou que possuem características de anotações funcionais por via experimental. Tal tendência foi impulsionada pelas recentes melhorias no sequenciamento *de novo* e ressequenciamento.[Chagoyen et al., 2015].

Estudos experimentais, como determinação de funções biológicas, cristalografia de proteínas ou detecção de sítios com importância funcional, não conseguem acompanhar a velocidade exponencial com que proteínas são depositadas. Sendo assim, a partir da necessidade de se estudar proteínas com pouquíssima informação disponível, abordagens computacionais chamam a atenção com resultados que podem servir como guias para futuros estudos experimentais [Pazos & Bang, 2006].

Segundo Chakraborty & Chakrabarti [2014], análises de evolução molecular por abordagens computacionais tendem a ser efetivas na inferência de função de proteínas, pois famílias de proteínas homólogas geralmente compartilham uma função principal, enquanto que as especificidades costumam variar dentro de um subconjunto destas proteínas. Segundo a teoria neutra da evolução molecular [Kimura, 1984], a maioria dos aminoácidos de uma proteína pode passar por mutações aleatórias sem nenhuma alteração na função principal da mesma, sendo apenas alguns poucos sítios sob uma restrição evolucionária mais rigorosa, refletindo uma maior conservação de propriedades de sequência e estrutural.

O alinhamento múltiplo de sequências (AMS) pode ser definido por um alinhamento de três ou mais sequências biológicas, geralmente proteína, DNA ou RNA. Na maior parte dos casos, o conjunto de sequências utilizado para construção do AMS possui um relacionamento evolutivo pelo qual se compartilham uma linhagem e descendem de um ancestral comum. Um estudo comparativo entre as sequências de um AMS fornece uma grande quantidade de informações sobre características estruturais e funcionais de seus membros. Já é bem estabelecido que, com raríssimas exceções, proteínas homólogas partilham a mesma estrutura 3D global, além de muitos aspectos funcionais, que são herdadas por um ancestral comum. Por conseguinte, a comparação de sequências é comumente utilizada para predição de estrutura 3D e função de proteínas, simplesmente por comparar padrões extraídos do alinhamento com características de proteínas já conhecida [Chagoyen et al., 2015]. O AMS é de fato uma importante ferramenta, sendo utilizado em diversas análises *in silico*, incluindo análises de domínio, reconstrução filogenética, descoberta de *motifs* entre outras. É considerado um dos métodos de modelagem mais amplamente utilizados na biologia [Van Noorden et al., 2014], sendo o gerador de AMS, ClustalW [Thompson et al., 1994], uma das publicações cientificas mais citada de todos os tempos [Chatzou et al., 2015].

É muito perceptível que certos resíduos possuem envolvimento na determinação da estrutura tridimensional de macromoléculas, pois estes afetam sua topologia de enovelamento e, portanto, sua estabilidade global [Villar & Kauvar, 1994]. Por exemplo, cisteínas são de extrema importância para manter a estabilidade estrutural em pequenas proteínas [Ramakrishnan & White, 1992], ao passo que argininas e lisinas

fazem o papel de aumentar a compactação da enzima [Nandi et al., 1993]. Prolinas também são importantes para a estrutura, já que, na maioria dos casos, elas alteram grosseiramente a topologia do enovelamento induzindo dobras em domínios de α-hélice, tais como em regiões de transmembrana [von Heijne, 1991]. Estes resíduos tendem a aparecer conservados em um alinhamento múltiplo de sequências.

Posições completamente conservadas em AMSs são interpretadas como resíduos de importância estrutural e funcional para a proteína, uma vez que nenhuma mudança foi permitida nessas posições durante o processo evolutivo. Estas posições foram o primeiro indicador de funcionalidade [Zuckerkandl & Pauling, 1965; Choi et al., 2012] e estão relacionados com todo tipo de sítio funcionais: sítios ativos, sítios de ligação com ligante, proteína-proteína, DNA, etc. [Valdar & Thornton, 2001]. Nem todo resíduo conservado está relacionado a função, mas muitos são conservados devido a requisitos estruturais. As posições conservadas podem, até certo ponto, serem identificada através do aminoácido conservado; alguns resíduos tendem a ter papéis estruturais (Trp, Leu, Gly, Cys), enquanto outros tendem a fazer parte de sítios ativos e de ligação (Asp, Ser, Cys, His) [Villar & Kauvar, 1994; Ouzounis et al., 1997; Pazos & Bang, 2006].

Além dos aminoácidos altamente conservados, outro tipo de posição mostra um padrão mais sútil de conservação. A posição é claramente preservada, mas o tipo de aminoácido é diferente em subgrupos de proteínas distintas no AMS (Fig. 1.1a). Estes subgrupos podem ser definidos por diversos critérios, como filogenia, fenótipo e função. O fato dessas posições serem conservadas pode ser visto como um indicativo de importância funcional, ao passo que, o tipo de aminoácido, por ser diferente para diversas subfamílias, indica que se trata de uma características "específica da subfamília", ou seja, são posições importantes para a função utilizada para definir as subfamílias. Se subfamílias são definidas de acordo com critérios funcionais, estas posições estarão relacionadas com a especificidade funcional. A figura 1.1b ilustra a relação entre posições completamente conservadas e de especificidades da subfamília. Posições conservadas estão presentes em núcleos estruturais e sítios ativos. Já as posições específicas também costumam ser encontradas nos sítios ativos perto das posições conservadas (conferindo especificidade para substratos com diferentes características) e em outras regiões da proteína que são relacionadas com a especificidade, como sítios de interação proteína-proteína [Pazos & Bang, 2006].

Evolução natural produz enovelamentos de proteínas complexos com um alfabeto de 20 aminoácidos. No entanto, acredita-se que durante a síntese primordial da proteína apenas um pequeno número tenha sido envolvido [Davis, 2002]. Diversos estudos têm demonstrado que com uso de uma redução de alfabeto correta, pode ser suficiente para codificar proteínas nativas [Walter et al., 2005; Regan & DeGrado, 1988; Schafmeisterll

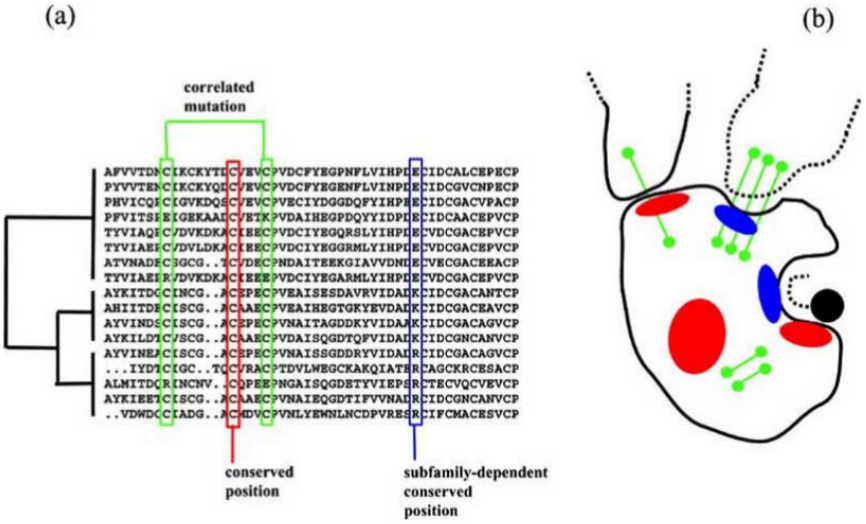


Figura 1.1: **Informações extraídas de AMSs relacionadas à estrutura e função da proteína.** a) Diferentes tipos de conservação de aminoácidos. b) Modelo ilustrando as relações entre essas posições e características estruturais e funcionais. Posições conservadas (vermelho) estão no núcleo e nos sítios ativo da proteína. Posições conservadas da subfamília (azul) também estão presentes no sítio ativo conferindo especificidade. O verde representa pares de resíduos correlatos, muitas vezes apontando para superfície de interação [Pazos & Bang, 2006].

et al., 1997]. Investigações experimentais tem fortemente sugerido que o enovelamento de proteínas pode ser alcançado com menos letras do que os 20 aminoácidos que ocorrem naturalmente [Chan, 1999; Plaxco et al., 1998]. Por exemplo, a estrutura nativa e propriedades físicas da proteína Rop permanecem mantidas quando seus 32 resíduos do núcleo hidrofóbico são formados apenas por alaninas e leucinas [Munson et al., 1996; Liu et al., 2002]. Outro exemplo foi reportado por Heim et al. [2015], os autores conseguiram construir pequenas proteínas transmembranares ativas 26 aminoácidos e utilizando apenas leucinas e isoleucinas.

Além das análises de conservação de aminoácidos já abordadas anteriormente, outra medida utilizada para se extrair informação de um AMS é a partir de correlação de resíduos. Diversas técnicas foram desenvolvidas com intuito de explorar correlação entre aminoácidos em famílias de proteínas [Lesk & Chothia, 1980; Valencia & Pazos, 2002; Jones et al., 2015; Iserte et al., 2015]. É logico se postular que a distribuição de tipos de aminoácidos numa dada posição em um alinhamento múltiplo de sequências é a manifestação de mudanças evolutivas sob restrições impostas pela função. Além disso, é provável que, por razões funcionais, a co-evolução de uma rede de resíduos em uma sequência também ocorra. Sendo assim, tais correlações devem aparecer com sinais estatisticamente significante ao analisar um AMS [Dima & Thirumalai, 2006].

Correlação entre aminoácidos pode ser interpretada como um resultado da tendência de ocorrerem mutações em certas regiões da proteína de forma coordenada quando sua interface necessita ser preservada por motivos funcionais ou estruturais. Assim, mudanças ocorridas durante a evolução na interface de dimerização de um dado monômero A seria compensada por mudanças em um monômero B, a fim de preservar a interface de interação [Filizola et al., 2002]. Análises de correlação podem levar a padrões relacionados a: estrutura tridimensional da proteína [Göbel et al., 1994; Lapedes et al., 1999], efeito filogenético [Lapedes et al., 1999] ou correlação funcional [Singer et al., 1994; Oliveira et al., 2002; Afonso et al., 2013; Bleicher et al., 2011].

Estruturas complexas de rede descrevem uma ampla variedade de sistemas de alta importância tecnológica e intelectual. Por exemplo, uma célula pode ser representada como uma rede complexa de substâncias químicas ligadas por meio de reações; a *internet* é uma rede complexa de computadores e roteadores ligados por vários links físicos ou sem fio; redes sociais também podem ser representada por meio de redes complexas, sendo os usuários representados por nós e as arestas representadas pelos relacionamentos. Estes são apenas alguns dos muitos exemplos que levou a comunidade cientifica a investigar mecanismos que determinam a topologia das redes complexas [Albert & Barabási, 2002]. Segundo Aftabuddin & Kundu [2007], análises de redes têm se tornado cada vez mais reconhecida como uma poderosa ferramenta para estudar sistemas complexos, ajudando a entender a interação entre componentes individuais para então caracterizar todo o sistema.

O estudo das relações entre sequência, estrutura e função de proteínas a partir da perspectiva de redes de aminoácidos (AAN - Amino Acid Networks) utilizando teoria dos grafos pode ser uma área promissora de investigação. A utilização desse tipo de abordagem aplicada a problemas biológicos podem fornecer resultados de interpretação intuitiva acerca das relações complexas nestes sistemas [Lin & Lapointe, 2013]. Isto pode ser exemplificado por muitos estudos anteriores relativos a uma série de temas biológicos importantes, tais como reações catalisadas por enzimas [Andraos, 2008; Chou & Forsén, 1980; Zhou & Deng, 1984], sistemas de metabolismo de fármacos [Chou, 2010], interações proteína-proteína [Kurochkina & Choekyi, 2011; Zhou, 2011a,b; Zhou & Huang, 2013], predição de estruturas secundárias [Ding et al., 2015], detecção de grupos de aminoácidos funcionais [Bleicher et al., 2011; Afonso et al., 2013].

A biologia de sistemas depende de representações precisas de redes de interação, mas estas são muitas vezes difícil de modelar. Interações podem ser condicionais ou contextuais e nem sempre podem ser captadas em um determinado estudo. Abordagens complementares com base de dados experimentais, bem como análises baseada em sequência e evolução são necessárias a fim de descrever um sistema com um grau

suficiente de detalhes para que possa ser suficientemente compreendido. Redes de aminoácidos baseadas em evolução de resíduos de proteínas são denominadas redes de coevolução [Tillier & Charlebois, 2009]. A motivação para o uso de redes de coevolução com objetivo de localizar sítios de importância funcional é baseada em resultados experimentais, em que a mutação da maior parte dos resíduos de uma proteína pouco afeta sua funcionalidade, enquanto que para alguns poucos aminoácidos específicos pode quebrar inteiramente a função [Lee et al., 2008].

Algumas propriedades de rede como: percolação, *clusters*, *hubs*, cliques e comunidades, têm sido estudados em um alto nível de detalhamento indicando que são capazes de extrair informações sobre fatores de estrutura e função de proteínas [Brinda et al., 2010; Petersen et al., 2012; Vijayabaskar & Vishveshwara, 2010; Ding et al., 2015; Bleicher et al., 2011].

Em teoria dos grafos, comunidades são definidas por grupos de vértices que provavelmente compartilham propriedades comuns e/ou desempenham funções semelhantes dentro do grafo[Fortunato, 2010]. Além da identificação de grupos correlacionados, a detecção de comunidades permite classificar os vértices de acordo com sua posição estrutural dentro de seu grupo. Assim, vértices centrais, aqueles que compartilham um grande número de arestas com outros de sua comunidade, podem ter uma importante função de controle e estabilidade dentro do seu grupo, enquanto que os vértices na fronteira desempenham um papel de mediação ou troca de informações entre as diferentes comunidades [Csermely, 2008].

Pode-se também estudar o grafo em que os vértices são as próprias comunidades e as arestas são definidas de acordo com conexões entre alguns de seus vértices no grafo original ou quando seus grupos se sobrepõe. Estudos indicam que redes de distribuição de comunidades têm um grau diferente em relação aos grafos completos [Palla et al., 2005]. No entanto, a origem de suas distribuições podem ser explicadas pelo mesmo mecanismo [Pollner et al., 2005].

1.1 Objetivos

1.1.1 Objetivos Gerais

Produção de um sistema computacional para análises de famílias de proteínas através de conservação de aminoácidos e decomposição de redes de coevolução, capaz de visualizar aminoácidos conservados com possibilidade de utilização de reduções de alfabeto, além de gerar, decompor e extrair informações a partir de redes de coevolução.

Para auxiliar na sua função, também foi gerado uma *API online* (*webservice*) cujo o objetivo é cruzar e fornecer dados dos bancos PFAM, Uniprot e PDB em tempo real.

1.1.2 Objetivos Específicos

- Instalar e manter atualizado os repositórios dos bancos de dados;
- Implementar uma *API online* para fazer a comunicação do cliente com os bancos de dados;
- Elaborar a arquitetura de classes e métodos do *software* cliente;
- Elaborar e implementar metodologia;
- Implementar interface gráfica de usuário;
- Executar bateria de estudos de caso;

Capítulo 2

Revisão Teórica

2.1 Proteínas

As proteínas são grandes biomoléculas, ou macromoléculas, consideradas o principal produto da informação genética, formadas a partir da tradução do RNAm. Constituem-se de cadeias longas de aminoácidos que podem variar de centenas a milhares de unidades e são responsáveis por diversas funções essenciais em organismos vivos. Geralmente atuam dentro de células e são necessárias para manter estrutura, função e regulação de tecidos do corpo. Os principais exemplos de função de proteínas são anticorpos: ligam-se a partículas desconhecidas, tais como vírus e bactérias, com intuito de proteger o corpo; enzimas: são responsáveis pela maioria das reações químicas que ocorrem nas células, além de ajudar na formação de novas moléculas através da leitura da informação genética armazenada no DNA; proteínas mensageiras: transmitem sinais para coordenar processos biológicos entre diferentes células, tecidos e orgãos; componentes estruturais: proteínas que fornecem estrutura e suporte para as células; transporte e armazenamento: estas proteínas são responsáveis por ligar e transportar átomos e pequenas moléculas no interior das células e por todo o corpo.

Existem 20 tipos diferentes de aminoácidos codificado pelo genoma (22 incluindo selenocisteína e pirrolisina, aminoácidos muito raros na natureza) e que são combinados para constituir uma proteína. Essa sequência de aminoácidos determinam a estrutura tridimensional e a função especifica das proteínas. Os aminoácidos são compostos orgânicos com características bem definidas e compartilhadas entre si. A base de um aminoácido, denominada esqueleto peptídico, é formada por um grupo amino, um grupo ácido carboxílico e um átomo de carbono que liga estes dois grupos, denominado carbono alfa ($C\alpha$). O diferencial de cada resíduo consiste de um grupamento ligado ao $C\alpha$, denominado cadeia lateral. A figura 2.1 ilustra cada um dos 20 aminoácidos e sua

estrutura organizados por características físico-químicas de sua cadeia lateral.

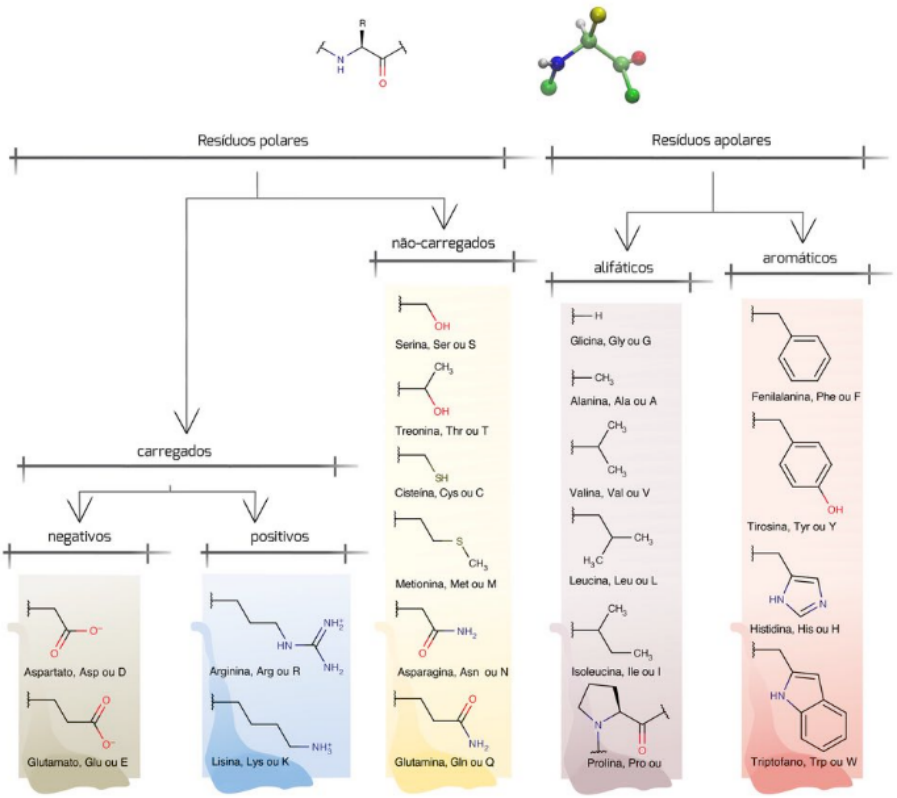


Figura 2.1: **Tabela dos 20 aminoácidos e suas propriedades.** Estrutura dos aminoácidos codificados pelo genoma, organizados segundo as propriedades de suas cadeias laterais. No topo o esqueleto peptídico é representado como encontrado dentro de uma proteína, tanto em sua forma 2D quanto 3D. [Ferreira, 2005].

A estrutura de uma proteína se constitui de uma representação tridimensional dos átomos de sua molécula. Para serem capazes de desempenhar sua função, as proteínas enovelam-se em uma ou mais conformações espaciais específicas motivadas por uma série de interações não covalentes, tais como pontes de hidrogênio, interações iônicas, forças de Van Der Walls e interações hidrofóbicas. Por convenção, a organização estrutural de proteínas foi dividida em quatro níveis: estrutura primária, secundária, terciária e quartenária. Estes níveis seguem uma organização hierárquica, ou seja, a informação de um nível é importante ou necessária para o nível de complexidade seguinte, porém não é o único fator. Por exemplo, normalmente é considerado que a informação contida na sequência de aminoácidos (estrutura primária) é determinante para sua estrutura secundária, porém não a única determinante. Concessões podem ser feitas para permitir uma estrutura terciária ou quaternária mais estável. Assim, uma determinada região em hélice pode ser parcialmente desestruturada para facilitar a formação de um determinado domínio [Ferreira, 2005].

2.1.1 Estrutura Primária

A estrutura primária de uma proteína se refere a sequência linear de aminoácidos na cadeia polipeptídica. Ela se mantém conectada através de ligações covalentes, tais como ligações peptídicas, originadas durante o processo de biossíntese ou tradução da proteína. A cadeia polipeptídica é delineada começando de um grupo amino (N-Terminal) e terminando por um grupo carboxila (C-Terminal). Esta estrutura é determinada de acordo com o gene correspondente a proteína específica. Sua sequência de DNA é transcrita em um RNAm, o qual é posteriormente lido pelo ribossomo em um processo denominado tradução. A estrutura primária da insulina foi descoberta por Frederick Sanger (pelo o qual recebeu o prêmio Nobel), assim estabelecendo que proteínas possuem sequências de aminoácidos bem definidas [Sanger & Tuppy, 1951a,b].

A estrutura primária é geralmente representada por um padrão de letras, nos quais aminoácidos são identificados por um código de uma ou três letras, figura 2.1. Esta sequência de letras representa uma informação de natureza unidimensional, em que a única dimensão descrita é a ordem de aparecimento dos resíduos. Apesar da baixa complexidade, a estrutura primária de proteínas é uma fonte volumosa de informações. Tais informações advêm principalmente da comparação de sequências em busca de padrões específicos associados a determinada característica ou função. Uma vez identificado, esses padrões podem ser utilizados para buscar outras proteínas desconhecidas e com características semelhantes. Essas comparações ainda permitem o estudo da evolução de biomoléculas e de seus organismos, contribuindo no entendimento de como a vida se originou e atingiu seu estágio atual de complexidade [Ferreira, 2005].

2.1.2 Estrutura Secundária

A estrutura secundária consiste de padrões repetitivos de organização espacial originados de interações entre aminoácidos vizinhos e moléculas do solvente (Figura 2.2). Os grupos de estrutura secundária mais comuns em proteínas incluem as alças, hélices e folhas β.

Diferentes sequências de aminoácidos podem levar a uma mesma estrutura secundária, incluindo com propriedades em comum, por exemplo, uma alça é frequentemente uma estrutura muito flexível, já folhas e hélices tendem a ser mais rígidas.

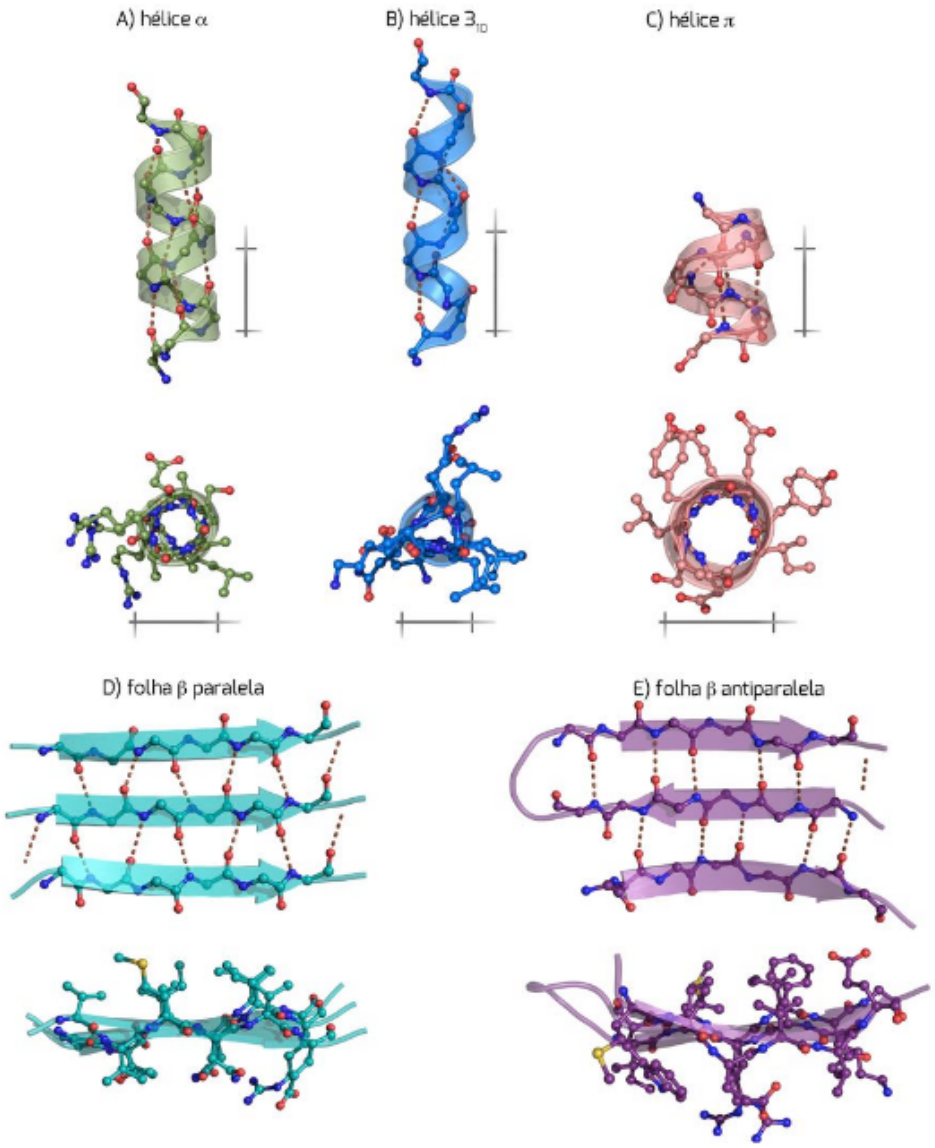


Figura 2.2: **Exemplos de estruturas secundárias.** A) hélices β, B) hélices 3_{10}, C) hélices π, D) folhas β paralelas e E) folhas β antiparalelas. Os traços entre átomos representam ligações de hidrogênio entre átomos do esqueleto peptídico. [Ferreira, 2005].

2.1.3 Estrutura Terciária

A estrutura terciária descreve como os elementos da estrutura secundária se organizam no espaço tridimensional. Esta organização não se dá de forma aleatória, mas sim por um fenômeno denominado enovelamento. Neste processo, uma série de interações, tanto entre partes da própria cadeia polipeptídica, quanto entre o polipeptídio com moléculas vizinhas de água, se combinam para que a proteína adote uma conformação mais estável. As interações responsáveis pelo enovelamento incluem: ligações

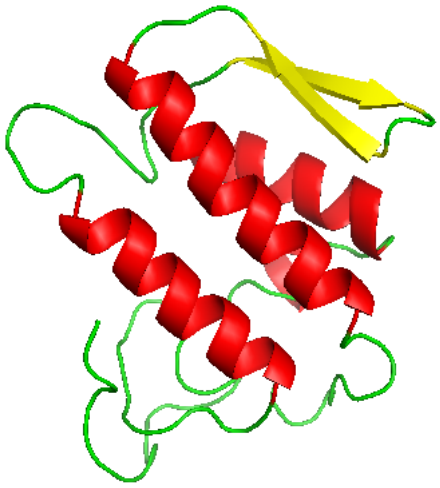


Figura 2.3: **Exemplo de estrutura terciaria.** Estrutura terciária da proteína fosfolipase A2 humana (pdb 1bbc), visualizada no formato cartoon e colorida pelos elementos da estrutura secundária. Hélices de vermelho, folhas β de amarelo e alças de verde.

iônicas, ligações de hidrogênio, interações de van der waals e pontes dissulfeto. A figura 2.3 ilustra a proteína fosfolipase A2 humana colorida por elementos da estrutura secundária.

2.1.4 Estrutura Quartenária

Nem todas as proteínas apresentam este quarto nível de organização estrutural de biomoléculas, ele somente está presente quando há mais de uma cadeia polipeptídica no complexo proteico. A estrutura quartenária consiste de uma estrutura tridimensional de múltiplas subunidades de proteínas e como estas se encaixam. Portanto, esta é regida pela mesma combinação de interações designadas na estrutura terciária. Complexos de dois ou mais polipeptídios são chamados de multímeros, ou mais especificamente, dímero para complexos de duas subunidades, trímero caso tenha três subunidades, tetrâmero, pentâmero e hexâmero para respectivamente, quatro, cinco e seis subunidades. Geralmente estas subunidades são necessárias para realização de determinadas funções em condições nativas. A figura 2.4 mostra a estrutura da proteína RAD51, muito estudada por sua possível relação a diferentes tipos de câncer, colorida por cada uma de suas cadeias.

2.1.5 Domínios

Domínios são regiões de proteínas com a capacidade de evoluir, possuir função específica e existir de forma independente do resto da cadeia. Cada domínio forma

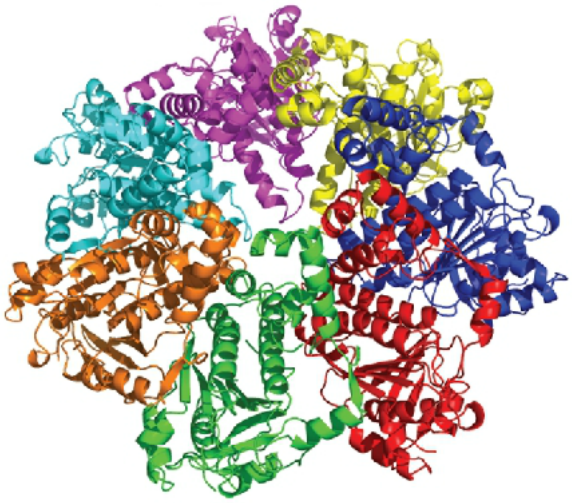


Figura 2.4: **Estrutura quartenária da proteína RAD51.** Estrutura quartenária da proteína RAD51 (pdb 1pzn), visualizada no formato cartoon e colorida por cada uma de suas cadeias.

uma estrutura tridimensional compacta e muitas vezes capaz de se enovelar e estabilizar independentemente. Muitas proteínas possuem mais de um domínio, da mesma forma, um mesmo domínio pode estar presente em diversas proteínas. Em proteínas de multidomínio, cada domínio pode tanto ter sua própria função quanto trabalhar de forma combinada com seus vizinhos.

Um exemplo de proteínas de multidomínio é a enzima piruvato cinase (figura 2.5), uma enzima glicolítica que desempenha um papel importante na regulação do fluxo da frutose-1,6-bifosfato para o piruvato. Ela contém um domínio de ligação de nucleotídios formado por folhas β em azul, um domínio de ligação do substrato em cinza e um domínio regulatório em verde [George & Heringa, 2002].

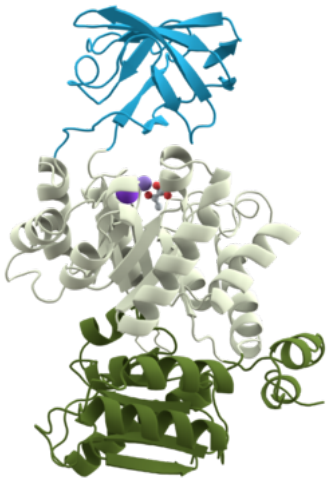


Figura 2.5: **Representação dos três domínios da enzima piruvato cinase.** Os três domínios estão representados pelas cores azul, cinza e verde [George & Heringa, 2002].

2.2 Alinhamento Múltiplo de Sequências

O alinhamento múltiplo de sequências (AMSs) é uma técnica clássica da biologia computacional para visualizar e extrair padrões de um conjunto de três ou mais sequências biológicas. Estas sequências geralmente compartilham uma relação estrutural ou evolucional. Durante o processo de geração do alinhamento, cada sequência do conjunto a ser alinhado é posta em linhas e o algoritmo buscará pela melhor correspondência dos resíduos, permitindo a inserção de *gaps* (caracter "-") de tal forma que resíduos equivalentes permaneçam na mesma coluna e os não equivalentes podem tanto ser postos na mesma coluna como um *mismatch* (incompatibilidade) quanto de frente à um *gap* nas outras sequências. Ao final deste processos, todas as sequências possuirão o mesmo tamanho, possibilitando uma fácil visualização de similaridade, pois caracteres idênticos ou similares estarão representado na mesma coluna (figura 2.6).

O significado preciso de equivalência para avaliar AMSs geralmente é dependente de contexto. Para o filogeneticista, resíduos equivalentes são aqueles que possuem uma relação evolutiva, descendem de um ancestral comum. Para o bioinformata estrutural, serão aqueles com posições análogas pertencentes a enovelamentos homólogos. Para o biólogo molecular, serão aqueles resíduos que desempenham papéis funcionais semelhantes em suas proteínas correspondentes. A figura 2.6 exemplifica essa variedade de tipos de informações que o AMS pode fornecer.

Segundo Ferreira [2005], se duas sequências distintas podem ser alinhadas com um certo grau de similaridade, é possível então assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, portanto podem ser con-

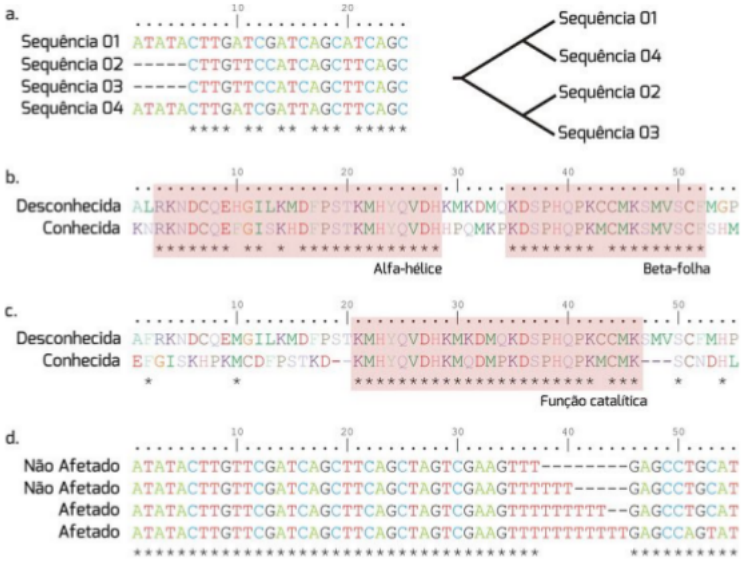


Figura 2.6: **Aplicações dos métodos de alinhamento de sequências biológicas.** a) Reconstrução de árvore filogenética a partir do alinhamento de quatro sequências de nucleotídios. b) Predição de estruturas secundárias de uma proteína desconhecida a partir de um alinhamento com uma sequência cuja estrutura tridimensional é conhecida. c) Predição de função de um domínio de uma proteína conhecida, observando a sequência de uma proteína cuja função é conhecida. d) Comparação de sequências de um determinado gene de indivíduos afetados e não afetados por uma doença genética. Os asteriscos identificam colunas com total similaridade de caracteres. [Ferreira, 2005].

siderado sequências homólogas. Contudo, é importante ressaltar que homologia não requer estritamente alta identidade entre as sequências. Em biologia molecular, homologia é um conceito essencialmente qualitativo utilizado para definir sequências que se originaram de um ancestral comum.

Existem diversos métodos disponíveis na literatura para busca por sequências homólogas, geração e visualização de alinhamentos múltiplos [Manohar & Singh, 2011; Do & Katoh, 2008]. Além destes métodos, existem também diversos bancos de dados públicos para baixar AMSs com diversas características: TIGRFAM [Haft et al., 2012], SUPFAM [Pandit et al., 2004], SUPERFAMILY [de Lima Morais et al., 2010] e PFAM [Finn et al., 2015].

2.3 Bancos de Dados Biológicos

Um banco de dados constitui-se de uma coleção organizada de dados de forma digital. A maioria dos bancos segue o modelo relacional de dados proposto por Codd [1970]. Neste modelo, os dados são organizados em tabelas compostas por linhas,

também chamadas de registro ou tupla, e colunas, além de possuirem sempre uma chave de identificação única (chave primária). Geralmente as tabelas representam um tipo entidade, por exemplo proteína ou dna, as colunas representam atributos da tabela, como nome ou organismo, e as linhas representam instâncias daquele tipo entidade, por exemplo: "Rad51" e "humano". As tabelas podem ser conectadas uma com as outras por relacionamentos possibilitando uma interação entre tabelas diferentes.

Com o avanço das tecnologias de sequenciamento de nova geração provocado nas últimas décadas, cada vez mais genomas sendo sequenciado e anotado, e interações de proteínas e genes acumuladas, uma quantidade imensa de dados biológicos tem sido acumulado. Logo, os bancos de dados biológicos tem sido de extrema utilidade para armazenar, organizar e disponibilizar tais informações.

Os bancos de dados biológicos são fomentados por experimentos científicos, publicações na literatura, tecnologia de alto rendimento (*high-throughput*) e análises computacionais. Estes bancos incluem informações relacionadas a radiologia, sequências de biomoléculas, genomas, proteomas, metabolomas, estruturas cristalográficas, interações entre biomoléculas, *microarray*, modelos matemáticos, *primers*, fenótipos, taxonomia, entre outros. Três grandes bancos de dados muito utilizados em estudos computacionais de proteínas são: Uniprot [Consortium et al., 2008], PFAM [Finn et al., 2015] e Protein Data Bank [Berman et al., 2000].

O Uniprot é um banco de dados de sequências de proteínas, inclui também diversos tipos de anotações relativas à função, taxonomia, localização subcelular, patologias, classificação, além de referências cruzadas com outros bancos de dados. O banco também é dividido em duas seções, sendo o Swiss-Prot limitado a sequências manualmente anotadas e revisadas, e o TrEMBL composto por todo tipo de sequências, incluindo anotadas automaticamente e sem revisão. O Uniprot possui atualmente 551.193 sequências armazenadas no Swiss-Prot e 62.148.086 sequências armazenadas no TrEMBL.

O PFAM armazena alinhamentos múltiplos de sequências homólogas. Os alinhamentos são construídos utilizando cadeias de markov e as sequências são selecionadas a partir de buscas por proteomas de referências, no Uniprot ou no NCBI. Também é possível baixar o perfil HMM utilizando para construção do alinhamento. Atualmente na versão 29, armazena alinhamentos para 16.295 famílias de porteínas.

O Protein Data Bank (PDB) é um banco de dados de estruturas cristalográficas tridimensionais de biomoléculas grandes, como proteínas e ácidos nucleicos. As estruturas armazenadas no PDB são geradas por cristalografia de raio-X, espectroscopia por ressonância magnética nuclear, crio-microscopia eletrónica ou métodos híbridos, sendo a grande maioria dos dados gerado por cristalografia de raio-x. O PDB armazena atualmente 119.137 estruturas, sendo 110.653 relativas a proteínas.

2.4 Teoria dos Grafos

A teoria dos grafos é uma área da matemática e da ciência da computação utilizada para modelar relações entre objetos de um determinado conjunto utilizando uma estrutura denominada grafo. Um grafo é representado por um par de conjuntos G = {V,A}, onde V representa um conjunto de vértices (também chamados por nós ou pontos) e A define um conjunto de arestas (também denominada arcos ou linhas) formado por pares conectados de vértices. Os grafos são geralmente representados por modelos gráficos, onde os vértices são representados por círculos ou pontos e as arestas por linhas ou setas (figura 2.7).

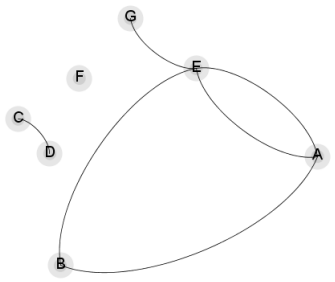


Figura 2.7: **Exemplo de grafo com 7 vértices e 5 arestas.** Grafo G = {V,A}, sendo V = {A,...,G} e A = {{A,B},{A,E},{B,E},{C,D}, {E,G}}.

Este tipo de modelagem têm sido utilizado em diversas áreas do conhecimento com o papel de modelar, visualizar e extrair informação de sistemas complexos. Algumas aplicações clássicas de grafos incluem: trajetos entre cidades, roteamento de veículos, mapeamento de páginas na web, redes de computadores e representação de máquinas de estado finito.

2.4.1 Terminologias

Um grafo pode ser representado graficamente de duas maneiras, utilizando arestas direcionadas ou não direcionadas. Em um grafo direcionado, as arestas apresentam uma relação estritamente orientada de um vértice para o outro, não sendo válido o caminho contrário. Neste caso, para representar uma relação bilateral é necessário a utilização de duas arestas entre o mesmo par de vértices, sendo uma para cada direção. Nos grafos não direcionados, todas as relações são consideradas bilaterais (figura 2.8).

O grau de um vértice é definido como o número de arestas incidentes a este. O grau máximo de um grafo G é frequentemente denotado por $\triangle(G)$, enquanto o grau mínimo é denotado por $\delta(G)$. Grafos em que todos os vértices possuem o mesmo grau k, são denominados k-regular (figura 2.9).

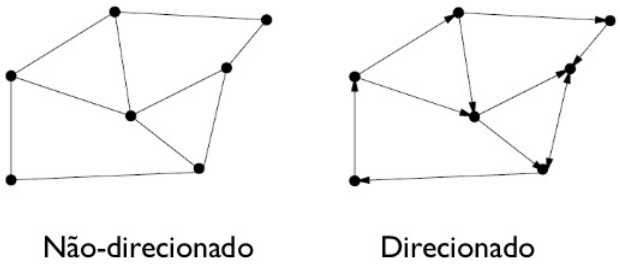


Figura 2.8: **Direcionalidade de grafo.** Exemplos de um grafo direcionado e um não direcionado.

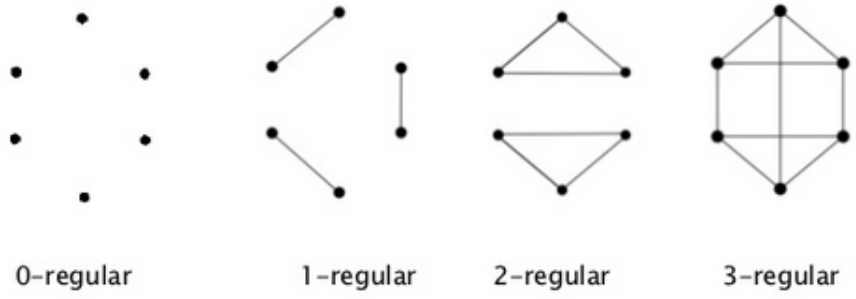


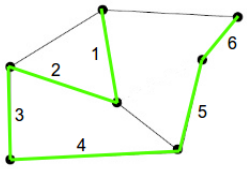
Figura 2.9: **Grafo regular.** Exemplos de um grafos regulares de grau 0, 1, 2 e 3.

Um caminho em um grafo consiste de uma sequência de vértices tal que para cada vértice há uma aresta para o próximo vértice da sequência. Existem alguns tipos diferentes de caminhos, por exemplo um ciclo é um caminho em um grafo que inicia e termina no mesmo vértice. Caminhos sem repetição de vértices são chamados de caminhos simples. Caminhos que passam por todos os vértices do grafo exatamente uma vez, são chamados de caminho hamiltoniano. De forma semelhante, o caminho euleriano é um caminho que passa por cada aresta uma única vez. O comprimento de um caminho é determinado pelo seu número de arestas. A figura 2.10 ilustra alguns exemplos de caminhos.

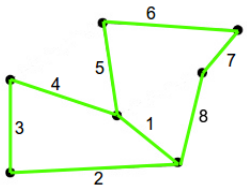
2.4.2 Conectividade

Conectividade é um conceito na teoria dos grafos para determinar o quão seus vértices estão conectados. Análises de conectividade são importantes para clusterização e quantificação de robustez das redes.

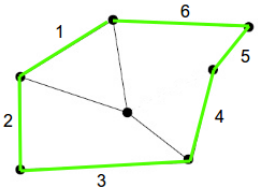
Dois vértices de um grafo G são ditos conectados se existe um caminho em G ligando ambos. Grafos os quais para quaisquer dois de seus vértices existe um caminho interligando-os são denominados conexos. Todo grafo desconexo é formado por



(a) Caminho hamiltoneano de comprimento 6.



(b) Caminho euleriano de comprimento 8.

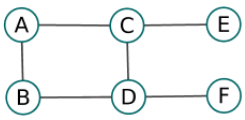


(c) Ciclo de comprimento 6.

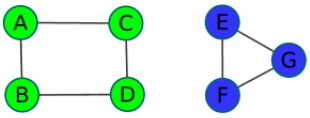
Figura 2.10: **Exemplos de caminhos.** Caminho hamiltoneano, caminho euleriano e ciclo.

pelo menos dois subgrafos conexos disjuntos, cada um destes subgrafos é denominado componente conexo.

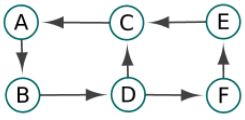
No caso dos grafos orientados, é dito fortemente conexo quando o grafo possui um caminho ligando quaisquer dois de seus vértices. Caso o grafo não seja fortemente conexo, porém mantém uma conexidade ao transformar todas as suas arestas em arestas não direcionadas, é denominado fracamente conexo. A figura 2.11 demonstra alguns exemplos de grafos com conectividades diferentes.



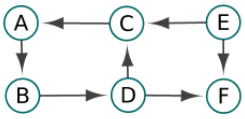
(a) Grafo conexo.



(b) Grafo desconexo com dois componentes conexos.



(c) Grafo fortemente conexo.



(d) Grafo fracamente conexo.

Figura 2.11: **Exemplos de conectividade.** 4 grafos com diferentes conectividades.

2.5 Redes Complexas

Sistemas complexos representam uma abordagem relativamente nova da ciência que estuda como as relações entre diversas partes dão origem a comportamentos coletivos e como o sistema interage e se relaciona ao seu ambiente. As três principais abordagens no que diz respeito ao estudo de sistemas complexos são: como as interações dão origem a padrões de comportamento, o espaço de possibilidades e a formação de sistemas complexos a partir da geração de padrões e evolução. [Bar-Yam, 2002]. Um sistema complexo pode ser visto como um sistema composto por diversos componentes que interagem uns com os outros, logo é útil representar estes sistemas a partir de modelos baseados na teoria dos grafos, com uma rede onde os nós representam os componentes e as arestas suas interações. Estas redes são denominadas redes complexas. O estudo de redes complexas é frequentemente baseado em análises de conectividade da rede, seja buscando padrões de interação entre os nós ou no agrupamento de grupos de vértices relacionados.

O anexo A contém alguns exemplos de sistemas complexos modelados por teoria dos grafos. A.1 ilustra a conectividade entre usuários do Facebook® ao longo do planeta. A.2 representa um sistema de cadeia alimentar de espécies do lago Little Rock em Wisconsin nos EUA. Neste modelo, os nós representam espécies nativas do lago e as arestas interações predatórias. A altura representa o nível trófico, sendo observado fitoplânctons nos níveis mais baixos e peixes nos mais altos. Em A.3 a rede foi modelado de acordo com palavras chaves extraídas da internet com ligação a eleição presidencial dos EUA de 2012 . A coloração das arestas identifica ligações positivas na cor verdes e negativas na cor vermelha, já o tamanho do vértice está relacionado ao seu grau. Por fim, a modelagem representada no anexo A.4 representa um sistema de interações proteína-proteína em leveduras. Nós vermelhos indicam proteínas essenciais, sua remoção causa morte celular; nós na cor laranja indicam alguma importância, ou seja sua remoção reduz o crescimento celular; nós com a coloração verde representam proteínas de menor significância e amarelo proteínas desconhecidas.

2.5.1 Redes Aleatórias

Por mais de 40 anos, a ciência tratou todas as redes complexas como sendo completamente aleatórias Barabási & Bonabeau [2003]. Paradigma que teve suas raízes baseada no trabalho de dois matemáticos, Paul Erdős e Alfréd Rényi, que em 1959, propuseram a teoria das redes aleatórias, na qual sistemas complexos poderiam ser efetivamente representados por uma rede conectando seus nós de forma aleatória [Erdős

& Rényi, 1959]. Paul Erdős e Alfréd Rényi definiram uma rede aleatória como sendo uma rede composta por N vértices conectados por n arestas que são escolhidas de forma aleatória a partir do conjunto de todas as possíveis conexões.

A construção de uma rede aleatória parte de um conjunto de N vértices isolados. A cada etapa sucessiva, uma quantidade de arestas de acordo com uma probabilidade p (p = 0: nenhuma aresta; p = 0.5: 50% do conjunto total de arestas; p = 1: grafo completo) são adicionadas aleatoriamente até que o grafo completamente conectado seja gerado, ou seja p = 1 (figura 2.13). O objetivo principal da teoria das redes aleatórias é observar o surgimento de padrões de propriedades em redes com diferentes probabilidades de conexão p. Erdős e Rényi descobriram que muitas propriedades importantes de grafos aleatórios aparecem de forma repentina, ou seja, dado uma probabilidade p, quase todos os grafo apresentam uma propriedade Q, por exemplo, cada par de nós está ligado por um caminho de arestas consecutivas Erdos & Rényi [1961]; Albert & Barabási [2002].

Erdős e Rényi estudaram diversos tipos de propriedades que poderiam ser extraídas de grafos aleatórios como padrões de conectividade, ciclos, árvores, subgrafos completos, componentes conexos e *clusters*.

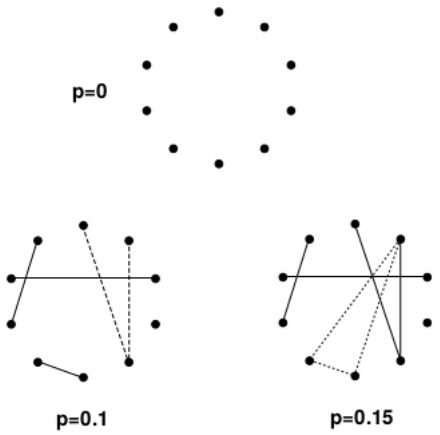


Figura 2.12: **Construção de uma rede aleatória segundo o modelo de Erdős e Rényi.** O processo inicia com N = 10 vértices isolados na rede p = 0. Na parte inferior da figura é ilustrado dois estágios diferentes do processo de construção da rede. Em ambos os grafos é possível observar a ocorrência de componentes conexos, sendo todos árvores em p = 0.1 (uma árvore de ordem 3 representada em linhas tracejadas). Em p = 0.15, é observado um cluster conectado por metade dos nós, além de um ciclo (representado pelas linhas tracejadas) [Albert & Barabási, 2002].

2.5.2 Redes Livre de Escala

Em 1998, Barabási & Albert desenvolveram um projeto para mapear a Internet, modelando cada página como um nó e as arestas representavam *links* entre duas páginas. Eles esperavam encontrar uma rede aleatória, pois, segundo os próprios autores, as pessoas seguem exclusivamente seus interesses ao decidir quais sites vincular em suas páginas e dado a enorme variedade de interesses de diferentes pessoas e ao enorme número de páginas disponíveis, o padrão resultante deveria aparecer bastante aleatório. Foi desenvolvido um programa robô, que navegava pulando de página em página na Internet coletando todos os *links* disponíveis. Apesar do *software* ter atingindo apenas uma pequena fração de toda a *web*, a rede gerada revelou que algumas poucas páginas altamente conectadas estão basicamente sustentando a *web*. Mais de 80% das páginas possuíam menos de 4 conexões, enquanto apenas 0,01 com mais de 1.000. Barabási & Albert observaram que o grau de distribuição da rede seguia a lei da potência, assim, a probabilidade de um nó ter k ligações decai quando k aumenta seguindo a fórmula $P(k) \sim k^{-\gamma}$, sendo γ o expoente livre da escala e determina a probabilidade de P(k) da ocorrência de nós com grau k na rede. De um modo geral, nas redes livre de escala, a maioria dos nós possui poucas ligações em contraste com uma pequena quantidade de nós com com uma grande quantidade de ligações (*hubs*) e que tendem a estarem conectados uns com os outros. Este tipo de rede foi denominada rede livre de escala (*Scale-Free Networks*).

Ao longo dos anos, pesquisadores encontraram estruturas livre de escala em diferentes tipos de sistemas complexos: rede social expressando aparição de atores em filmes [Barabási & Albert, 1999], redes de citações em revistas científicas [Eom & Fortunato, 2011], estrutura física da Internet [Percacci & Vespignani, 2003], redes de transporte aéreo [Guimera et al., 2005], redes sociais [Krawczyk et al., 2011], propagação de epidemias [Pastor-Satorras & Vespignani, 2001], redes metabólicas [Ma & Zeng, 2003], atividade cerebral [Hanson et al., 2016; Uzuntarla et al., 2015], co-expressão genica [Gibson et al., 2013], interação proteína-proteína [Jeong et al., 2001], entre muitos outros exemplos. A figura 2.14 ilustra uma representação do sistema aéreo americano utilizando a teoria das redes aleatórias e a teoria das redes livre de escala.

Barabási & Albert propuseram duas razões para justificar a ocorrência de *hubs* nas redes livre de escala: mecanismos de crescimento e fixação preferencial. A primeira se refere ao fator de crescimento da rede, pois muitas redes não têm um modelo constante. Por exemplo, em 1990 a *web* possuía uma única página, hoje já passa de três bilhões. O surgimento de novos nós tendem a conectar em sítios mais conectados. A segunda justificativa é da fixação preferencial, ou seja, não se pode tratar todos os nós

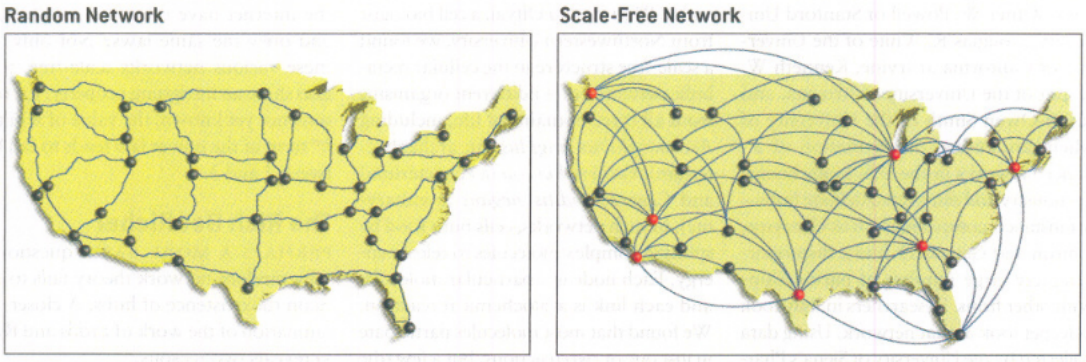


Figura 2.13: **Rede aleatória e rede livre de escala.** O sistema aéreo americano por duas representações, sendo a da esquerda pela teoria das redes aleatórias, com uma distribuição de conexões entre nós mais próxima da linearidade. Em contraste, à direita, um modelo de redes livres de escala com a presença de alguns nós *hubs* representado pela cor vermelha. [Barabási & Bonabeau, 2003].

igualmente. Por exemplo, apesar de haver mais de 3 bilhões de possíveis páginas para *linkar*, o desenvolvedor da página tem seu conhecimento limitado a uma pequena fração da rede e tende a conectar sua página à outras páginas mais populares. Com o tempo, essas páginas já conectadas tendem a possuir cada vez mais conexões. Estes processos foram denominados "rich get richer" (rico se torna mais rico) e tende a favorecer os nós mais antigos, que eventualmente acabam se tornado *hubs*.

Crescimento da rede e fixação preferencial podem inclusive ajudar a explicar a presença de *hubs* em sistemas biológicos. Wagner & Fell descobriram que as moléculas mais conectadas na rede metabólica de *Escherichia coli* frequentemente possuíam uma origem evolutiva mais antiga, sendo algumas moléculas possíveis remanescentes da era pré-DNA e outras componentes das mais ancestrais vias metabólicas.

É dito que as redes livre de escala são resistentes a falhas acidentais, porém vulneráveis a ataques coordenados, ou seja, a remoção de nós aleatórios tem alta probabilidade de remover um nó de baixa importância para topologia, já que estes são muito mais abundantes que os *hubs*. Entretanto, a remoção de um nó *hub* pode afetar drasticamente a topologia da rede.

Capítulo 3

Metodologia

O produto deste trabalho consiste em um sistema para estudo e análises de famílias de proteínas. O sistema inclui métodos para filtragem de alinhamentos, análises de conservação de resíduos, reduções de alfabeto, análises de resíduos correlacionados, além de organização e visualização de dados. O sistema desenvolvido foi chamado de PFStats, uma continuação aos pacotes desenvolvidos por Bleicher et al. [2011].

Dados biológicos são massivamente armazenados em diversos bancos de dados, porém o escopo de uma análise de bioinformática geralmente ultrapassa o nível de informação de apenas um deles. Como foi relatado em Stein [2003], existe uma dificuldade de integração entre os diversos bancos de dados biológicos necessários para um estudo de caso real. Além desta dificuldade de integração, outro problema é relativo ao custo para requisitar mais de um banco via Internet. Pensando nisso, o PFStats foi projetado para utilizar uma arquitetura "cliente-servidor", sendo implementados duas aplicações paralelas. A maior parte de suas funcionalidades se encontra no *software* cliente, aquele que é distribuído para os usuários. Enquanto que do lado do servidor foi desenvolvido uma *API web*, utilizada pelo *software* cliente com o objetivo de realizar consultas que interpolam o escopo de mais de um banco de dados clássico.

O PFStats está atualmente disponível para os sistemas Windows e Linux (baseado em Debian) nas versões para *desktop* e *clusters*. A versão *desktop* possui uma interface gráfica de usuário que facilita sua utilização, porém pode não ser viável para AMSs de tamanho extremamente grande, já na versão *clusters*, o sistema é utilizado através do terminal de linha de comando. Foi utilizado o *framework* Qt (https://www.qt.io/) para desenvolvimento da interface gráfica, as bibliotecas OpenMP [Dagum & Enon, 1998] para paralelização de código, a arbor.js para visualização das redes (http://arborjs.org/) e 3DMol para visualização de estruturas cristalográficas [Rego & Koes, 2014].

3.1 API Web

Para o desenvolvimento deste módulo foi utilizado um servidor *web* pertencente ao grupo de biologia computacional da UFMG (http://www.biocomp.icb.ufmg.br/biocomp/). Neste servidor foi criado repositórios de três bancos de dados biológicos clássicos para comunicar com o PFStats: PFAM, UniprotKb e PDB.

A comunicação entre os bancos e o cliente é realizada através de um pequeno *webservice REST* desenvolvido na linguagem java e com auxílio do *framework* Jersey (https://jersey.java.net/). O REST é um protocolo de comunicação para *webservices* de maior simplicidade. Ele utiliza a própria HTTP e não impõe restrições ao formato da entrada e da saída (JSON, XML e texto puro são formatos usualmente utilizados em *webservices REST*). No caso deste trabalho, o sistema foi desenvolvido com intuito de apenas fornecer uma API online para o cliente, sendo assim, o método de envio foi limitado ao POST e os formatos de entrada e saída a texto puro.

Atualmente o *webservice* conta com seis funcionalidades. Três destas são utilizadas para encontrar sequências que podem ser interessantes para comparação por serem proteínas já com um certo grau de conhecimento a nível experimental: filtro de sequências por escore de anotação do Uniprot, por proteínas que já tenham uma estrutura cristalografia disponível e por sequências que já tenham anotações experimentais referentes a resíduos da sequência. Além destes, também está disponível um filtro por taxons com o objetivo de remover possíveis viés taxonômico do alinhamento, a possibilidade de recuperar códigos PDB referente a um conjunto de sequências e por fim retornar uma lista contendo todas as características anotadas para todas as sequências do conjunto de entrada, este método é utilizado pelo cliente no módulo de busca por anotações dos resultados obtidos que será abordado posteriormente.

3.2 Obtenção do Alinhamento

A qualidade do AMS utilizado como entrada é fundamental para obter resultados satisfatórios com esta metodologia, sendo que este deve ser constituído por sequências homólogas e que tenha uma quantidade razoável de sequências para que haja uma representatividade mínima a que possa ainda ser extraído informações relevantes. O alinhamento pode ser tanto gerado pelo usuário, quanto fornecido por algum banco dados, desde que estejam nos formatos *Fasta*, *Selex* ou *Stockholm*, que são compreendidos pelo PFStats. Além disso, o sistema aceita AMSs tanto formado por sequências comuns (resíduos representados por letras maiúsculas e *gaps* por '-') quanto por alinhamentos

que incluem informações relativas ao perfil HMM utilizado para gerá-lo (inclusão de letras minúsculas e '.'). É possível também carregar o AMS pela própria interface do PFStats apenas sabendo seu código do PFAM.

3.3 Filtragem de Dados

A etapa de filtragem é essencial para melhorar a representatividade de um AMS. Problemas como fragmentos de sequências, regiões mal alinhadas, redundância (correlação devido ao efeito filogenético) ou enviesamento podem ser aliviados ao aplicar determinados filtros. O PFStats implementa 5 tipos de filtros: taxonômico, cobertura por comparação com sequência de referência, cobertura a partir do perfil HMM, identidade mínima e identidade máxima.

Os filtros por cobertura são utilizados para remover fragmentos de sequências, ou seja, sequências do alinhamento que possuem baixa cobertura em relação a uma com as outras. No caso da filtragem de cobertura por comparação com uma sequência de referência, além da sequência, é necessário informar a fração mínima de cobertura, assim todas as sequências do AMS com tamanho inferior a essa fração da sequência de referência são removidas.

Alguns AMSs incluem informações adicionais relativas ao perfil HMM que foi utilizado para gerá-lo. Nestes casos, os aminoácidos podem ser representados tanto por letras maiúsculas, resíduos considerado bem alinhado, quanto letras minúsculas, resíduos que não foram possíveis de alinhar. Além das letras, estes AMSs também costumam representar os *gaps* pelos caracteres '.' (*dot*) ou '-' (*dash*). Os *dashes* representam um salto numa coluna que o perfil HMM esperava encontrar. Já os *dots* são inserções com intuito de posicionar as sequências para que todas tenham um mesmo tamanho e suas colunas estejam corretamente alocadas. Para alinhamentos que possuem esse tipo de informação adicional, o PFStats permite uma filtragem de fragmentos que não requer o uso de uma sequência de referência. Nesta filtragem a sequência é removida se seu número de aminoácidos dividido pelas número de posições válidas do perfil HMM (letras maiúsculas e *dashes*) for maior que o valor de cobertura mínima.

AMSs podem conter um viés muito grande para determinado taxon, por exemplo, um alinhamento contendo 70% de sequências de cordados e apenas 30% de diversos outros filos pode gerar resultados direcionados para características apenas do filo dos cordados, acrescentado de ruídos gerado pelos dados dos outros 30%. Para tratar esse tipo de enviesamento, existe a possibilidade de filtrar o alinhamento por taxons. O filtro taxonômico utiliza uma rotina implementada no *webservice* para remover todas

as sequências de proteínas que não pertencem a uma determinada linhagem taxonômica. O repositório do UniprotKb armazena a linhagem taxonômica de cada proteína, portanto este filtro realiza uma requisição ao banco para cada proteína do alinhamento e verifica se possui determinado taxon.

A última modalidade de filtros disponível no PFStats é por identidade: máxima e mínima. O filtro de identidade mínima requer uma sequência de referência e uma taxa de identidade a ser utilizada, assim o método remove todas as sequências do AMS que obtiverem uma identidade em relação a sequência de referência inferior a taxa de identidade. Por exemplo, utilizando 15% de identidade mínima e a sequência de referência PA2GA_HUMAN, iria remover todas as sequências com menos de 15% de identidade com a PA2GA_HUMAN. O objetivo desse filtro é remover todas as proteínas que são tão diferentes da sequência de referência que poderiam introduzir ruído nas análises. Já o filtro de identidade máxima, compara cada sequência contra todas as outras e remove a menor sempre que a identidade entre ambas é superior a uma taxa pré estabelecida. O objetivo desse filtro é tentar reduzir enviesamento filogenético causado pelo acúmulo de sequências muito semelhantes no alinhamento.

Vale ressaltar que existem dois tipos de abordagens para filtragem do AMS presente no PFStats: com ou sem sequência de referência. A utilização de uma sequência de referência pode ser benéfica caso o usuário tenha em mente uma sequência bem conhecida da literatura e que tenha domínios bem representativos no alinhamento. Já para a filtragem sem sequências de referência pode elucidar características mais abrangentes do alinhamento, porém requer um AMS que inclua informações sobre o perfil HMM. Por fim, é importante que a quantidade de sequências no AMS após a bateria de filtros seja suficiente para manter uma representatividade e qualidade.

3.4 Resíduos Conservados

Uma das análises disponíveis no PFstats é a observação de aminoácidos extremamente conservados em um alinhamento, ou seja, apesar de um longo período evolutivo, tais posições obtiveram pouca ou nenhuma mutação e isto pode estar relacionado a sua importância funcional ou estrutural. Encontrar sítios conservados em um AMS pode ser aparentemente uma tarefa trivial, porém não basta apenas olhar para a frequência da ocorrência de cada aminoácido por posição, deve-se também levar em consideração os *gaps*. Um alinhamento, pode por exemplo, conter uma coluna com ocorrência de apenas um tipo de aminoácido, mas uma frequência incrivelmente alta de *gaps*. O PFStats utiliza uma metodologia proposta por Dima & Thirumalai [2006] para definir

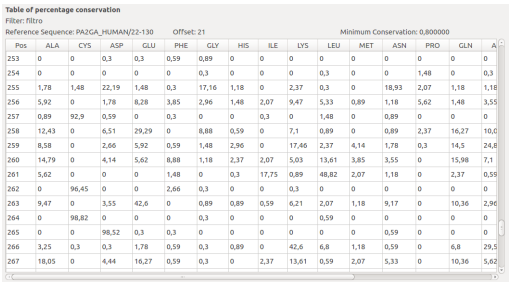
escores de conservação (ΔG_i^{stat}) para cada posição de um alinhamento. O método é baseado na definição de conservação proposta por Lockless & Ranganathan, segundo os autores, conservação de resíduos em uma determinada posição pode ser definido como o desvio global de frequências de aminoácidos numa posição de seus valores médios. Sendo assim, para um AMS evolutivamente bem amostrado, em que sequências adicionais não altere significativamente as distribuições por posições, a probabilidade de qualquer aminoácido x, na posição i em relação a outra posição j está relacionada a energia livre de separação das posições i e j para um aminoácido x, calculada pela distribuição de Boltzmann. Dima & Thirumalai [2006] propuseram a seguinte fórmula para o cálculo de um escore de conservação: $\frac{\Delta G_i^{stat}}{kT*} = \sqrt{\frac{1}{C_i}\sum_{x=1}^{20}\left[p_i^x ln\left(\frac{P_i^x}{p_x}\right)\right]^2}$, sendo kT uma constante arbitrária de energia (utilizamos com valor 10), C_i indica o número de tipos de aminoácidos que aparecem ao menos uma vez na coluna i ($Ci \leq 20$), p_x representa a frequência média do aminoácido x em todo AMS, já p_x^i é a frequência média de x na coluna i.

Além de calcular, o *software* permite visualizar as frequências de cada resíduos por posição, os resíduos com alto escore de conservação no alinhamento e uma visualização da estrutura colorida por conservação, caso o usuário informe um PDB associado a uma sequência do AMS. A figura 3.1 ilustra alguns exemplos destas visualizações.

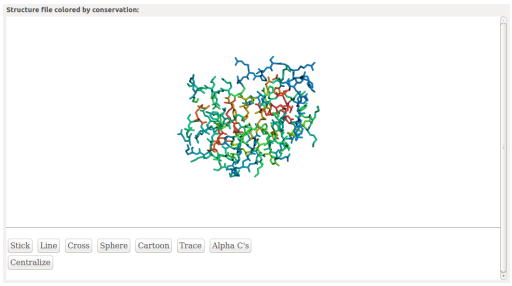
3.5 Resíduos Correlacionados

Outra abordagem pra extração de informação em um AMS implementado no PFStats é a predição de grupos de resíduos correlacionados proposta por Bleicher et al. [2011]. Dizemos que dois aminoácidos x e y estão correlacionados quando a ocorrência de um aminoácido x numa posição i, implica num aumento (ou diminuição no caso de anti-correlação) da frequência de y numa posição j. Dessa forma, dado um número de sequências em um AMS que possuem X em i (n_A) e um número de sequências que possuem Y em j (n_B), pode-se comparar como a frequência de Y em j se comporta em relação a n_A e ao alinhamento global utilizando uma distribuição binomial cumulativa. Contudo, a comparação de um número mínimo de sequências que possuem um determinado aminoácido em uma dada posição com todo o alinhamento global pode causar uma pertubação de representatividade. Por exemplo, uma correlação entre duas posições observadas em um subalinhamento de 10 sequências em relação ao alinhamento global de 500 sequências pode ser devida ao acaso e não refletir uma correlação funcional. Este problema pode ser contornado com a utilização da metodologia proposta por Dima & Thirumalai [2006], que consiste de um subalinhamento de tamanho mínimo

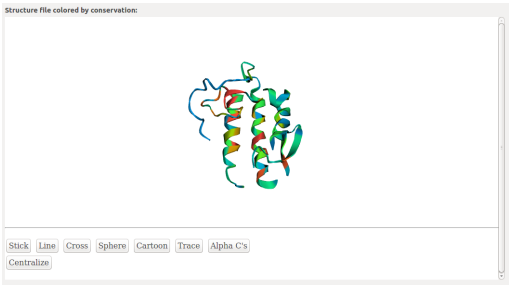
(a) Visualização do AMS colorido por resíduos e filtrados por apenas colunas com alta conservação de resíduos. As posições em negrito representam colunas extremamente conservadas (frequência próxima de 100%).



(b) Tabela de frequência de aminoácidos por posições doa aminoácidos, é possível observar a cisteína 264 e o ácido aspártico 265 com frequências muito próximas de 100%.



(c) Visualização da estrutura 1BBC no formato *sticks* colorida por ΔG_i^{stat}. Resíduos com cores mais próximas do vermelho representam alta conservação e mais próximos do azul, baixa conservação.



(d) Visualização da estrutura 1BBC no formato *cartoon* colorida por ΔG_i^{stat}. Resíduos com cores mais próximas do vermelho representam alta conservação e mais próximos do azul, baixa conservação.

Figura 3.1: **Visualizações de dados de conservação de resíduos.** Exemplos de resultados de análises de conservação utilizando a família PF00068 (Pfam) com a sequência de referência PA2GA_HUMAN.

que mantenha sua representatividade. Os dados de correlação são utilizados para construir uma rede de aminoácidos que por fim é decomposta em comunidades altamente correlacionadas.

3.5.1 Calculo do Subalinhamento Mínimo Representativo

A escolha de uma fração do AMS que mantenha sua representatividade é uma tarefa de suma importância para o cálculo de correlações entre duas posições do alinhamento. Apesar de ainda ser uma escolha manual, o *software* implementa uma abordagem estatística para o auxílio nesta decisão

A entropia de Shannon é muito utilizada para medir diversidade em sequências biológicas [Durbin et al., 1998; Valdar, 2002; Litwin et al., 1992]. A fórmula da entropia

de Shannon, já aplicada ao problema de diversidade de sequências de aminoácidos, é dada por $S = -\sum_{i=1}^{20} p_i log_2(p_i)$, no qual, i representa cada um dos 20 diferentes aminoácidos e p_i sua frequência, $\frac{n_i}{N}$. Para definir um escore de diversidade de todo o AMS, é calculado a média das entropias de cada posição: $score = \frac{\sum_{i=0}^{N} S_i}{N}$.

Tendo definido um escore de diversidade de sequências para um alinhamento, pode-se utilizar a metodologia de Dima & Thirumalai [2006] para amostrar a representatividade de subalinhamentos de diversos tamanhos diferentes, em busca de a partir qual tamanho um alinhamento deixa de ser representativo. O método consiste em calcular escores de diversidade de sequência para diversos subalinhamentos gerados por sequências aleatória e diferentes tamanhos, sendo que para cada tamanho, o cálculo é realizado um número N de vezes (100 por padrão) e um valor médio é utilizado para plotar o gráfico. Segue abaixo o pseudo-código deste método para elucidar seu funcionamento:

```
1  def minss(AMS,N):
2    values = []
3    para i = 1 ate i = 100: #Para cada tamanho do alinhamento
4      meanScore = 0
5      para i = 1 ate i = N: #Repetições para garantir a representatividade
6        subams = rnd(AMS, i) #Retorna um subalinhamento aleatorio de tamanho
7        meanScore += score(subams) #Calcula o score do subalinhamento
8      meanScore = meanScore / N #Calcula a média dos escores
9      values.add(meanScore)
10   retorna values
```

O resultado gerado por este método pode ser plotado em um gráfico de escores médio por tamanho do alinhamento (apêndice A.5). A partir deste gráfico o usuário pode observar pela curva de divergência de diversidade e assim escolher qual tamanho de subalinhamento pode utilizar para o calculo das correlações, mantendo sua representatividade.

3.5.2 Calculo das Correlações

Tendo em mãos o tamanho mínimo em que um subalinhamento aleatório mantenha sua representatividade, pode-se utilizar da distribuição binomial cumulativa para calcular os escores de correlação, também chamados de valor-p. Valor que representa a probabilidade da correlação entre duas posições ocorrer ao acaso.

O subalinhamento mínimo utilizado neste método será denominado AMSm. Considerando n_A como o número de sequências em AMSm que possuem o aminoácido X na posição i, n_B o número de sequências em AMSm que possuem o aminoácido Y na

posição j e N, o número de sequências do AMSm. O número esperado de sequências que possuem Y em j para o subconjunto tendo X em i, considerando que não há uma correlação entre os dois, pode ser calculado por $n_A(n_b/N$. Caso o número observado, $n_{B|A}$, seja maior que o esperado, será utilizado a equação (3.1) para calcular o valor-p de uma correlação, caso contrário, é utilizado a equação (3.2) para calcular o valor-p de uma anti-correlação.

$$p = \sum_{n=n_{B|A}}^{n_A} \frac{n_A!}{n!(n_A-n)!}(n_B/N)^n(1-n_B/N)^{n_A-n} \tag{3.1}$$

$$p = \sum_{n=0}^{n_{B|A}} \frac{n_A!}{n!(n_A-n)!}(n_B/N)^n(1-n_B/N)^{n_A-n} \tag{3.2}$$

Em casos de correlação significativa, o valor-p será muito pequenos, portanto para facilitar sua utilização, será considerado no formato logarítmico. Além disso, afim de diferenciar escores de correlação de anti-correlação, para correlação positiva, o escore será definido por $-log(p)$, e para uma correlação negativa, por $log(p)$. Sendo assim, um valor-p de 10^{-10}, se tratar de uma correlação positiva, terá escore de 10, caso contrário seu escore será -10.

Para definir o que será realmente considerado uma correlação (ou anti-correlação), mais duas variáveis são empregadas: o escore mínimo e a alteração de frequência mínima. O escore mínimo demarca qual o mínimo escore, em módulo, uma correlação deve obter para ser considerada na construção da rede. O segundo parâmetro exige uma variação mínima na mudança de frequência de uma posição. Assim, n_B deve ter um aumento (ou redução) por uma determinada fração mínima em relação a n_A. Por padrão, o PFStats mantém o escore mínimo com 10, assim são consideradas correlações aquelas que atingirem um escore acima de 10 e anti-correlações as que atingem abaixo de -10. A varição de frequência padrão é 0.3, assim para que uma correlação positiva seja considera, deve ocorrer um aumento de frequência para um valor maior que 70%, já para as negativas, uma redução para uma frequência abaixo de 30%.

3.5.3 Montagem da Rede de Coevolução

A partir dos dados de correlação, é construído uma rede não direcional, em que os vértices representam um par aminoácido-posição e as arestas uma correlação (ou anti-correlação) entre dois pares aminoácido-posição, sendo o peso de cada aresta, seu escore de correlação (figura 3.2). Esta tipo de rede é denominado rede de coevolução, pois modela grupos de resíduos altamente correlacionados (ou anti-correlacionados) a

partir de dados de um alinhamento de sequências de origem evolutiva em comum.

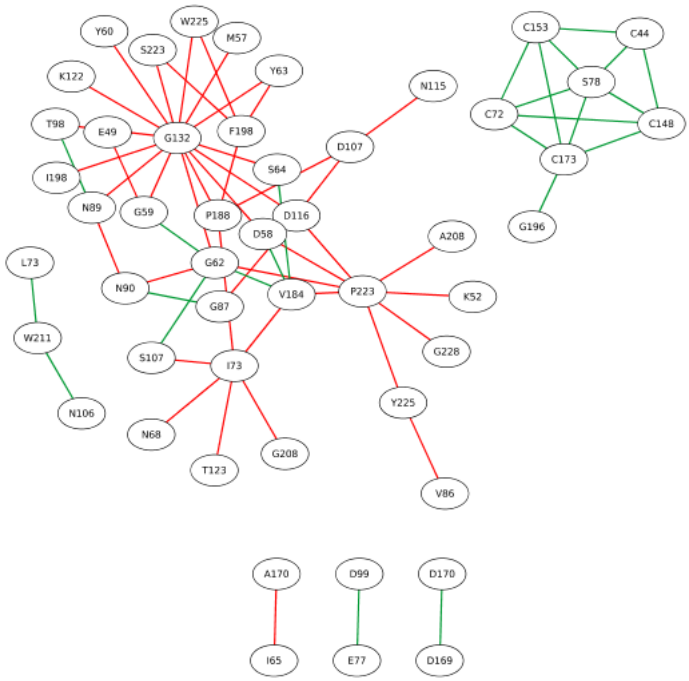


Figura 3.2: **Exemplo de rede de coevolução.** Rede de coevolução gerada a partir de dados de correlação para a família PF00062. As arestas de cor verde representam correlação positiva, já as de cor vermelha indicam correlação negativa.

A partir da figura 3.3, é possível visualizar a ocorrência de *hubs* (região central), grupos altamente correlacionados, grupos de vértices que possuem grande quantidade de ligação entre si e pouca com o resto da rede, além de componentes conexos. Estas ocorrências podem indicar que grupos de resíduos tendem a aparecer juntos e com a metodologia correta, informações relevantes podem ser extraídas.

3.5.4 Decomposição em Comunidades

Com o objetivo de agrupar grupos de vértices que compartilham propriedades em comum, o PFStats utiliza uma abordagem de decomposição em comunidades. Uma comunidade pode ser definida como um subgrafo cujos os vértices possuem alta conectividade entre si, porém uma baixa com o resto da rede. A abordagem utilizada neste trabalho é a decomposição por componentes conexos, nesta metodologia, é considerado apenas as correlações positivas e cada subgrafo conexo restante é tido como uma comunidade. A figura 3.3 ilustra a decomposição em comunidades da rede ilustrada na figura 3.2.

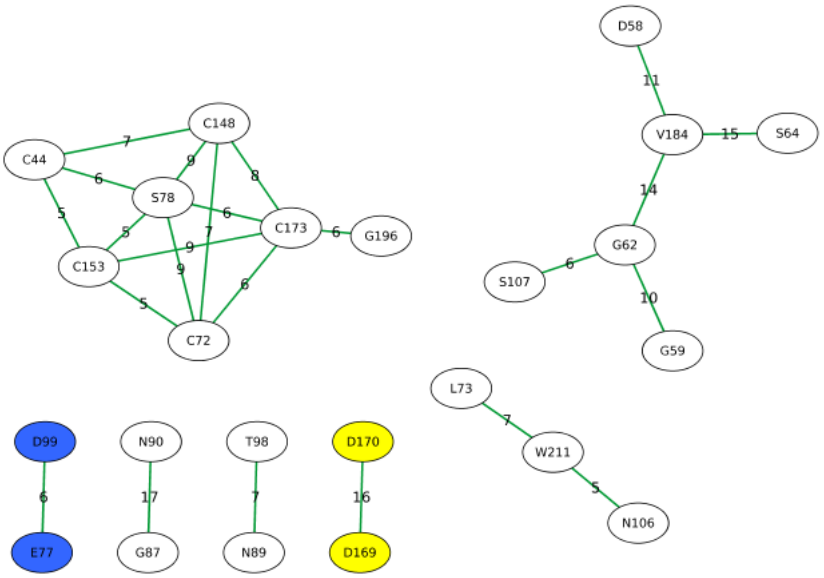


Figura 3.3: **Exemplo de decomposição em comunidades de uma rede de correlação de resíduos.** A decomposição da rede de correlações gerada para a família PF00062 resultou em 7 comunidades conexas. Comunidade envolvida na ligação de cálcio está destacada em amarelo e o sítio ativo em azul.

Além da análise de cada comunidade individualmente, o *software* permite estudar como as comunidades estão conectadas entre sí. Apesar da decomposição ser realizada por componentes conexos, ou seja, em teoria sem nenhuma conectividade com vértices de fora, no processo de montagem da rede há um filtro por escore mínimo para se considerar uma comunidade, logo, apesar de improvável, é possível que haja conectividade significativa entre mais de uma comunidade. Portanto é possível plotar uma rede de comunidades correlacionadas (figura 3.4), nas quais, os vértices são representado por comunidades e as arestas indicam correlação (ou anti-correlação). A fim de quantificar uma correlação entre uma comunidade A e uma comunidade B, é utilizado a equação 3.3, na qual N e M representam o número de vértices das comunidades A e B respectivamente e $w(a_n, b_m)$ o escore de correlação do vértice a_n para o vértice b_m.

$$\Delta AB = \frac{1}{NM} \sum_{a_n, b_m} w(a_n, b_m) \tag{3.3}$$

Outro possível análise implementada no PFStats é a quantificação de quanto cada sequência se encaixa em cada comunidade. Para isto é utilizado a equação de aderência, equação 3.4. A função recebe como parâmetro uma sequência S e uma comunidade A, N_A representa o número de sequências na comunidade A, a função $\delta_S(a_i, a_j)$ retorna

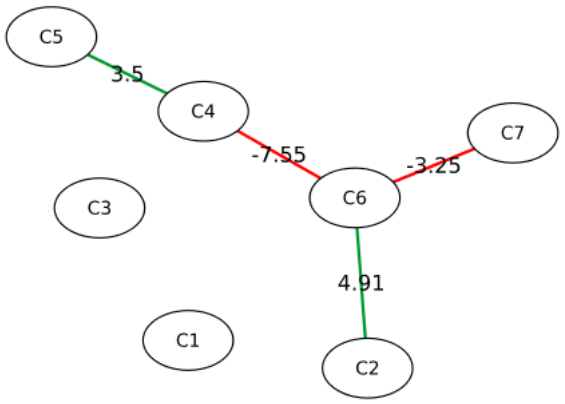


Figura 3.4: **Rede de correlação entre comunidades.** Rede de correlação entre as comunidades decompostas na figura 3.4.

1 caso ambos os vértices a_i e a_j estejam presentes na sequência S e 0 caso contrário. Quanto maior o valor de aderência para uma sequência, mais representativa ela está na comunidade A. Este método tem tanto a utilidade de auxiliar na identificação do significado biológico das comunidades por revisão bibliográfica, guiando quais sequências são mais representativas para cada comunidade, quanto para auxiliar na em aplicações de anotação gênica, identificando outras sequências que possuem as mesmas propriedades que geraram a comunidade A, para que se possa comparar com uma sequência o qual seu significado biológico já foi elucidado.

$$Adh(S,A) = \frac{1}{N_A(N_A-1)} \sum_{a_i,a_j \epsilon A} w(a_i,a_j)\delta_S(a_i,a_j) \tag{3.4}$$

Caso o usuário tenha carregado um arquivo de estrutura PDB referente a alguma sequência do alinhamento, também é possível visualizar a estrutura com as comunidades destacada tanto na própria interface do programa (apêndice A.9) quando baixar em um formato pdb para visualizar em *softwares* externos como o Pymol (https://www.pymol.org/). Outra possibilidade é exportar a rede de correlações em um formato csv para que possa ser plotada em softwares externos como o Cytoscape (http://www.cytoscape.org/).

3.6 Busca por Anotações no Uniprot

O principal passo para identificação do significado biológico das comunidades geradas nesta metodologia consiste em uma extensa revisão na literatura. Entretanto, o PFStats disponibiliza uma etapa anterior capaz de identificar funcionalidades de

alguns resíduos e direcionar a busca na literatura. Este método consiste em buscar no Uniprot por todas as anotações referentes aos pares de resíduo-posição resultantes dos métodos anteriores (conservação e correlação) para todas as sequências do alinhamento. Isto é feito com auxílio da API Web, que recebe uma lista de nomes de sequências e para cada uma destas, realiza uma consulta ao repositório local do Uniprot retornando uma lista de características para o cliente. O cliente então analisa esta lista, tendo o cuidado de converter as posições referentes do alinhamento para as posições referentes a cada sequência. Finalmente, o usuário pode observar o resultado tanto organizado por comunidades, quanto organizado por sequências. As figuras no apêndice A.13a e A.13b ilustram um exemplo de resultado deste método, é possível observar que neste caso que, pelo menos para a proteína *PA2GA_HUMAN*, a comunidade 1 parece ter uma alta relação a ligação de cálcio.

3.7 Redução de Alfabeto

Como já foi relatado anteriormente, algumas trocas de aminoácido podem não afetar o enovelamento ou a atividade desta proteína, sendo assim, o PFStats permite ao usuário executar todos os métodos utilizando alfabetos reduzidos. Dessa forma, as análises podem considerar grupos de aminoácidos ocupando as posições do alinhamento, em vez de utilizar uma identificação para cada um dos 20 tipos diferentes. A figura no apêndice A.8b ilustra um AMS colorido por um alfabeto de 6 grupos, é possível observar nesta imagem que aminoácidos diferentes assumem a mesma cor e são considerados conservados, por exemplo, H,Y,W e F representados na cor vermelha.

O PFStats disponibiliza 20 diferentes alfabetos retirados da literatura para utilização nas análises, além da possibilidade de criação de alfabetos customizados. O primeiro alfabeto disponível, T20, que se trata do alfabeto tradicional, ou seja cada um dos 20 aminoácidos representados por um caractere único.

Três alfabetos foram criados baseado no estudo de propriedades de aminoácidos e consequências de substituição de Betts & Russell [2007]. O primeiro (T2) e mais simples consiste de um alfabeto de apenas 2 letras, formado por aminoácidos que preferem estar em um ambiente aquoso, hidrofílicos e os que não, hidrofóbicos. Os outros dois (T5, T6) alfabetos são baseado na presença de quatro grupos: cadeia lateral alifática, cadeia lateral aromática, aminoácidos polares e aminoácidos pequenos. Os autores justificam essa classificação por motivos de especificidades de cada grupo, por exemplo: os alifáticos são muito pouco reativos, por isso raramente estão diretamente envolvidos na função da proteína; os resíduos aromáticos são capazes de participar de interações

de *stacking*, além de terem um papel fundamental para algumas ligações específicas; os aminoácidos polares têm uma preferência hidrofílica e aqueles que estão no interior da proteína, geralmente participam de ligações de hidrogênio essenciais para substituir a água; o último grupo é dos aminoácidos de tamanho pequenos, pois a substituição de um aminoácido de tamanho pequeno por um grande, pode causar problemas no enovelamento.

Pommié et al. [2004], em seu trabalho com imunoglobulinas, propuseram reduções de alfabetos de um modo diferente de Betts & Russell [2007]. Em vez de agrupar os aminoácidos por propriedades completamente diferentes, os autores criaram três alfabetos, cada um com um específico para uma propriedade. O primeiro deles (3IMG), similar ao T2, também agrupa os resíduos por hidropatia, porém considerando 3 caracteres: hidrofóbicos, hidrofílicos e neutros. O segundo alfabeto (5IMG) trabalha com 5 caracteres e agrupa os aminoácidos de acordo com seus volumes em *angstrom* cúbicos: muito pequeno (60-90 Å^3), pequeno (108-117 Å^3), médio (138-154 Å^3), grande (162-174 Å^3) e muito grande (189-228 Å^3). O último alfabeto de Pommié et al. (11IMG) agrupa os resíduos por propriedades químicas da cadeia lateral. Foram definidas 11 classes, sendo cinco delas contendo um único aminoácido com suas características específicas e as outras seis: alifáticos, presença de enxofre, hidroxilas, resíduos ácidos, amidas e bases.

Alguns outros trabalhos chegaram a alfabetos reduzidos através de abordagens computacionais, principalmente através do uso de matrizes de similaridade. Murphy et al. chegou a 5 alfabetos reduzidos utilizando correlações através da matriz de similaridade BLOSUM50 (Murphy15, Murphy10, Murphy8, Murphy4 e Murphy2). Li et al. [2003] realizou um trabalho semelhante, utilizando a matriz BLOSUM62 e chegando a quatro alfabetos (Li10, Li5, Li4 e Li3). Wang & Wang [1999] também utilizaram uma abordagem computacional calculando o potencial de contato através comparando com a matriz MJ e chegaram a quatro alfabetos reduzidos, sendo dois de cinco caracteres (Wang5 e Wang5v), um de três caracteres (Wang3) e um de dois caracteres (Wang2). A figura 3.5 ilustra o esquema de uma redução de alfabeto recursiva, partindo de um alfabeto de 20 caracteres até chegar em um de apenas dois.

As tabelas completas de cada alfabeto reduzido disponível no PFStats podem ser vistas no apêndice B.

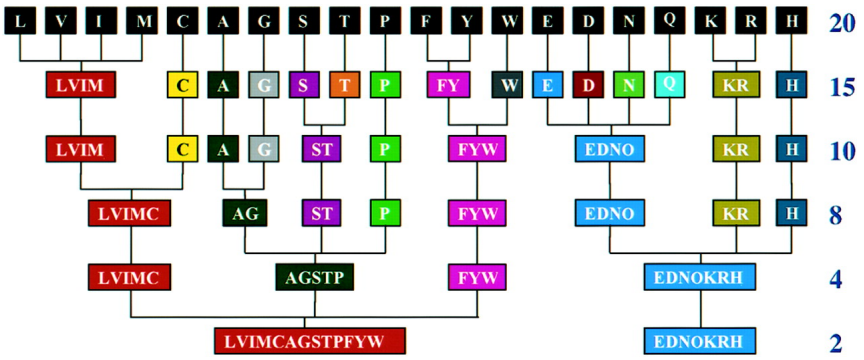


Figura 3.5: **Redução de alfabeto de Murphy et al. [2000].** Esquema para redução de alfabeto por junção de grupos com alta similaridade baseado na matriz BLOSUM50.

3.8 Geração de Sub-Alinhamentos

Uma última funcionalidade implementada neste trabalho, consiste de um gerador de sub-alinhamentos. Neste método, é possível gerar sub-alinhamentos que possuem ao menos uma fração de determinados resíduos ou comunidades, além de sub-alinhamentos de determinado taxon. Ainda é possível quantificar a presença de determinado organismos no alinhamento, afim de observar abundância de um único taxon.

Capítulo 4

Discussão e Resultados

Neste trabalho foram realizado quatro estudos de caso com as famílias: *lisozimas de tipo C/alfa-lactoalbulminas*, *fosfolipases A2*, proteínas reguladoras de nitrogênio P-II e a família de domínio de ligação de DNA dos receptores nucleares. Para cada uma destas famílias, foram construídas e decompostas diversas redes, testando diferentes parâmetros e abordagens. Por fim, a busca pelo papel biológico das comunidades geradas foi realizada tanto pela ferramenta de consulta a anotações do Uniprot disponível no PFStats, quanto por consultas na literatura.

4.1 Lisozimas de tipo C/Alfa-Lactoalbulminas

A família das *Lisozimas de tipo C/Alfa-Lactoalbulminas*, também conhecida por família 22 das glicosídeo hidrolases [Davies & Henrissat, 1995], é basicamente compostas por duas subclasses: a *Alfa-Lactoalbulmina*, uma proteína reguladora no leite, e a *lisozima de tipo C* (EC 3.2.1.17), uma enzima com atividade catalítica bacteriolítica.

As *Alfa-Lactoalbulminas* não possuem a atividade catalítica da *lisozima*, porém podem se associar ao *b-1,4-galactosil-transferase*, formando um heterodímero funcional chamado *lactose-sintetase*, essencial para produção de leite.[Hall & Campbell, 1986]. Além disso, todas as *Alfa-Lactoalbulminas* possuem a capacidade de se ligar a íons de cálcio [Stuart et al., 1986], característica que é restrita a apenas algumas *Lisozimas C* [Nitta et al., 1987].

Apesar de toda essa divergência funcional, tanto a *Lisozima C* quanto a *Alfa-lactoalbulmina* são semelhantes em termos de estrutura primária (sequência de aminoácidos) e terciária (enovelamento tridimensional) e provavelmente evoluíram a partir de um ancestral comum [NITTA & SUGAI, 1989]. Possuem entre 35% a 40% de resíduos conservados, incluindo as posições das quatro ligações dissulfeto.

Nesta análise foram construídas 9 redes, variando entre três abordagens para filtragem de dados: utilizando uma *Lisozima* como sequência de referência, utilizando uma *alfa-lactoalbulmina* como sequência de referência e sem utilizar sequência de referência (filtragem pelo perfil HMM). Além da variação das abordagens de filtragem, também foi testado três tipos de AMS diferentes disponíveis no PFAM: *full*, *uniprot* e *ncbi*.

Os parâmetros utilizados para o filtro do AMS foram 80% de cobertura mínima, 15% de identidade mínima (não utilizado na abordagem por perfil HMM) e 70% de identidade máxima. Os alinhamentos *full*, *uniprot* e *ncbi* possuíam respectivamente 637, 1577 e 2603 sequências antes da filtragem. Apesar da diferença de tamanhos dos alinhamentos utilizados, após a aplicação dos filtros, os tamanhos ficaram mais próximos: o *full* obteve entre 135 e 137 sequências; o *uniprot* entre 218 e 224 sequências e o *ncbi* terminou com entre 169 e 268 sequências. Já para montagem e decomposição da rede, os parâmetros utilizados foram 5 para o escore mínimo de correlação, 20% de subalinhamento mínimo e 0.3 de variação de frequência.

A tabela 4.1 ilustra as comunidades já decompostas para cada uma das 9 abordagens. Os prefixos *Lysc*, *Lalba* ou *Hmm* se referem ao tipo de filtragem utilizado, sendo respectivamente: filtragem por uma sequência de referência do tipo *Lisozima C*, no caso a *LYSC3_PIG*; filtragem por uma sequência de referência do tipo *alfa-lactoalbulmina*, no caso a *LALBA_SHEEP* e filtragem sem sequência de referência, utilizando apenas o perfil HMM. Já os sufixos *full*, *uniprot* e *ncbi* se referem ao tipo de AMS utilizado. As sequências de referências utilizadas na filtragem do alinhamento foram escolhidas pelo critério de haver informação experimental disponível e estar presente em todos os três alinhamentos utilizados.

Lysc/full	Lysc/uniprot	Lysc/ncbi	Lalba/full	Lalba/uniprot	Lalba/ncbi	Hmm/full	Hmm/uniprot	Hmm/ncbi
G37 G40 D36 S42 V118	D36 S42 V118	G40 D36 S42 V118	G37 G40	N62 P65	S42 D36 V118	G37 G40	S42 V118	**E53** **D71** **S69**
W130 Y72	L49 W130 N78	N62 P65	S42 V118	S42 V118	N62 P65	S42 V118	**E53** **D71**	S42 V118
N62 P65	N62 P65	**Q109** **D110**	W130 Y72	**E53** **D71**	N93 K100	W130 Y72	N78 W130	N78 W130
Q109 **D110**	**E53** **D71**		N62 P65	N78 W130	**Q109** **D110**	N62 P65	**Q109** **D110**	**Q109** **D110**
	Q109 **D110**		**Q109** **D110**	**Q109** **D110**		**Q109** **D110**		

Tabela 4.1: Comunidades geradas para a família PF00062. A numeração utilizada se refere a sequência *LYSC3_PIG*.

As comunidades geradas não obtiveram uma grande variação. Todas elas encontraram um par de resíduos que fazem parte do sítio de ligação de cálcio, representado na cor vermelha. Já o sítio ativo das *Lisozimas de tipo C*, representado na cor verde, apareceu apenas nas redes construídas a partir do alinhamento baseado no Uniprot, com uma única exceção de um alinhamento baseado no NCBI e filtrado pelo perfil HMM. A figura 4.1 demonstra a representação gráfica da rede *lysc/uniprot*.

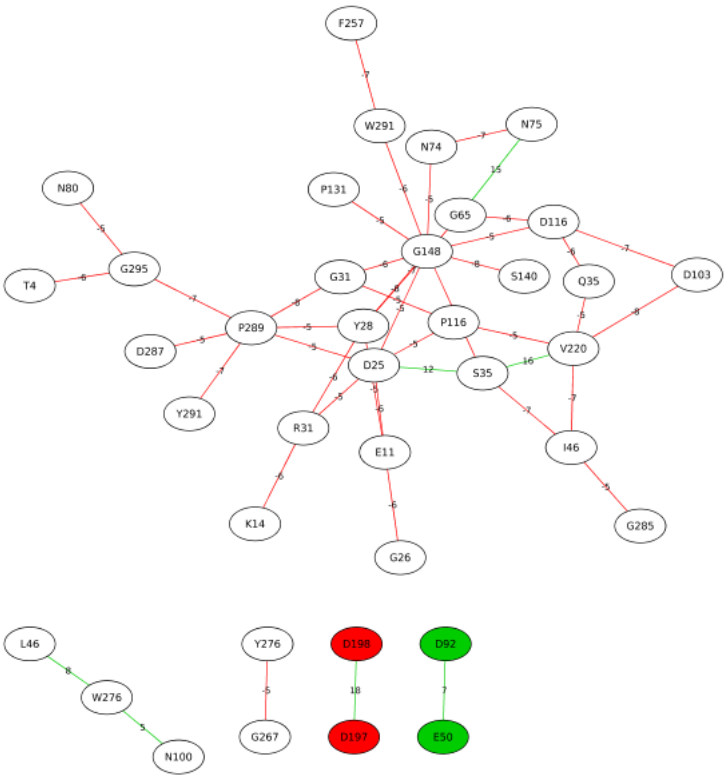
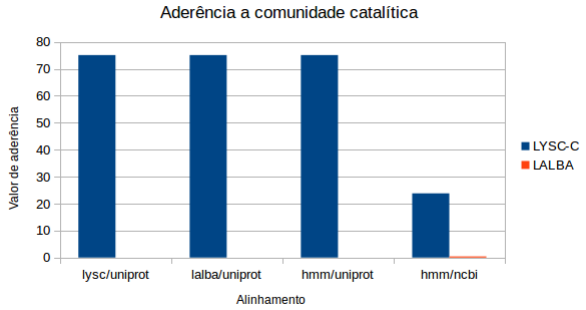


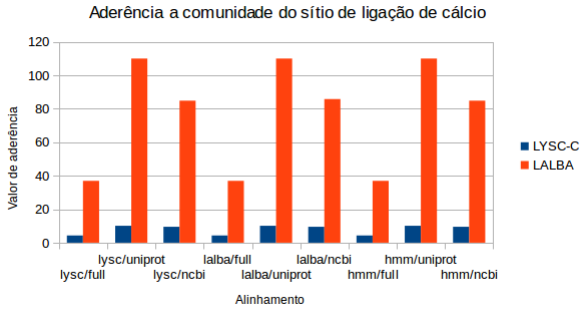
Figura 4.1: **Rede *lysc/uniprot*.** Vértices na cor verde representam o sítio catalítico das *Lisozimas C*, já os vértices na cor vermelha, indicam resíduos que compõe o sítio de ligação de cálcio. Arestas na cor verde indicam correlação positiva, vermelhas indicam correlação negativa. A rede está utilizando a numeração do alinhamento.

Os resíduos N62 e N78 fazem parte da fenda catalítica das *lisozimas C* [Ito et al., 1999], o que pode ser um indicativo de importância para a atividade destas enzimas, apesar de não haver nada na literatura relacionado a seus pares P65 e W130. Alguns experimentos futuros com mutagênese sítio-dirigida podem ser interessante para confirmar ou refutar estas hipóteses. Outro par de resíduos interessante e sem informações na literatura é o S42 e V118. Ambos apareceram em todas as redes, em alguns casos acompanhados de outros resíduos em sua comunidade e sempre com alto escore de correlação.

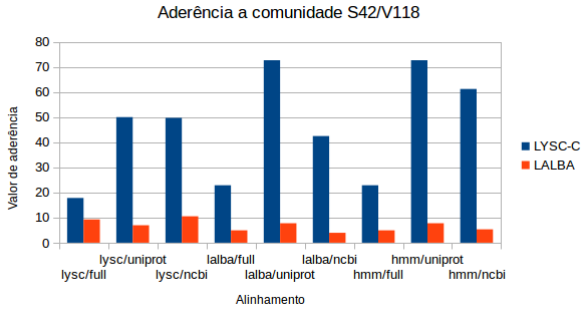
A fim de observar a ocorrência das comunidades encontradas em ambas subclasses, foi separado um grupo de sequências de *Lisozimas C* e de *Alfa-Lactoalbulminas* presentes em cada um dos alinhamentos utilizados. Foi então calculada a aderência média às comunidades do sítio ativo, do sítio de ligação e daquelas que possuem o par S42 e V118. Os resultados foram plotados nos gráficos ilustrados na figura 4.2.



(a) Aderência média a comunidades do sítio ativo da *Lisozima*.



(b) Aderência média a comunidades do sítio de ligação de cálcio.



(c) Aderência média a comunidades que possuem o par S42/V118.

Figura 4.2: **Gráficos de aderência para comunidades do sítio ativo, sítio de ligação e resíduos S42/V118.** Na cor azul está representado as *Lisozimas C* e na cor vermelha as *Alfa-Lactoalbulminas*.

Como já era esperado, a comunidade do sítio catalítico ocorre apenas nas *Lisozimas* (figura 4.2a), de forma também já esperada, os valores de aderência média das *Lisozimas* a comunidades do sítio de ligação de cálcio são ínfimos ao comparado com os da *Alfa-Lactoalbulminas* (figura 4.2b). Isso ocorre, como já foi dito, pelo fato de todas as *Alfa-Lactoalbulminas* possuírem este sítio, o que ocorre em apenas um pequena fração das *Lisozimas*. Já sobre as comunidades dos resíduos S42/V118, não foi encontrado nada na literatura a respeito de nenhum dos resíduos envolvidos em suas comunidades. Contudo, é possível observar no gráfico de aderência (figura 4.2c) que sua presença é muito mais comum nas *Lisozimas* do que nas *Alfa-Lactoalbulminas*. Ao observar esta comunidade na estrutura (figura 4.3), percebe-se que todos os resíduos

envolvidos, representados na cor vermelha, estão em dois *loops* paralelos, um outro resíduo, D119 representado na cor rosa, que não aparece em nenhuma comunidade, porém está presente em uma destas fitas, é responsável pelo sítio de ligação com o substrato. O que pode indicar uma hipótese de que os resíduos desta comunidade podem ter importância na ligação com o substrato e consequentemente na atividade catalítica.

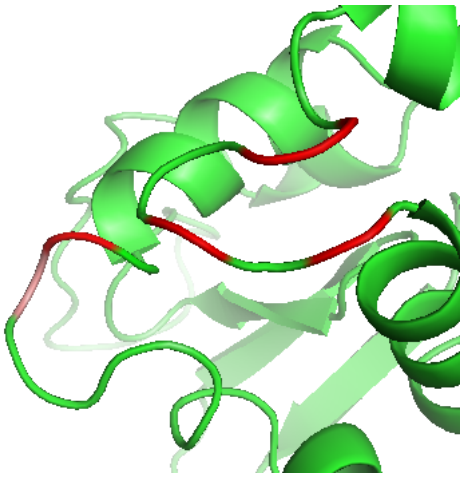


Figura 4.3: **Comunidade 1 da rede *lysc/full* ilustrada na estrutura de uma *Lisozima C*.** Os resíduos da comunidade estão representados na cor vermelha. Na cor rosa se encontra o resíduo responsável pelo sítio de ligação com o substrato.

4.2 Fosfolipases A2

As fosfolipases A2 (EC 3.1.1.4), também chamadas de PLA2s, são enzimas que catalisam a hidrólise da ligação 2-acil éster de fosfolipídeos, liberando ácidos graxos e lisofosfolipídeos [Van Deenen et al., 1963]. Além de possuirem um papel fundamental no metabolismo de lipídeos, o ácido araquidônico, um dos produtos da reação catalisada pelas PLA2s, possui grande interesse em pesquisas pelo fato de poderem ser modificados em compostos com atividade anti-inflamatória [Dennis, 1994].

Enzimas PLA2 são bastante encontradas em tecidos de mamíferos, bem como em venenos de insetos, aracnídeos e cobras, sendo inclusive a maior parte dos componentes tóxicos presentes em venenos de serpentes [Nicolas et al., 1997; Kini, 2005]. Devido ao aumento da presença e atividade da PLA2 resultante de uma picada de cobra, inseto ou aracnídeo, o ácido araquidônico é liberado de forma desproporcional pelos fosfolipídeos que compõe a membrana celular, resultando em inflamação e dor no local [Argiolas & Pisano, 1983].

As fosfolipases A2 incluem diversas famílias não relacionadas, porém com atividade enzimática em comum. Neste caso de uso, utilizamos o AMS PF00068, extraído do PFAM, que incluem apenas as PLA2s secretadas. Estas enzimas precisam do Ca^{2+} para sua atividade [Six & Dennis, 2000].

A bateria de testes realizada neste caso de uso foi feita de forma semelhante ao caso anterior, porém como neste caso não há mais de uma subclasse discretizada, foi construído apenas 6 redes. Três utilizando filtragem por uma sequência de referência *PA2G5_HUMAN* e três utilizando o perfil HMM, sendo que para ambos os casos utilizando os diferentes tipos de AMS disponíveis no PFAM (*full*, *uniprot* e *ncbi*). Os parâmetros para filtragem, montagem e decomposição da rede também foram idênticos ao caso de uso anterior, com uma única exceção do escore mínimo para considerar uma correlação de 10 e não 5. A tabela 4.2 mostra as comunidades encontradas para cada uma das seis redes.

Pla2/full	Pla2/uniprot	Pla2/ncbi	Hmm/full	Hmm/uniprot	Hmm/ncbi
G49 **G51** **C63** **H67** **D68** **C117**	**G49** **G51** **C63** **H67** **D68** **C117**	**G49** **G51** **C63** **H67** **D68** **C117**	**G49** **G51** **C63** **H67** **D68** **C117**	**G49** **G51** **C63** **H67** **D68** **C117**	**G49** **G51** **C63** **H67** **D68** **C117**
L25 **L114** **L118**	**L25** **L114** **L118** **I29**	**L25** **L114** **L118** **I29**	**L25** **L114** **L118**	**L25** **I29** **L114** **L118**	**L22** **L25** **D24** **L114** **L118** **I29**
K31 **F90**	**K31** **F90**	**K31** **F90**	**K31** **F90**	**K31** **F90**	**K31** **F90**
H69 G78	T60 H69	T60 H69 N121	H69 G78		T60 H69 N121 R82 S124

Tabela 4.2: Comunidades geradas para a família PF00068. A numeração utilizada se refere a sequência *PA2G5_HUMAN*.

Novamente, as redes geradas não obtiveram grande variação. A comunidade 1, representada na cor verde, foi retratada de forma idêntica em todas as seis redes. Já a comunidade 2, de azul, obteve três leucinas predominantes: L25, L114 e L118, uma isoleucina que apareceu em quatro das seis redes e mais dois resíduos que apareceram apenas na rede *hmm/ncbi*. A comunidade 3, na cor vermelha, também foi representada da mesma forma em todas as seis redes, formada pelo par K31 e F90. E por fim, a comunidade quatro foi a que mais variou, chegando a nem ser identificada na rede

hmm/uniprot. A figura 4.4 ilustra graficamente a rede gerada usando filtragem por perfil hmm e o banco ncbi.

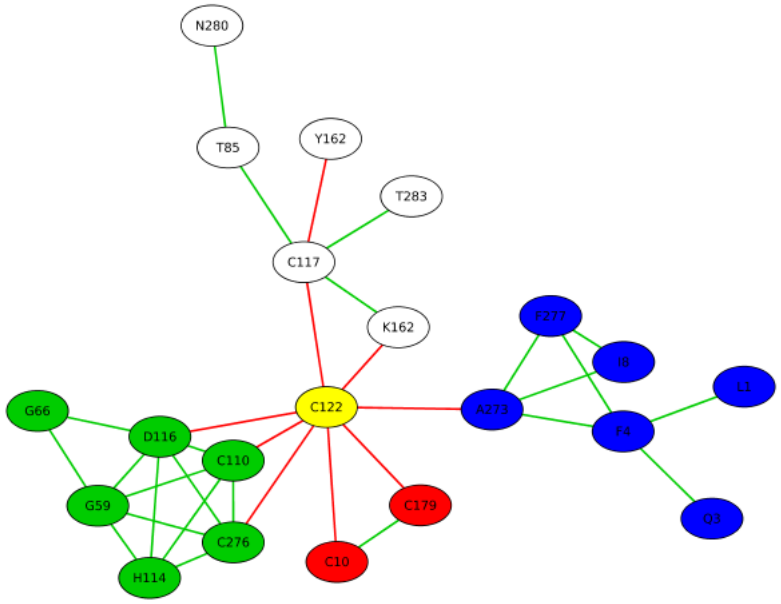
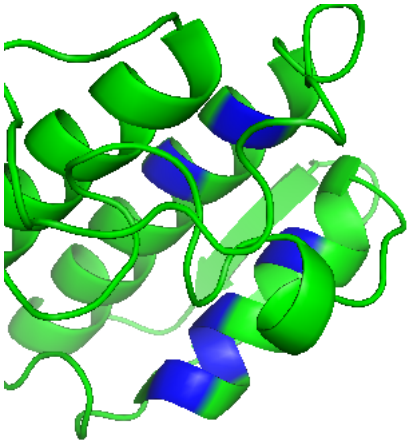


Figura 4.4: **Rede *hmm/ncbi*.** Arestas na cor verde indicam correlação positiva, vermelhas indicam correlação negativa. Está sendo utilizada a numeração do alinhamento.

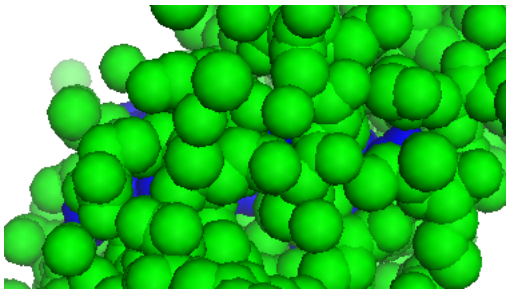
A comunidade 1, representada na cor verde tanto na tabela quanto na figura 4.4, se trata do agrupamento catalítico das PLA2s. A comunidade é formada por duas cisteínas que fazem uma ponte dissulfeto, uma histidina do sítio ativo e mais duas glicinas e um aspartato que além de serem resíduos do sítio de ligação de cálcio, também fazem parte da rede catalítica [Scorr et al., 1991; Han et al., 1997].

A comunidade 2, representada na cor azul, se trata do núcleo hidrofóbico das PLA2s [Zhao et al., 1998]. Além de serem resíduos majoritariamente hidrofóbicos, é possível observar a localização dos seis resíduos preditos para esta comunidade na estrutura da *PA2GA_HUMAN* (1BBC) na figura 4.5.

A comunidade 3 está indicada na tabela 4.2 como sendo formada por uma lisina e uma fenilalanina, porém ao analisar a figura 4.4, a comunidade aparece como formada por duas cisteínas. Isto ocorre pelo fato de a sequência de referência utilizada para numerar na tabela ser uma exceção, ela não possui as duas cisteínas naquelas posições. Esta comunidade divide as fosfolipases em duas classes: as proteínas que possuem a ponte dissulfeto formada pelos resíduos C10 e C179 (figura 4.6a); e as que possuem uma interação cátion-pi formada pelo par K10 e F179 (figura 4.6b). As interações cátion-pi desempenham um papel importante na natureza, principalmente em relação



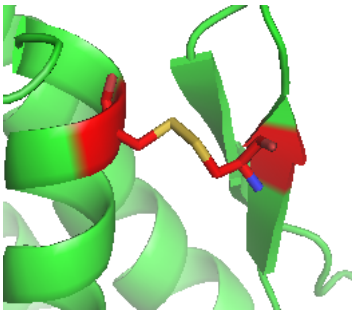
(a) Estrutura visualizada na representação *cartoon*



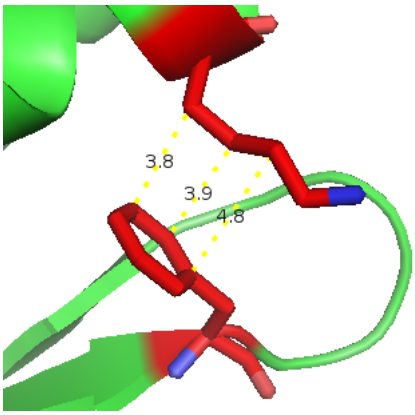
(b) Estrutura visualizada na representação *spheres*

Figura 4.5: **Comunidade 2 representada na estrutura da *PA2GA_HUMAN* (1BBC).** Os resíduos da comunidade 2 na rede *hmm/ncbi* (maior gerada) estão marcados na cor azul.

ao enovelamento da proteína e na catálise enzimática [Ma & Dougherty, 1997]. Das 91 sequências que possuem o par K10 e F179, cerca de 85% pertecem a família das víboras, *Viperidae*, já em relação a ponte dissufeto, está presente em 611 sequências do AMS ncbi, a maior parte destas sequências, cerca de 36%, são também de serpentes, porém da família *Elapidae*, ao contrário da subclasse dos resíduos K10 e F179, esta ainda conta com proteínas de aves, mamíferos e outros taxons. Uma hipótese para isto é que a presença deste cátion-pi pode estar ligada a especificidade do veneno das víboras em relação as *elapidaes*.



(a) **Comunidade 3 representada na estrutura da *PA2GA_HUMAN* (1BBC).** As duas cisteínas marcadas na cor vermelha fazem uma ponte dissulfeto.



(b) **Comunidade 2 representada na estrutura da *PA2B_DABRR* (1VIP).** Interação cátion-pi formada por uma lisina e uma fenilalanina.

A comunidade 4 parece ser formada por efeito de conservação, os cinco resíduos

que aparecem na rede *hmm/ncbi* possuem alta taxa de frequência. Cerca de 80% das sequências possuem ao menos quatro destes resíduos, pelo fato de ser um grupo de resíduos muito comum no AMS, podem ter sido unidos na geração da rede simplesmente por este motivo. Além do mais, estes resíduos se encontram espalhados pela estrutura e nada foi encontrado na literatura e bancos de dados os unindo.

Ao analisar a rede na figura 4.4, um outro resíduo que não está em nenhuma comunidade chama atenção. A cisteína 122 forma um *hub* de anti-correlações com todas as quatro comunidades, um comportamento atípico para uma família sem subclasses distintas. A ocorrência desta cisteína na posição 122 parece reduzir a ocorrência dos resíduos que formam as quatro comunidades. Ao separar as sequências que possuem a C122 no alinhamento do *ncbi*, sobra-se 219 sequências, sendo 217 relacionadas a espécie *Oikopleura dioica*. Foi então realizado uma comparação dos valores de aderência de um conjunto de PLA2s com um de *Oikopleura dioicas* para cada uma das quatro comunidades encontradas na rede *hmm/ncbi*. Como já era esperado após observar a rede, as *Oikopleura dioicas* não estão relacionadas à nenhuma das quatro comunidades. É possível ver o gráfico de aderência na figura 4.7.

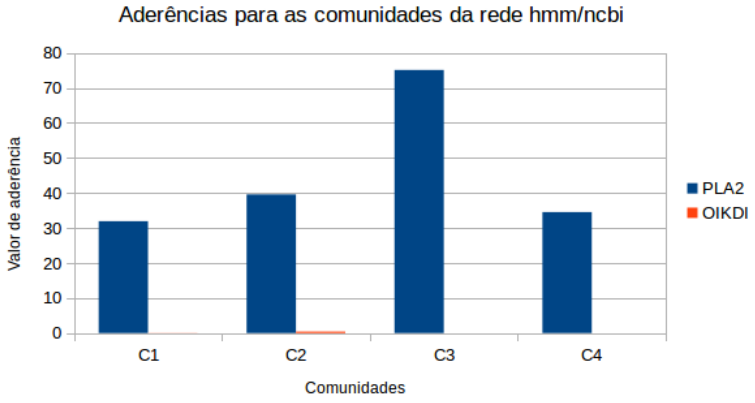


Figura 4.7: **Aderência entre PLA2s experimentalmente observadas com sequências da espécie *Oikopleura dioicas*.**

4.3 Proteínas reguladoras de nitrogênio P-II

A família das proteínas reguladoras de nitrogênio P-II (PII) é formada por um conjunto de proteínas de sinalização procarióticas homólogas, compostas por cerca de 110 aminoácidos extremamente bem conservados, e que estão envolvidas na regulação do metabolismo do nitrogênio [Conroy et al., 2007].

Proteínas PII possuem a capacidade de atuar em transduções de sinais afim de regular a atividade de uma enorme variedade de proteínas alvos através de interações proteína-proteína. Estes alvos incluem proteínas de membrana, enzimas e fatores de transcrição, sendo na grande maioria dos casos envolvidos em algum aspecto do metabolismo do nitrogênio celular [Huergo et al., 2013]. As proteínas são pós-translacionalmente modificadas através da adição de um grupo *uridil* pela enzima *uridiltransferase*. A presença ou ausência deste grupo modula o comportamento da proteína.

A abordagem de testes realizada para essa família foi diferente das anteriores. Neste caso foi utilizado apenas um AMS, que foi o PF00543 (Uniprot), extraído do PFAM. Além disso, também foi aplicado uma única abordagem de filtro, sendo a filtragem pelo perfil HMM, com 80% de cobertura e 70% de identidade máxima. O objetivo neste caso foi observar a geração de diferentes redes , a partir do mesmo alinhamento com diferentes alfabetos. Baseado no AMS filtrado, foi gerado 19 redes, uma para cada redução de alfabeto disponível no PFStats. Os parâmetros para criação das redes foram escore mínimo de 30, sub-alinhamento mínimo de 20% e variação de frequência de 20%. Os resultados obtidos podem ser observados na tabela 4.3.

T20	M15	11IM	Li10	M10	M8	T6	T5	5IM	Li5	W5	W5v	Li4	M4	3IM	W3	Li3	T2	M2
Q86 G89 R115	Q86 G89 K115	H278 H281 H263 S282	I164 E278 E281 S282	K278 K281 K263 S282	K278 K281 K263 S282	K278 K281 T282	T86 G89 K182 K171	M278 M281 C282	I164 E278 E281 S282	I11 K278 K15 I164 A282 K281	A68 K171 E182	I164 E278 E281 S282	L169 E171 E182	G68 D171 I169 D182	I18 I58	I164 E278 E281 S282	Y279 P281 P282 P283	P164 E278 E281 P282
R281 T282	K278 K281 T282	N86 G89 H115	I11 E15 E115	G89 K115 F111	L169 E182 K171	T86 G89	K278 K281 T282		I11 E15 E115	K115 K171 I169 E182	K278 K281 A282	I11 E15 E115	E278 E281 A282	I164 D278 D281 G282	A281 A282	I11 E15 E115	Y18 Y58 P242	P169 E171 E182
	K171 E182	A169 D182 H171	I169 E171 E182	L169 E182 K171	F111 K115	K171 D182			I169 E171 E182	I18 I58 E261		I169 E171 E182	E15 E115	D15 D115		I169 E171 E182	P68 P171 Y169	E15 E115
	D161 K263	S242 G248 D261	I18 I58	S242 G248		D261 K263			I18 I58			I18 I58				I18 I58	Y11 P15	P171 P182

Tabela 4.3: Comunidades geradas para a família PF00543. A numeração utilizada se refere a posição no alinhamento. Resíduos representados na cor verde fazem parte do sítio de ligação de nucleotídios, enquanto que os resíduos na cor azul estão todos numa mesma folha beta.

É importante salientar que a utilização de valores mais altos de escore mínimo irá aumentar o tamanho de redes de alfabetos de tamanho médio. Isso ocorre pelo fato de que quanto maior o número de caracteres utilizados para representar um alfabeto, mais correlações podem ser encontradas, porém de menor força. O contrário ocorre para alfabetos muito pequenos, poucas correlações serão encontradas porém com um

escore muito grande. O objetivo principal deste caso de uso não é dizer qual é o melhor alfabeto a ser utilizado, mas sim como a utilização de diferentes representações podem ainda produzir redes informativas. Ao utilizar o escore mínimo de 10 para a rede T20 (alfabeto completo), a rede gerada se mantém com duas comunidades, contudo maiores. Comunidade 1 formada por L18, T242, G248, D261, K263, I264, R281 e T282; comunidade 2 formada por: Q86, G89, R115.

Os resíduos representados na cor verde na tabela 4.3 constituem-se daqueles que estão indicados nos bancos de dados Uniprot ou NCBI como aminoácidos do sítio de ligação de nucleotídios (ECO:0000250). Já na cor amarela, está indicado os resíduos de uma mesma comunidade que também fazem parte de uma mesma estrutura secundária, no caso uma folha beta. Esse tipo de informação é gerado automaticamente pelo PFStats através da busca nos repositórios. Um fato interessante de ser observado nessa tabela é que todas as redes, mesmo aquelas com alfabeto de 2 caracteres, separaram relativamente bem os diferentes grupos obtidos. Outro dado importante, é que ao analisar todas as redes como um todo, alguns padrões de conectividade se tornam perceptíveis, por exemplo, as redes Murphy15 ou 11IMGT possuem três comunidades distintas de ligação de nucleotídios, e ao comparar com as outras redes, tais grupos, como as posições 86, 89 e 115; ou 278, 281 e 282; tendem a aparecer juntos. Observando tais grupos na estrutura, é perceptível que a proteína realmente possui três loops de ligação de nucleotídios, como se pode ver na figura 4.8.

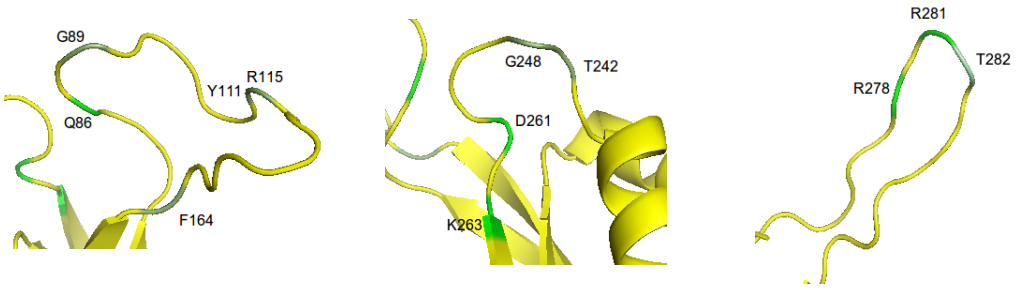


Figura 4.8: **Comunidades do sítio de ligação de nucleotídios representada na estrutura da *GLNB_ECOLI* (2PII).** Os resíduos agrupados pela decomposição das redes estão indicados em duas tonalidades de verdes. Os mais claros representam posições já ditas como sítio de ligação nos bancos de dados biológicos. No tom acinzentado estão os outros resíduos classificados que participam do *loop* de ligação.

Na figura 4.9 é possível visualizar os três resíduos representados na cor azul na tabela 4.3 compondo uma mesma folha beta. A T68 aparece inclusa nesta comunidade nas redes T2, 3IMGT e Wang5v, este resíduo faz parte de uma segunda folha beta paralela e pode ser visualizado num tom mais fraco de azul.

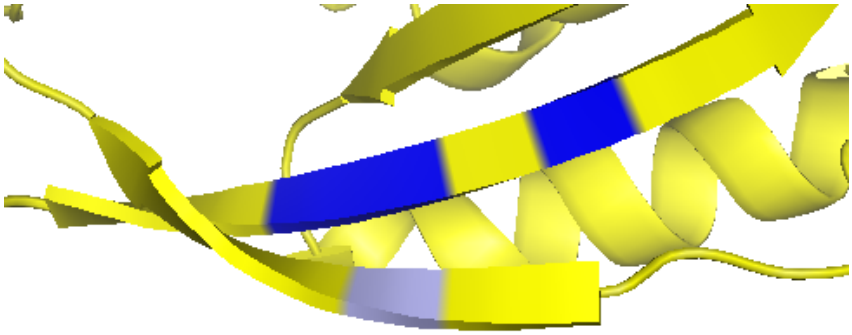


Figura 4.9: **Comunidade de uma folha beta representada na estrutura da *GLNB_ECOLI* (2PII).**

Para os outros grupos de resíduos como: 11, 15 e 115; e 18 e 58; não foi encontrado características funcionais ou estruturais relevantes.

4.4 Domínio de ligação de DNA dos receptores nucleares tipo 4

Os receptores nucleares (RNs) constituem uma classe de proteínas que se encontram no interior das células responsáveis pela detecção de moléculas de ácidos graxos, ácidos biliares, hormônios esteroides, tireoides e vitaminas A e D [Liu et al., 2015]. Em resposta, estes receptores trabalham em conjunto com outras proteínas afim de regular a expressão de genes específicos, controlando o desenvolvimento, homeostase e o metabolismo do organismo [Evans, 1988]. São classificados como fatores de transcrição por sua capacidade de se ligar diretamente ao DNA, regulando a expressão de genes adjacentes.

Uma propriedade única dos receptores nucleares é sua capacidade de interagir diretamente e controlar a expressão de DNA genômico. Como consequência, os mesmos desempenham papéis fundamentais tanto no desenvolvimento embrionário quanto na homeostase adulta [Olefsky, 2001].

Todos os receptores nucleares possuem algumas características estruturais em comum que incluem: um domínio N-Terminal variável (domínio A/B) contendo um subdomínio de função ativada 1, seguido de um domínio de ligação de DNA extremamente conservado (DBD) composto por outros dois subdomínios dedo de zinco, seguido pela região *Hinge* (domínio flexível que conecta o DBD com o LDB, também chamado de domínio E), um segundo domínio variável, um domínio também altamente conservado de ligação com o ligante e por fim alguns receptores nucleares também possuem

um domínio C-terminal variável, também chamado de domínio F [Olefsky, 2001]. Neste caso de uso foi utilizado o alinhamento de código PF00105 extraído do PFAM. Esse AMS além de focar no domínio específico de ligação de DNA, também se limita a um tipo de específico de RNs, denominado receptores nucleares de tipo 4. Estes receptores podem tanto ligar DNA como dímeros quanto como monômeros, porém apenas um único domínio de ligação de DNA se liga a um único sítio de HRE (elementos de resposta hormonal, pequenas sequência do DNA capaz de interagir com receptores hormonais específicos) [Novac & Heinzel, 2004].

Neste caso de uso as análises foram novamente focadas nas reduções de alfabeto, porém também variando parâmetros de criação das redes. Foi utilizado apenas um alinhamento, extraído do PFAM (PF00105 - Uniprot), a filtragem foi realizada pelo perfil HMM, por 80% de cobertura e 70% de identidade máxima. Alguns modelos foram filtrados após a aplicação da redução de alfabeto. Foram então geradas 130 redes, conforme os modelos ditos na tabela 4.4, para cada modelo foi gerada uma rede para cada metodologia de redução de alfabeto.

Modelo	Escore Mínimo	Subalinhamento Mínimo	Variação de Frequência	Reduções de Alfabeto
A	10	20%	30%	Após a filtragem
B	10	20%	40%	Após a filtragem
C	5	15%	50%	Após a filtragem
D	10	20%	30%	Antes da filtragem
E	10	20%	40%	Antes da filtragem
G	5	15%	50%	Antes da filtragem

Tabela 4.4: Modelos para geração das redes de coevolução.

Com tantas redes para serem estudadas, o passo seguinte consistiu de uma busca por padrões de conectividade. Buscar quais resíduos tendem a aparecer na mesma comunidade dentre todos os 130 grafos. Chegou-se então em um padrão de duas comunidades, sendo a primeira formada pelos resíduos: G8, D12, K16, A21, S24, G33, Y47, E81, G90, K106, G122, K202, Y227, I315, K332, R378, N384 e Q405; já a segunda é composta por Y414, K425, C426, L427, G432, M433 e K434. É possível observar uma das redes na figura 4.10.

Segundo Afonso et al., os *p-boxes* em RNs humanos possuem um padrão de sequência formado por CE153G154CK156G157 (numeração da hRXRA), com exceção para os receptores da classe *3-ketosteroid* (receptores glicocorticóides, mineralocorticóides, progesterona e andrógenos), que apresentam o padrão de sequência CG153S154CK156V157. Estas posições 153, 154, 156 e 157 na numeração do alinha-

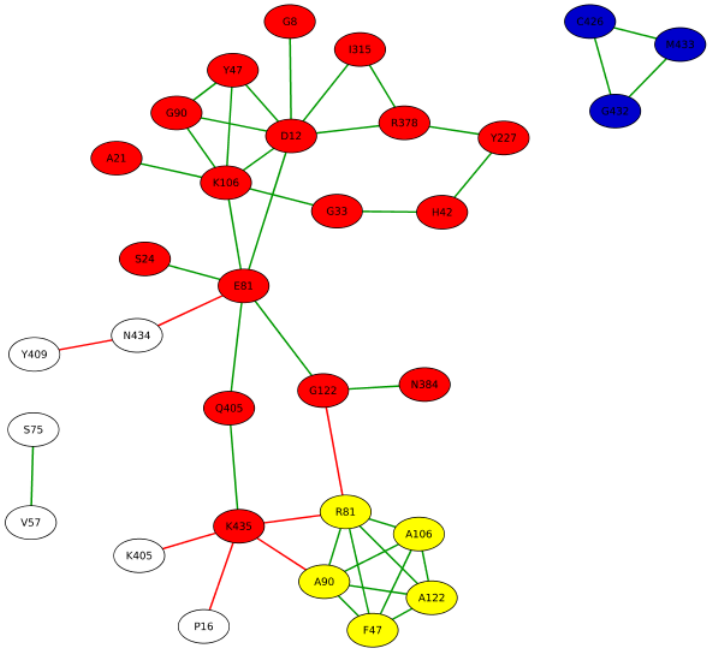


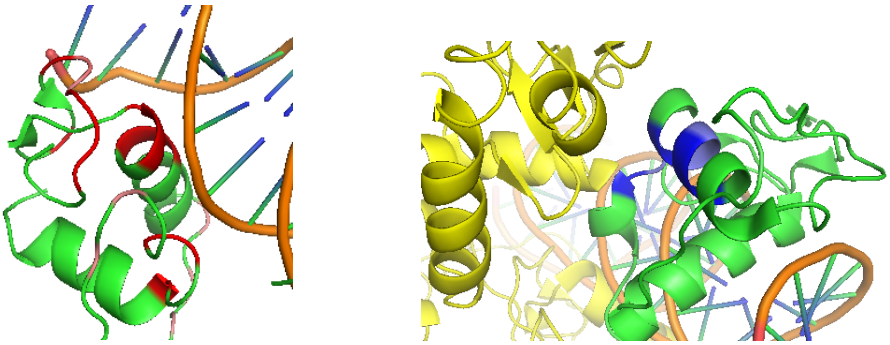
Figura 4.10: **Rede de coevolução construída pelo modelo C com alfabeto completo (T20).** Os vértices coloridos são aqueles que acabaram derivando as comunidades virtuais por padrões de conectividade de todas as redes.

mento, utilizada na rede da figura 4.10, se refere a 81, 90, 106 e 122, posições que apareceram na comunidade 1 geral e na comunidade 3 dos alinhamentos com alfabeto completo (T20). As duas cisteínas existentes nos dois padrões de sequências citados são estritamente conservadas em todos os DBDs, uma vez que são necessárias para ligar o zinco no primeiro dedo de zinco. Já as posições 153, 154 e 157 na numeração do RXRA são chamadas de primeira, segunda e terceira, respectivamente, posição da *p-box* e experimentos de mutagênese de sítio dirigido mostraram que substituição de amino ácido em alguma destas três posições pode causar efeitos profundos na sequência especifica de ligação destes receptores ao DNA [Nelson et al., 1995; Glass, 1994].

A comunidade formada pelos resíduos R81, A106, A122, F47 e A90 (representada com vértices na cor amarela na figura 4.10), representa um caso onde o uso de redução de alfabeto pode causar perda de informação, pois tais posições também estão contidas na comunidade maior, porém, com aminoácidos diferentes. Esta comunidade apareceu apenas nas redes de alfabeto completo. Um fato interessante acerca dessa comunidade é que ela obteve grau zero de aderência a todas as 10 subclasses que estudadas nesta análise. Porém ao analisar as sequências que apresentam pelo menos quatro destes seis resíduos, é obtido uma subclasse com 941 sequências, sendo cerca de 84% do filo

Nematoda. Afonso et al. demonstraram que estes resíduos estão relacionados a um P-Box especifico dos *nematodas*, presente principalmente nas espécies *Caenorhabditis elegans* e *Caenorhabditis briggsae*. A divergência nesse *p-box* provavelmente reflete interações a elementos responsivos diferentes.

Por se tratar de um alinhamento específico do domínio do sítio de ligação, estes resíduos apareceram bastante conectados nas redes. O domínio em questão também contém dois subdomínios de *zinc finger*, porém não foram constatada conectividades relevantes separando estes dois grupos. A primeira das duas comunidades virtuais se trata do sítio de ligação com nucleotídios. Grande parte dos resíduos inclusos nesta comunidade já são descritos como membros do sítio ou estão em posições estratégicas na estrutura. Já os resíduos da segunda comunidade estão localizados no fundo da proteína, numa região de interface de dimerização entre as proteínas da subclasse RXRA e PPAR. A figura 4.11 demonstra as duas comunidades representadas na estrutura tridimensional da proteína. Em 4.11a, é possível observar os resíduos do sítio de ligação de nucleotídios interagindo com uma molécula de DNA, enquanto que em 4.11b está ilustrada a interação dimérica entre uma proteína RXRA e uma PPAR. Os resíduos representados com uma coloração mais forte consistem de posições já definidas como sítio nos bancos de dados.



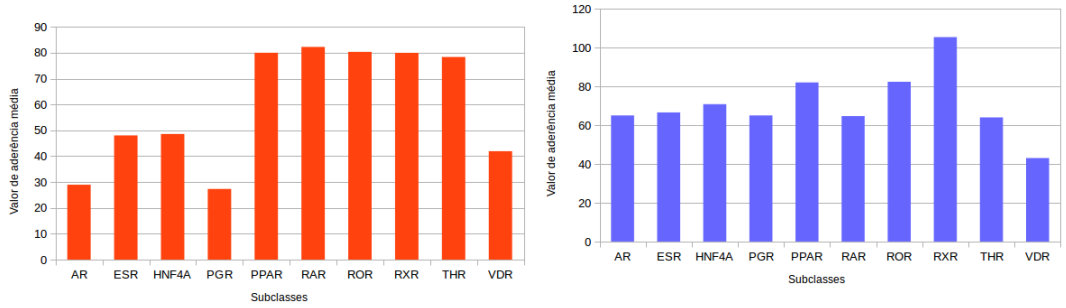
(a) Sítio de ligação da proteína RARA_HUMAN interagindo com uma molécula de DNA (2NLL).

(b) Heterodímero entre uma molécula RXRA e uma PPAR (3DZU)

Figura 4.11: **Comunidades do sítio de ligação de nucleotídios (vermelho) e de dimerização (azul) representadas na estrutura tridimensional.**

Por fim, foi realizada uma análise de aderência para ver como cada uma das 10 subclasses desta família compartilham as duas comunidades analisadas e o resultado pode ser visto na figura 4.12. Ambas as comunidades possuem aderência considerável para todas as subclasses, a comunidade 1 possui duas subclasses, AR e PGR, com valores médio de aderência abaixos da média, mas ainda sim consideráveis. O motivo provável dessa divergência pode ser observado na rede (figura 4.10), uma comunidade

de cinco resíduos anti-correlacionada com a comunidade 1. Estes cinco resíduos se tratam de uma variação de aminoácidos que as posições 47, 81, 90, 106 e 122 podem assumir. Já para a comunidade dois, os valores médios de aderência permaneceram equilibrados, com um salto maior para as RXRs, que consiste na subclasse que realiza a heterodimerização nestes resíduos.



(a) Comunidade do sítio de ligação.

(b) Comunidade do sítio de heterodimerização.

Figura 4.12: **Aderências médias das comunidades identificadas para cada subclasse da família.**

Capítulo 5

Considerações Finais

5.1 Conclusões

O objetivo deste trabalho foi desenvolver um sistema computacional para auxiliar no estudo e compreensão de proteínas a partir de dados evolutivos. O PFStats atualmente contém métodos para: filtragem de AMS, calcular conservação de resíduos, decomposição de redes de coevolução de resíduos, seleção de sequências por características, obtenção de alinhamento automático de PDB e sequência, cálculo do subalinhamento mínimo representativo, busca por anotação de resíduo automático, reduções de alfabetos, além de visualizações de resultados. Além disto, o *software* está disponível para os sistemas operacionais Windows e Linux, nas versões com interface gráfica, de fácil aprendizagem e utilização, e nas versões para linha de comando, para que possa ser utilizado em lotes e por servidores.

Foi realizado uma bateria de testes para 4 famílias de proteínas, visando observar padrões nas gerações de redes de coevolução, modificando parâmetros, alinhamentos de entrada e com o uso de alfabetos reduzidos. Chamamos de comunidade chave aquela, cujo seu papel biológico na proteína pode ser identificado através de consulta a bancos de dados e a literatura e não há presença de aleatoriedade, resíduos completamente fora do contexto.

A começar pela filtragem, tanto o método por sequência de referência quanto por perfil HMM levaram a conclusões próximas, após a decomposição em comunidades. No caso das fosfolipases A2, talvez pelo fato da rede não possuir subclasses muito distintas, as comunidades geradas para cada rede foram bem semelhantes. Já nas lisozimas C, as redes geradas foram um pouco mais divergentes. Porém, as diferenças não foram o suficiente para alterar ou identificar um novo papel biológico para a comunidade.

Outro parâmetro comparado neste trabalho são os diferentes tipos de alinhamen-

tos disponíveis no PFAM. Estes ocasionaram uma variação maior nas comunidades identificadas em cada rede, do que a abordagem de filtro. Em geral, as redes que utilizaram o AMS *full* identificaram menos resíduos de comunidade chave. Já as redes com *Uniprot* e *NCBI* obtiveram resultados satisfatórios nos testes realizados.

Um fato instigante é que todos os alfabetos reduzidos foram capaz de identificar resíduos comunidades chaves, mesmo aqueles alfabetos extremamente reduzidos, com dois caractere. Em muitos casos, alfabetos de 8 a 15 caracteres identificaram um número maior de comunidades chaves do que o alfabeto completo. Logo, atribuímos a nossa metodologia a utilização diversas redes com diferentes alfabetos e buscar por padrões de conectividade entre todas elas.

A ocorrência de um *hub* de anti-correlações nas redes da fosfolipase A2 permitiu a identificação de uma subclasse especifica na família, formada por 217 sequências da espécie *Oikopleura dioica* e que não possui nenhum dos resíduos responsáveis pela atividade catalítica nem do sítio de ligação.

Outra subclasse foi descoberta através da observação de conectividade das anti-correlações na rede. As redes com o alfabeto completo de 20 caracteres dos receptores nucleares IV identificou duas comunidades conectadas por algumas arestas de anti-correlação, ou seja a presença de uma destas comunidades tende a ocorrer na ausência da outra. O interessante foi que estas duas comunidades possuem as mesmas posições, porém com uma variação entre os amino ácidos, logo a utilização de alfabetos reduzidos faz com que as comunidades se unam e perca essa informação.

5.2 Trabalhos Futuros

Atualmente estamos com muitas ideias para continuidade deste trabalho. No quesito de aplicação, existe o plano de passar todo o sistema para a *web*, numa arquitetura de *webservice*. Além disso, há também uma ideia de gerar um banco de dados de comunidades e anotações para todo o banco do PFAM.

Em relação a metodologia, existe a intenção de substituir as abordagens de filtro e subalinhamento mínimo por uma metodologia de atribuição de pesos. Outro desejo é o de acrescentar outros algoritmos para decomposição em comunidades que não seja por componentes conexos. A decomposição por componentes conexos funciona bem para boa parte dos alinhamentos, porém, casos de complexidade muito grande como: alinhamentos com sequências muito longas, presença de grande quantidade de subclasses funcionais, presença de vários subdomínios, entre outros; podem levar a redes de complexidade muito alta incapaz de se dividir em componentes conexos sem grandes

perdas de informações.

Além da observação de padrões de comunidades, há também a vontade de estudar outros padrões nesse tipo de rede, como métricas de sistemas complexos, ocorrência de *hubs*, percolação e cliques, além da direcionalidade da rede.

Referências Bibliográficas

Afonso, M. Q. L.; de Lima, L. H. & Bleicher, L. (2013). Residue correlation networks in nuclear receptors reflect functional specialization and the formation of the nematode-specific p-box. *BMC genomics*, 14(Suppl 6):S1.

Aftabuddin, M. & Kundu, S. (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophysical journal*, 93(1):225--231.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Andraos, J. (2008). Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws-new methods based on directed graphs. *Canadian Journal of Chemistry*, 86(4):342--357.

Anfinsen, C. B. (1972). Studies on the principles that govern the folding of protein chains.

Anfinsen, C. B. & Corley, L. G. (1969). An active variant of staphylococcal nuclease containing norleucine in place of methionine. *Journal of Biological Chemistry*, 244(19):5149--5152.

Argiolas, A. & Pisano, J. J. (1983). Facilitation of phospholipase a2 activity by mastoparans, a new class of mast cell degranulating peptides from wasp venom. *Journal of Biological Chemistry*, 258(22):13697--13702.

Bar-Yam, Y. (2002). General features of complex systems. *Encyclopedia of Life Support Systems (EOLSS), UNESCO, EOLSS Publishers, Oxford, UK.*

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509--512.

Barabási, A.-L. & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5):50--59.

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

Betts, M. J. & Russell, R. B. (2007). Amino-acid properties and consequences of substitutions. *Bioinformatics for geneticists*, 2:311--342.

Bleicher, L.; Lemke, N. & Garratt, R. C. (2011). Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. *PloS one*, 6(12):e27786.

Brinda, K.; Vishveshwara, S. & Vishveshwara, S. (2010). Random network behaviour of protein structures. *Molecular Biosystems*, 6(2):391--398.

Chagoyen, M.; García-Martín, J. A. & Pazos, F. (2015). Practical analysis of specificity-determining residues in protein families. *Briefings in bioinformatics*, p. bbv045.

Chakraborty, A. & Chakrabarti, S. (2014). A survey on prediction of specificity-determining sites in proteins. *Briefings in bioinformatics*, p. bbt092.

Chan, H. S. (1999). Folding alphabets. *Nature structural biology*, 6(11).

Chatzou, M.; Magis, C.; Chang, J.-M.; Kemena, C.; Bussotti, G.; Erb, I. & Notredame, C. (2015). Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, p. bbv099.

Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R. & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688.

Chou, K. & Forsén, S. (1980). Graphical rules for enzyme-catalysed rate laws. *Biochemical Journal*, 187(3):829--835.

Chou, T.-C. (2010). Drug combination studies and their synergy quantification using the chou-talalay method. *Cancer research*, 70(2):440--446.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377--387.

Conroy, M. J.; Durand, A.; Lupo, D.; Li, X.-D.; Bullough, P. A.; Winkler, F. K. & Merrick, M. (2007). The crystal structure of the escherichia coli amtb–glnk complex reveals how glnk regulates the ammonia channel. *Proceedings of the National Academy of Sciences*, 104(4):1213--1218.

Consortium, U. et al. (2008). The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl 1):D190--D195.

Csermely, P. (2008). Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends in biochemical sciences*, 33(12):569--576.

Dagum, L. & Enon, R. (1998). Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering, IEEE*, 5(1):46--55.

Davies, G. & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure*, 3(9):853--859.

Davis, B. K. (2002). Molecular evolution before the origin of species. *Progress in biophysics and molecular biology*, 79(1):77--133.

de Lima Morais, D. A.; Fang, H.; Rackham, O. J.; Wilson, D.; Pethica, R.; Chothia, C. & Gough, J. (2010). Superfamily 1.75 including a domain-centric gene ontology method. *Nucleic acids research*, p. gkq1130.

Dennis, E. A. (1994). Diversity of group types, regulation, and function of phospholipase a2. *Journal of Biological Chemistry*, 269:13057--13057.

Dima, R. I. & Thirumalai, D. (2006). Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Science*, 15(2):258--268.

Ding, Y.; Wang, X. & Mou, Z. (2015). Communities in the iron superoxide dismutase amino acid network. *Journal of theoretical biology*, 367:278--285.

Do, C. B. & Katoh, K. (2008). Protein multiple sequence alignment. *Functional Proteomics: Methods and Protocols*, pp. 379--413.

Durbin, R.; Eddy, S. R.; Krogh, A. & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eom, Y.-H. & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PloS one*, 6(9):e24926.

Erdős, P. & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290--297.

Erdos, P. & Rényi, A. (1961). On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38(4):343--347.

Evans, R. (1988). The steroid and thyroid hormone receptor superfamily. *Science*, 240(4854):889--895. ISSN 0036-8075.

Ferreira, R. R. (2005). Introdução a bioinformática. *Centro de Biologia Genômica e Molecular da Universidade Federal do Rio Grande do Sul. Disponível em:< http://www. inf. ufrgs. br/˜ rrferreira/bioinf/Apresentacoes/introducaoBioinf. pdf>. Acesso em*, 10.

Filizola, M.; Olmea, O. & Weinstein, H. (2002). Prediction of heterodimerization interfaces of g-protein coupled receptors with a new subtractive correlated mutation method. *Protein engineering*, 15(11):881--885.

Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A. et al. (2015). The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, p. gkv1344.

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

George, R. A. & Heringa, J. (2002). An analysis of protein domain linkers: their classification and role in protein folding. *Protein Engineering*, 15(11):871--879.

Gibson, S. M.; Ficklin, S. P.; Isaacson, S.; Luo, F.; Feltus, F. A. & Smith, M. C. (2013). Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PLoS One*, 8(2):e55871.

Glass, C. K. (1994). Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers*. *Endocrine reviews*, 15(3):391--407.

Göbel, U.; Sander, C.; Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309--317.

Guimera, R.; Mossa, S.; Turtschi, A. & Amaral, L. A. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794--7799.

Haft, D. H.; Selengut, J. D.; Richter, R. A.; Harkins, D.; Basu, M. K. & Beck, E. (2012). Tigrfams and genome properties in 2013. *Nucleic acids research*, p. gks1234.

Hall, L. & Campbell, P. (1986). Alpha-lactalbumin and related proteins: a versatile gene family with an interesting parentage. *Essays in biochemistry*, 22:1.

Han, S. K.; Yoon, E. T.; Scott, D. L.; Sigler, P. B. & Cho, W. (1997). Structural aspects of interfacial adsorption a crystallographic and site-directed mutagenesis study of the phospholipase a2 from the venom of agkistrodon piscivorus piscivorus. *Journal of Biological Chemistry*, 272(6):3573--3582.

Hanson, S.; Mastrovito, D.; Hanson, C.; Ramsey, J. & Glymour, C. (2016). Scale-free exponents of resting state provide a biomarker for typical and atypical brain activity. *arXiv preprint arXiv:1605.09282*.

Heim, E. N.; Marston, J. L.; Federman, R. S.; Edwards, A. P.; Karabadzhak, A. G.; Petti, L. M.; Engelman, D. M. & DiMaio, D. (2015). Biologically active lil proteins built with minimal chemical diversity. *Proceedings of the National Academy of Sciences*, 112(34):E4717--E4725.

Huergo, L. F.; Chandra, G. & Merrick, M. (2013). Pii signal transduction proteins: nitrogen regulation and beyond. *FEMS microbiology reviews*, 37(2):251--283.

Iserte, J.; Simonetti, F. L.; Zea, D. J.; Teppa, E. & Marino-Buslje, C. (2015). I-coms: Interprotein-correlated mutations server. *Nucleic acids research*, 43(W1):W320--W325.

Ito, Y.; Kuroki, R.; Ogata, Y.; Hashimoto, Y.; Sugimura, K. & Imoto, T. (1999). Analysis of a catalytic pathway via a covalent adduct of d52e hen egg white mutant lysozyme by further mutation. *Protein engineering*, 12(4):327--331.

Jeong, H.; Mason, S. P.; Barabási, A.-L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41--42.

Jones, D. T.; Singh, T.; Kosciolek, T. & Tetchner, S. (2015). Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999--1006.

Kimura, M. (1984). *The neutral theory of molecular evolution*. Cambridge University Press.

Kini, R. M. (2005). Structure–function relationships and mechanism of anticoagulant phospholipase a 2 enzymes from snake venoms. *Toxicon*, 45(8):1147--1161.

Krawczyk, M. J.; Muchnik, L.; Mańka-Krasoń, A. & Kułakowski, K. (2011). Line graphs as social networks. *Physica A: Statistical Mechanics and its Applications*, 390(13):2611--2618.

Kurochkina, N. & Choekyi, T. (2011). Helix–helix interfaces and ligand binding. *Journal of theoretical biology*, 283(1):92--102.

Lapedes, A. S.; Giraud, B. G.; Liu, L. & Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*, pp. 236--256.

Lee, B.-C.; Park, K. & Kim, D. (2008). Analysis of the residue–residue coevolution network and the functionally important residues in proteins. *Proteins: Structure, Function, and Bioinformatics*, 72(3):863--872.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of molecular biology*, 136(3):225IN1231--230IN2270.

Li, T.; Fan, K.; Wang, J. & Wang, W. (2003). Reduction of protein sequence complexity by residue grouping. *Protein Engineering*, 16(5):323--330.

Lin, S.-X. & Lapointe, J. (2013). Theoretical and experimental biology in one-a symposium in honour of professor kuo-chen chou's 50th anniversary and professor richard giegé's 40th anniversary of their scientific careers. *Journal of Biomedical Science and Engineering*, 6(4):435.

Litwin, S.; Jores, R.; Perelson, A. & Weisbuch, G. (1992). In theoretical and experimental insights into immunology. *Berlin and New York: Springer-Verlag*.

Liu, S.; Downes, M. & Evans, R. M. (2015). Metabolic regulation by nuclear receptors. Em *Innovative Medicine*, pp. 25--37. Springer.

Liu, X.; Liu, D.; Qi, J. & Zheng, W.-M. (2002). Simplified amino acid alphabets based on deviation of conditional probability from random background. *Physical Review E*, 66(2):021906.

Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295--299.

Ma, H. & Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270--277.

Ma, J. C. & Dougherty, D. A. (1997). The cation-π interaction. *Chemical reviews*, 97(5):1303--1324.

Manohar, P. & Singh, S. (2011). Protein sequence alignment: A review.

Munson, M.; Balasubramanian, S.; Fleming, K. G.; Nagi, A. D.; O'Brien, R.; Sturtevant, J. M. & Regan, L. (1996). What makes a protein a protein? hydrophobic core designs that specify stability and structural properties. *Protein Science*, 5(8):1584--1593.

Murphy, L. R.; Wallqvist, A. & Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149--152.

Nandi, C. L.; Singh, J. & Thornton, J. M. (1993). Atomic environments of arginine side chains in proteins. *Protein engineering*, 6(3):247--259.

Nelson, C. C.; Hendy, S. C. & Romaniuk, P. J. (1995). Relationship between p-box amino acid sequence and dna binding specificity of the thyroid hormone receptor. the effects of half-site sequence in everted repeats. *Journal of Biological Chemistry*, 270(28):16981--16987.

Nicolas, J.-P.; Lin, Y.; Lambeau, G.; Ghomashchi, F.; Lazdunski, M. & Gelb, M. H. (1997). Localization of structural elements of bee venom phospholipase a2 involved in n-type receptor binding and neurotoxicity. *Journal of Biological Chemistry*, 272(11):7173--7181.

NITTA, K. & SUGAI, S. (1989). The evolution of lysozyme and α-lactalbumin. *European Journal of Biochemistry*, 182(1):111--118.

Nitta, K.; Tsuge, H.; Sugai, S. & Shimazaki, K. (1987). The calcium-binding property of equine lysozyme. *FEBS letters*, 223(2):405--408.

Novac, N. & Heinzel, T. (2004). Nuclear receptors: overview and classification. *Current Drug Targets-Inflammation & Allergy*, 3(4):335--346.

O'Leary-Driscoll, S. (2015). What is bioinformatics?

Olefsky, J. M. (2001). Nuclear receptor minireview series. *Journal of Biological Chemistry*, 276(40):36863--36864.

Oliveira, L.; Paiva, A. C. & Vriend, G. (2002). Correlated mutation analyses on very large sequence families. *Chembiochem*, 3(10):1010--1017.

Ouzounis, C.; Pérez-Irratxeta, C.; Sander, C. & Valencia, A. (1997). Are binding residues conserved? Em *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 401--412.

Palla, G.; Derényi, I.; Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814--818.

Pandit, S. B.; Bhadra, R.; Gowri, V.; Balaji, S.; Anand, B. & Srinivasan, N. (2004). Supfam: a database of sequence superfamilies of protein domains. *BMC bioinformatics*, 5(1):28.

Pastor-Satorras, R. & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200.

Pazos, F. & Bang, J.-W. (2006). Computational prediction of functionally important regions in proteins. *Current Bioinformatics*, 1(1):15--23.

Percacci, R. & Vespignani, A. (2003). Scale-free behavior of the internet global performance. *The European Physical Journal B-Condensed Matter and Complex Systems*, 32(4):411--414.

Petersen, S. B.; Neves-Petersen, M. T.; Henriksen, S. B.; Mortensen, R. J. & Geertz-Hansen, H. M. (2012). Scale-free behaviour of amino acid pair interactions in folded proteins. *PloS one*, 7(7):e41322.

Plaxco, K. W.; Riddle, D. S.; Grantcharova, V. & Baker, D. (1998). Simplified proteins: minimalist solutions to the ?protein folding problem? *Current opinion in structural biology*, 8(1):80--85.

Pollner, P.; Palla, G. & Vicsek, T. (2005). Preferential attachment of communities: The same principle, but a higher level. *EPL (Europhysics Letters)*, 73(3):478.

Pommié, C.; Levadoux, S.; Sabatier, R.; Lefranc, G. & Lefranc, M.-P. (2004). Imgt standardized criteria for statistical analysis of immunoglobulin v-region amino acid properties. *Journal of Molecular Recognition*, 17(1):17--32.

Ramakrishnan, V. & White, S. W. (1992). The structure of ribosomal protein s5 reveals sites of interaction with 16s rrna. *Nature*, 358(6389):768--771.

Regan, L. & DeGrado, W. F. (1988). Characterization of a helical protein designed from first principles. *Science*, 241(4868):976--978.

Rego, N. & Koes, D. (2014). 3dmol. js: molecular visualization with webgl. *Bioinformatics*, p. btu829.

Sanger, F. & Tuppy, H. (1951a). The amino-acid sequence in the phenylalanyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4):463.

Sanger, F. & Tuppy, H. (1951b). The amino-acid sequence in the phenylalanyl chain of insulin. 2. the investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 49(4):481.

Schafmeisterll, C. E.; LaPorte, S. L.; Miercke, L. W. & Stroud, R. M. (1997). A designed four helix bundle protein with native?like structure.

Scorr, D. L.; WHITE, P. & BROWNING, L. (1991). Structures of free and inhibited human secretory phospholipase a2 from inflammatory exudate.

Singer, M. S.; Oliveira, L.; Vriend, G. & Shepherd, G. M. (1994). Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. *Receptors & channels*, 3(2):89--95.

Six, D. A. & Dennis, E. A. (2000). The expanding superfamily of phospholipase a 2 enzymes: classification and characterization. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1488(1):1--19.

Stein, L. D. (2003). Integrating biological databases. *Nature Reviews Genetics*, 4(5):337--345.

Stuart, D.; Acharya, K.; Walker, N.; Smith, S.; Lewis, M. & Phillips, D. (1986). α-lactalbumin possesses a novel calcium binding loop.

Sudhahar, S.; Veltri, G. A. & Cristianini, N. (2015). Automated analysis of the us presidential elections using big data and network analysis. *Big Data & Society*, 2(1):2053951715572916.

Thompson, J. D.; Higgins, D. G. & Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673--4680.

Tillier, E. R. & Charlebois, R. L. (2009). The human protein coevolution network. *Genome Research*, 19(10):1861--1871.

Uzuntarla, M.; Yilmaz, E.; Wagemakers, A. & Ozer, M. (2015). Vibrational resonance in a heterogeneous scale free network of neurons. *Communications in Nonlinear Science and Numerical Simulation*, 22(1):367--374.

Valdar, W. S. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, 48(2):227--241.

Valdar, W. S. & Thornton, J. M. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Bioinformatics*, 42(1):108--124.

Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current opinion in structural biology*, 12(3):368--373.

Van Deenen, L.; De Haas, G. & Heemskerk, C. T. (1963). Hydrolysis of synthetic mixed-acid phosphatides by phospholipase a from human pancreas. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Enzymological Subjects*, 67:295--304.

Van Noorden, R.; Maher, B.; Nuzzo, R. et al. (2014). The top 100 papers. *Nature*, 514(7524):550--553.

Vijayabaskar, M. & Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophysical journal*, 99(11):3704--3715.

Villar, H. O. & Kauvar, L. M. (1994). Amino acid preferences at protein binding sites. *FEBS Letters*, 349(1):125--130. ISSN 1873-3468.

von Heijne, G. (1991). Proline kinks in transmembrane α-helices. *Journal of molecular biology*, 218(3):499--503.

Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478):1803--1810.

Walter, K. U.; Vamvaca, K. & Hilvert, D. (2005). An active enzyme constructed from a 9-amino acid alphabet. *Journal of Biological Chemistry*, 280(45):37742--37746.

Wang, J. & Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nature Structural & Molecular Biology*, 6(11):1033--1038.

Williams, R. J. & Martinez, N. D. (2000). Simple rules yield complex food webs. *Nature*, 404(6774):180--183.

Zhao, H.; Tang, L.; Wang, X.; Zhou, Y. & Lin, Z. (1998). Structure of a snake venom phospholipase a 2 modified by p-bromo-phenacyl-bromide. *Toxicon*, 36(6):875--886.

Zhou, G. & Deng, M. (1984). An extension of chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochemical Journal*, 222(1):169--176.

Zhou, G.-P. (2011a). The disposition of the lzcc protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. *Journal of Theoretical Biology*, 284(1):142--148.

Zhou, G.-P. (2011b). The structural determinations of the leucine zipper coiled-coil domains of the cgmp-dependent protein kinase iα and its interaction with the myosin binding subunit of the myosin light chains phosphase. *Protein and peptide letters*, 18(10):966--978.

Zhou, G.-P. & Huang, R.-B. (2013). The ph-triggered conversion of the prpc to prpsc. *Current topics in medicinal chemistry*, 13(10):1152--1163.

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity.

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, 97:97--166.

Apêndice A

Telas do PFStats

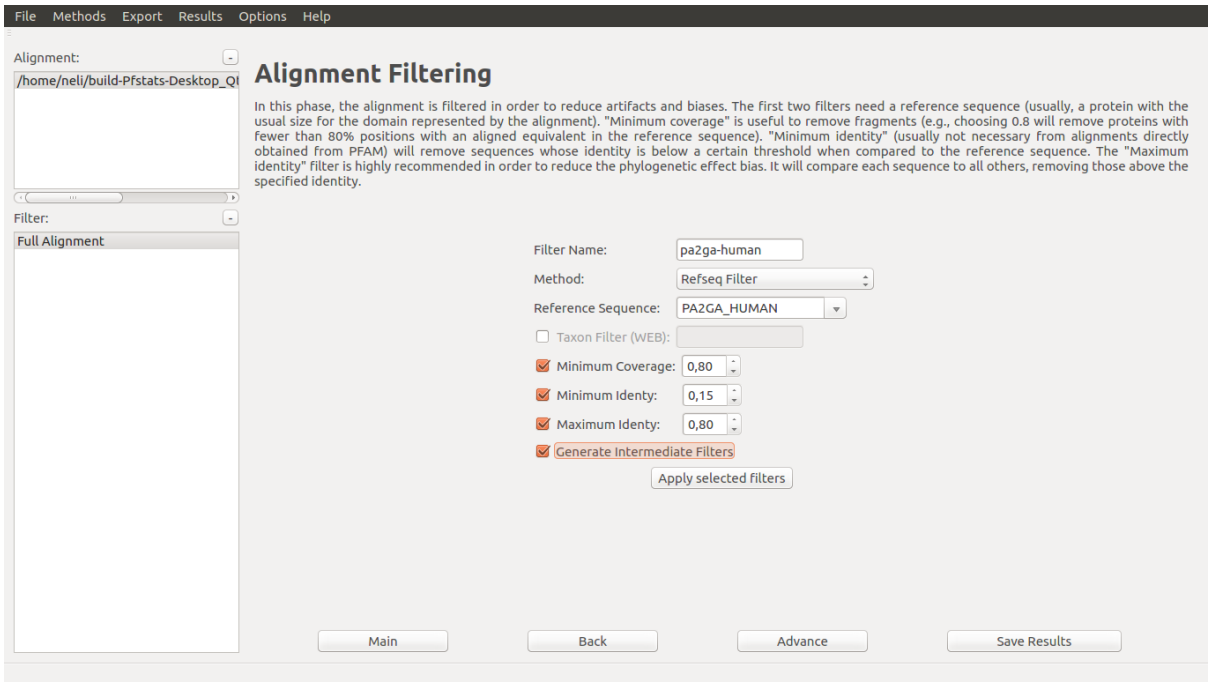


Figura A.1: Aplicação de filtros ao alinhamento.

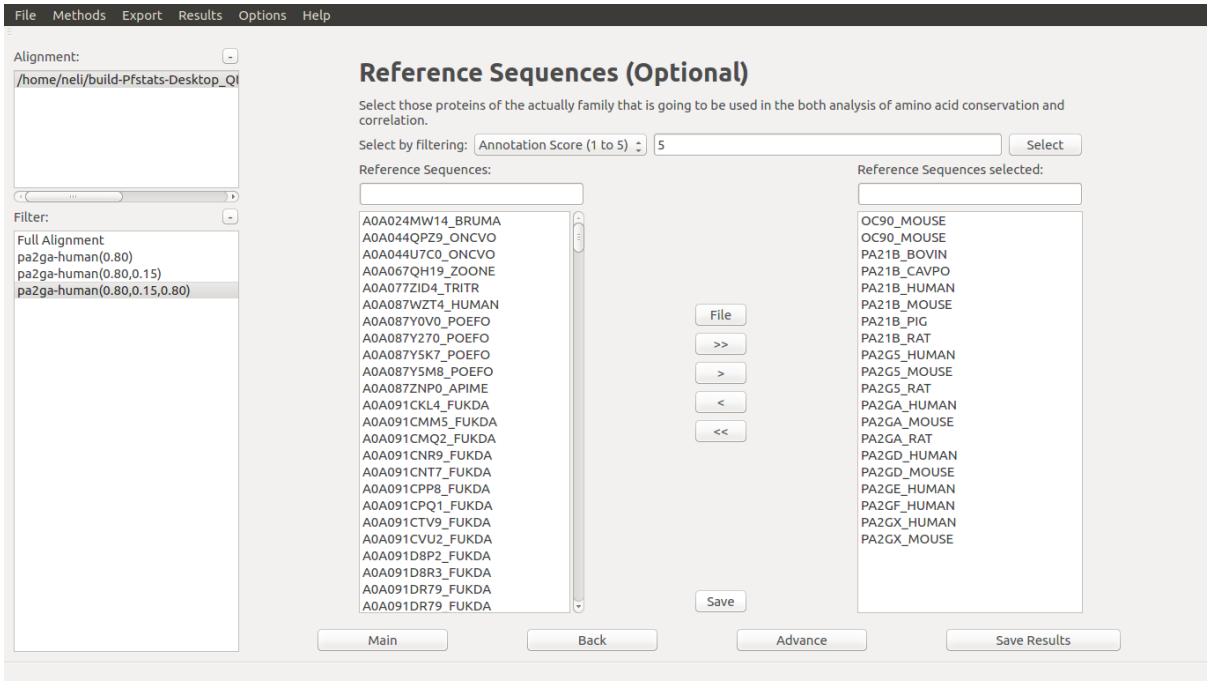


Figura A.2: Seleção de sequências de referência.

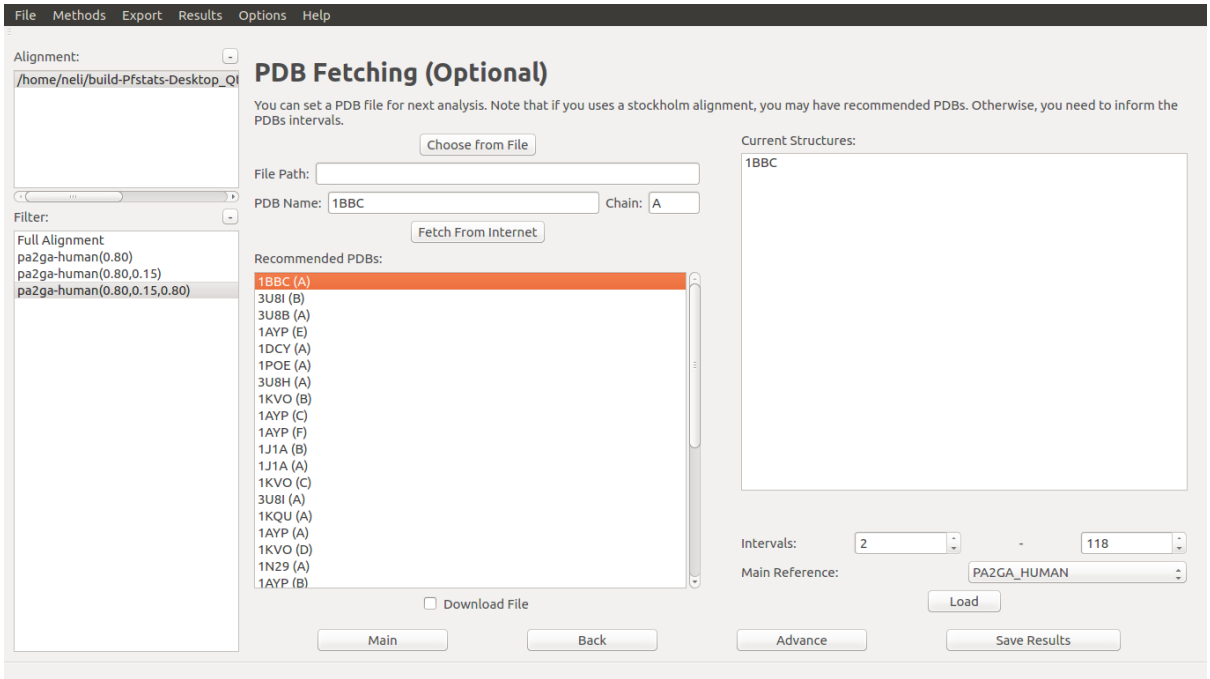


Figura A.3: Anexo de uma estrutura PDB a uma sequência para gerar visualizações.

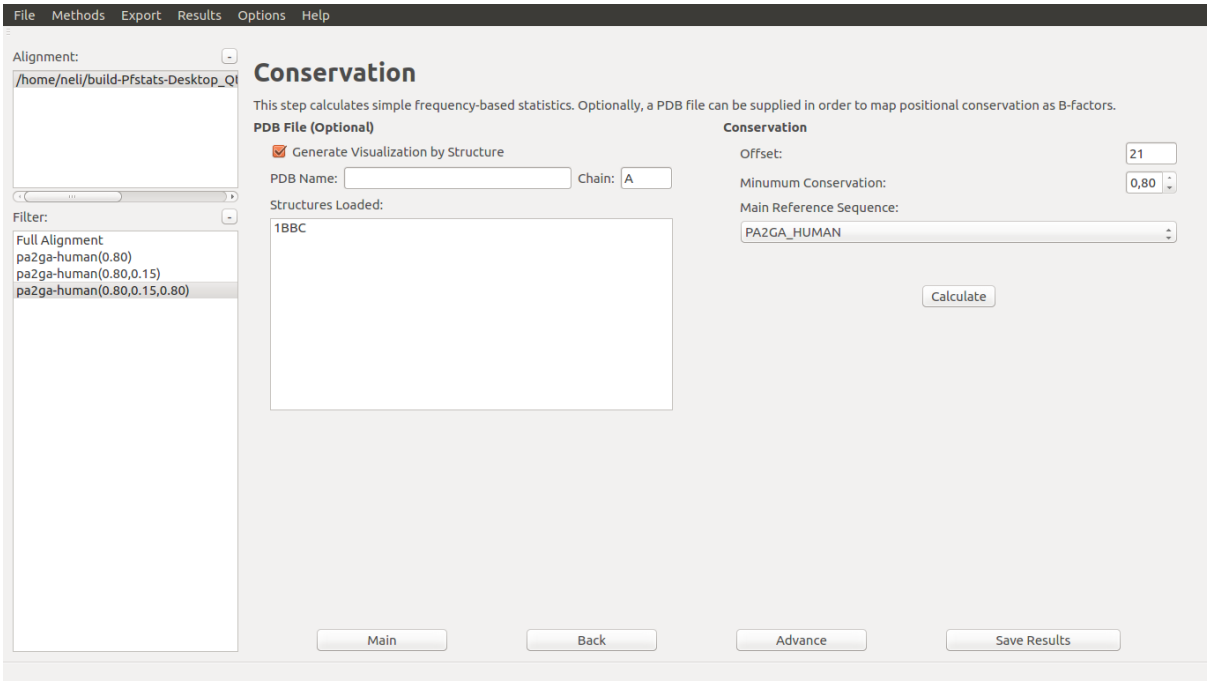


Figura A.4: Calcula de conservação de resíduos.

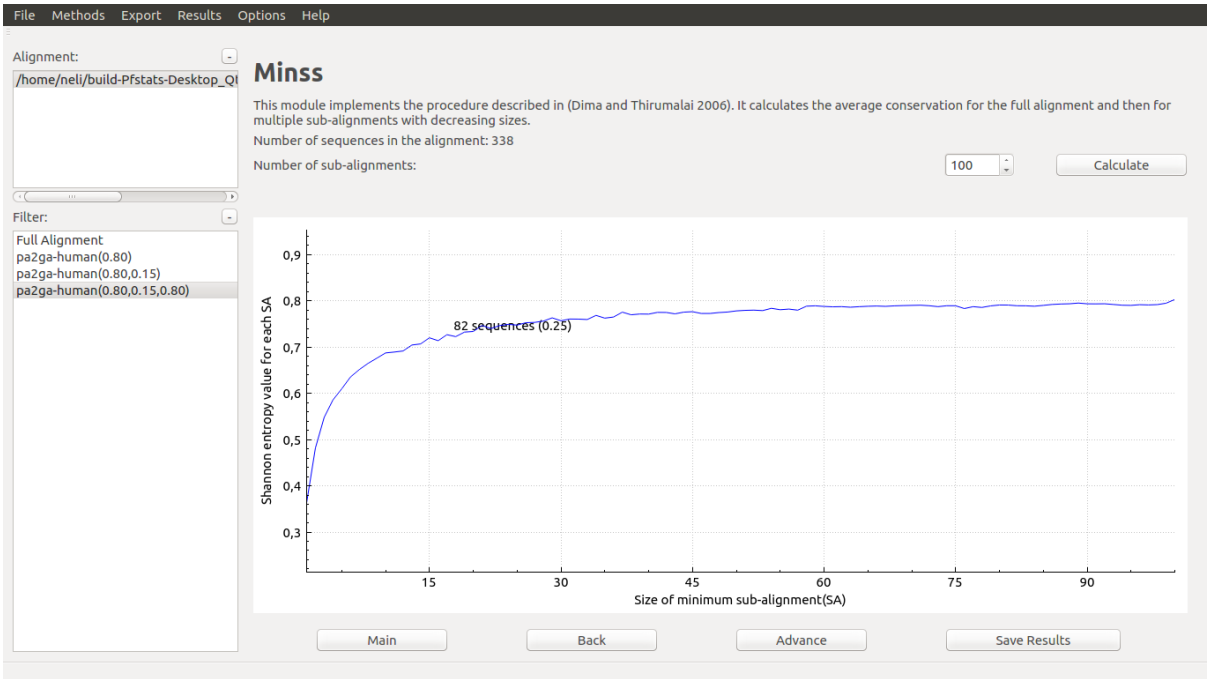


Figura A.5: Escolha do subalinhamento mínimo representativo

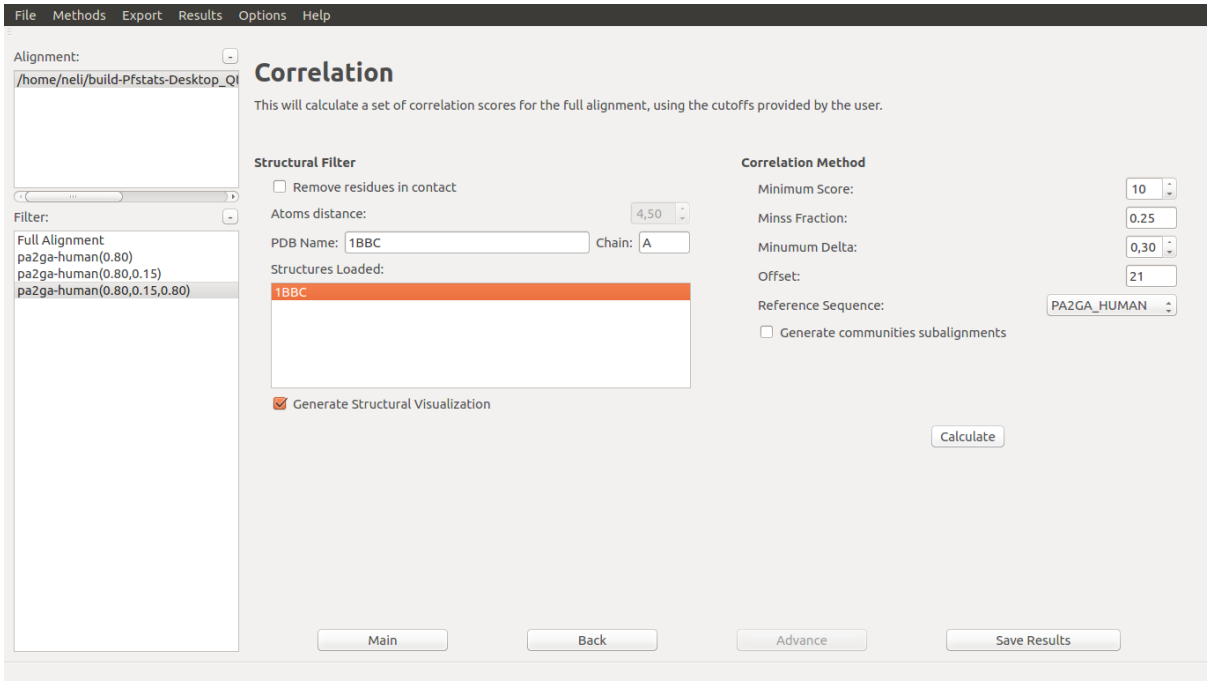


Figura A.6: Calculo das correlações e decomposição das comunidades.

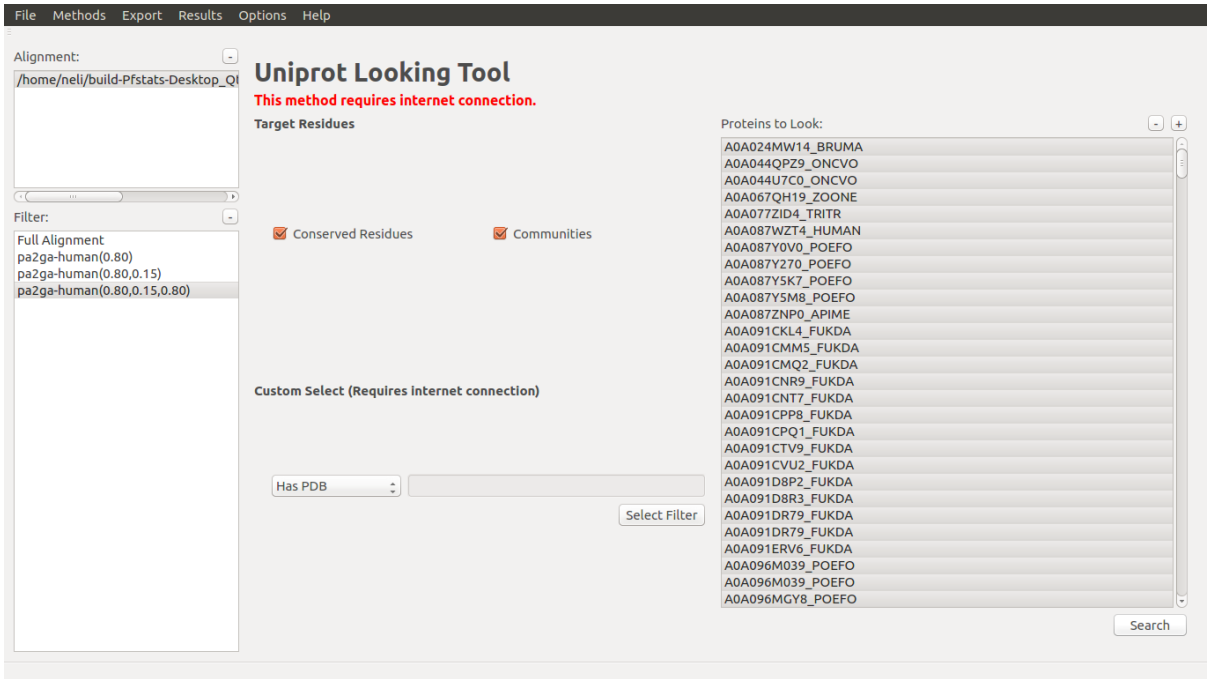
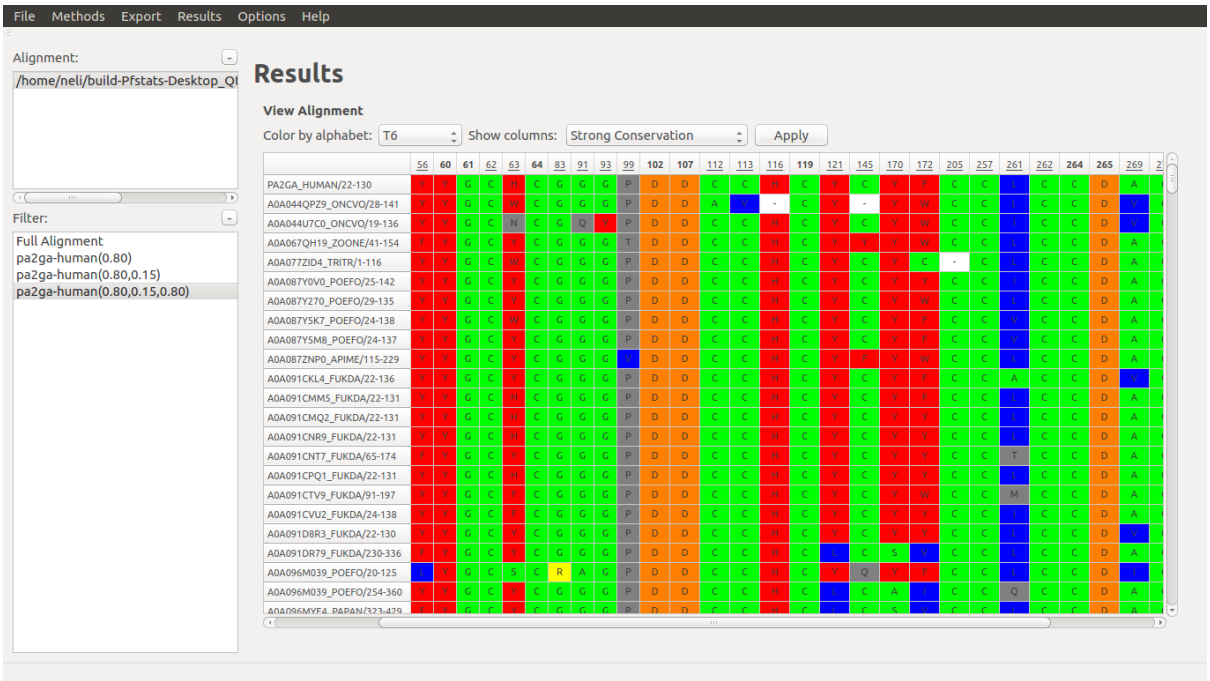


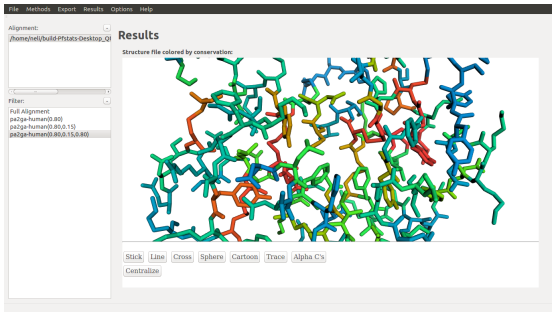
Figura A.7: Busca por anotações no Uniprot.

(a) Resíduos colorido por tipos de aminoácidos.

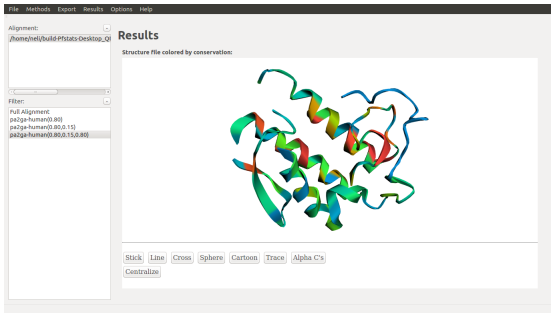


(b) Resíduos colorido por um alfabeto reduzido de 6 caracteres.

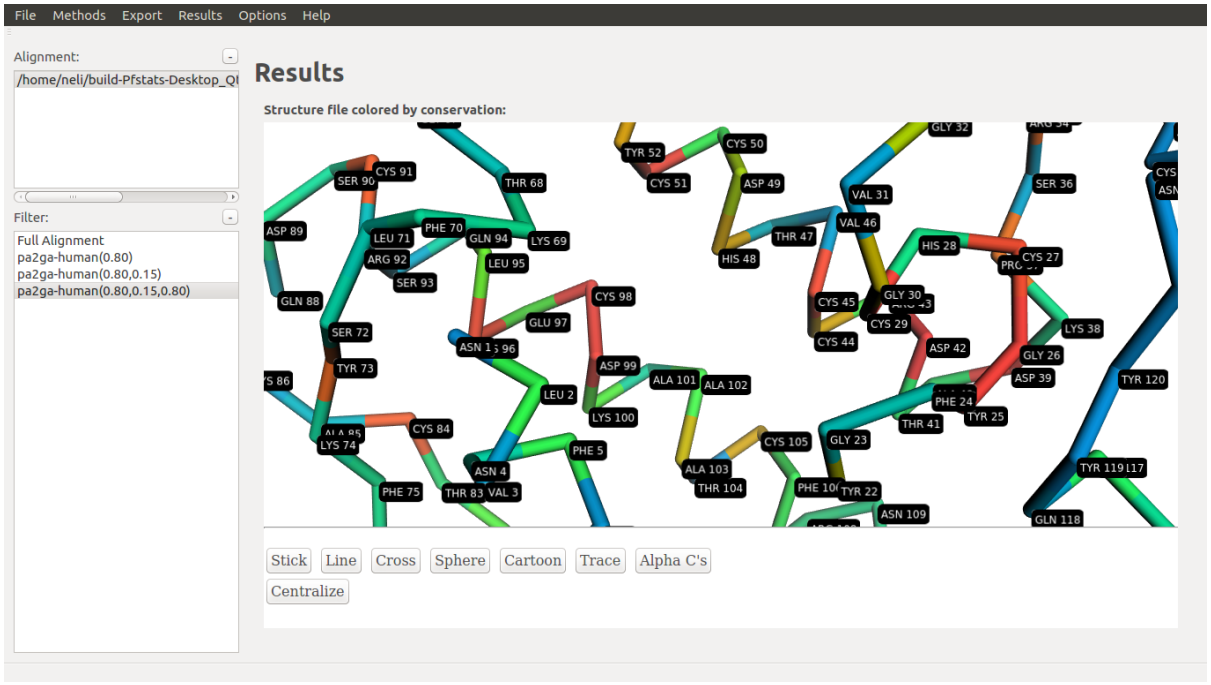
Figura A.8: Visualização de resíduos do alinhamento.



(a) Visualização por *sticks* com um *zoom* aproximado.

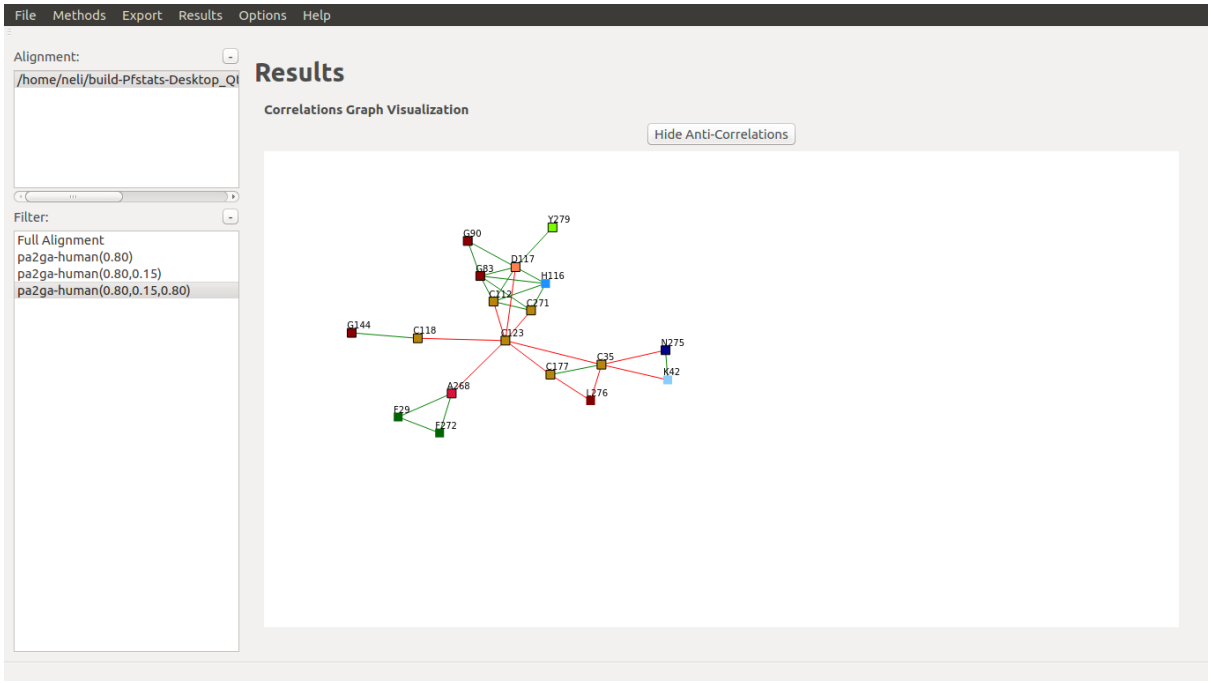


(b) Visualização no formato *cartoon*.

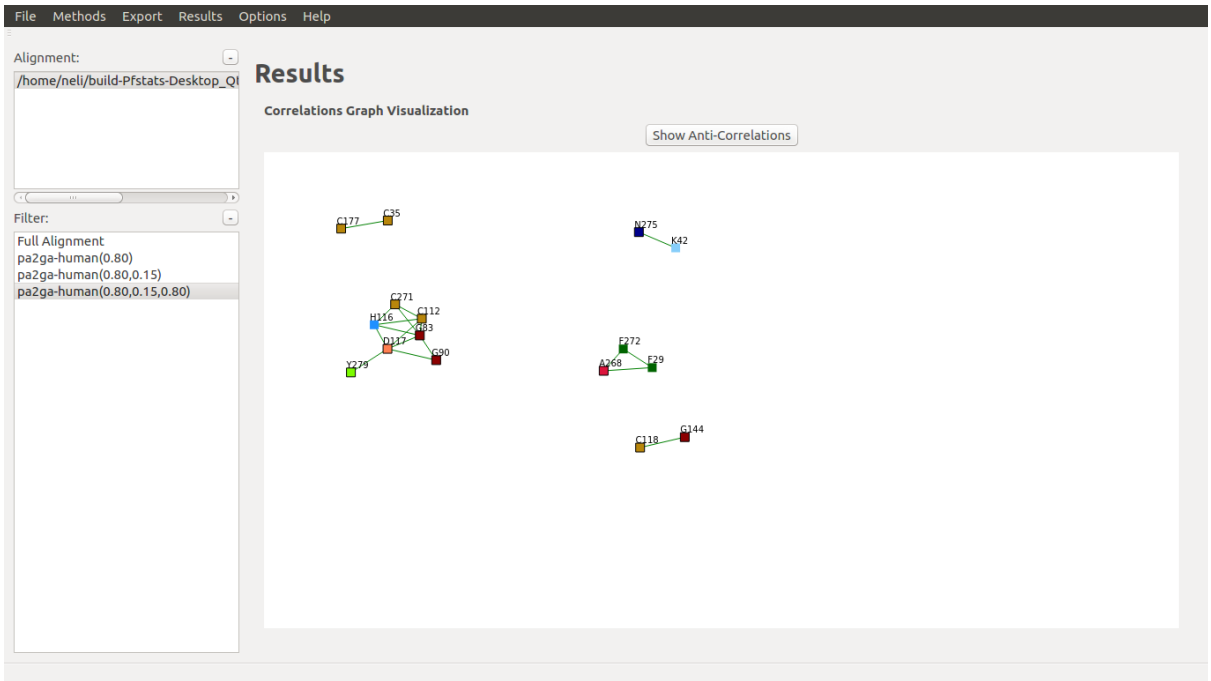


(c) Visualização da cadeia principal com seus resíduos rotulados.

Figura A.9: Visualização de resíduos conservados na estrutura. Cores mais próximas do vermelho indicam resíduos altamente conservados no AMS, cores mais próximas do azul indicam resíduos fracamente conservados.

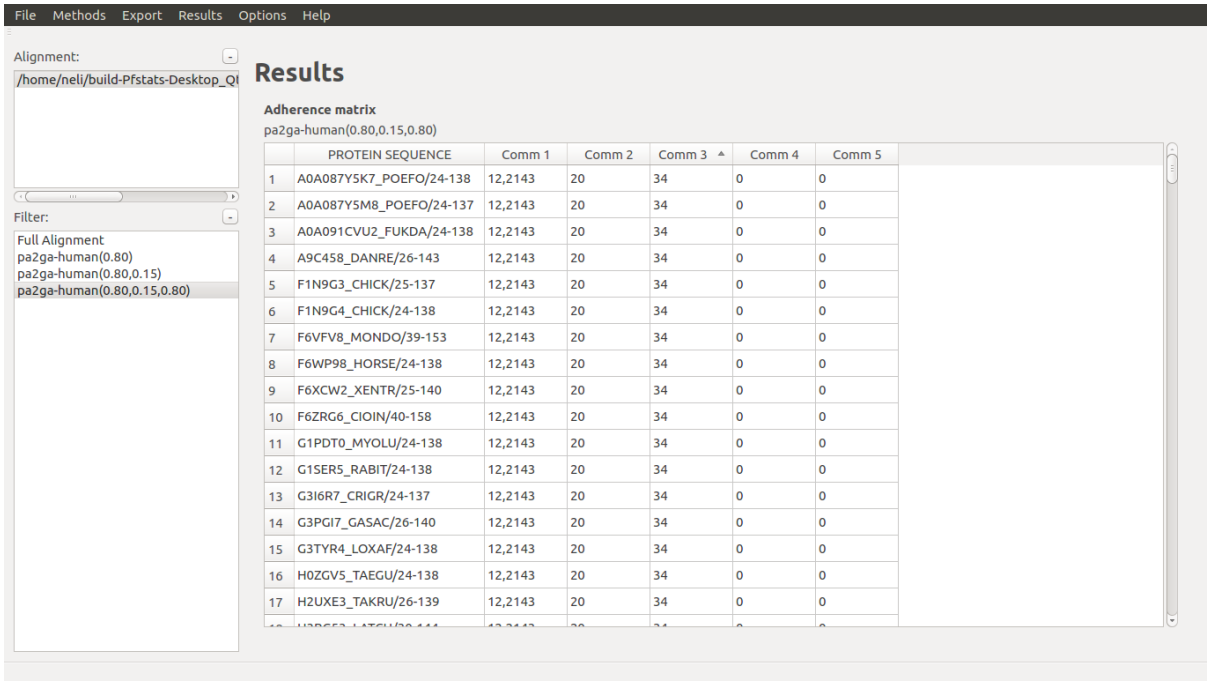


(a) Rede contendo correlações e anti-correlações.

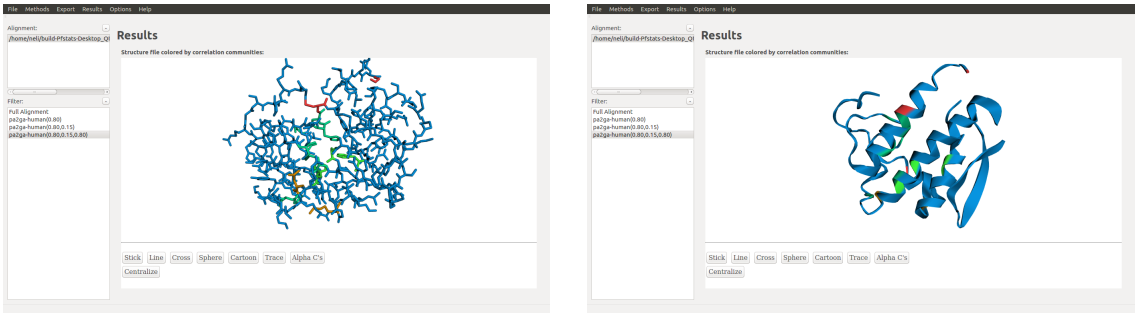


(b) Rede apenas com conexões positivas e componentes conexos separados.

Figura A.10: Visualização da rede de resíduos correlacionados. Os vértices estão coloridos por tipo de aminoácido enquanto as arestas verdes representam conexão positiva e as vermelhas, negativa.



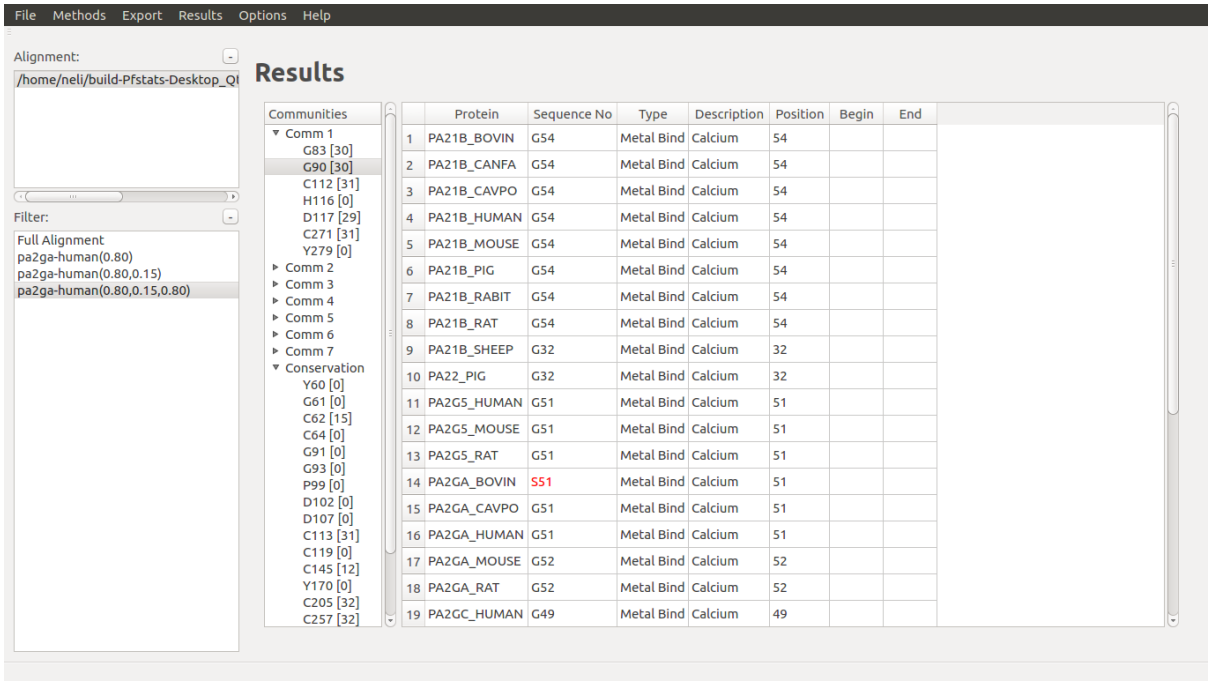
	PROTEIN SEQUENCE	Comm 1	Comm 2	Comm 3	Comm 4	Comm 5
1	A0A087Y5K7_POEFO/24-138	12,2143	20	34	0	0
2	A0A087Y5M8_POEFO/24-137	12,2143	20	34	0	0
3	A0A091CVU2_FUKDA/24-138	12,2143	20	34	0	0
4	A9C458_DANRE/26-143	12,2143	20	34	0	0
5	F1N9G3_CHICK/25-137	12,2143	20	34	0	0
6	F1N9G4_CHICK/24-138	12,2143	20	34	0	0
7	F6VFV8_MONDO/39-153	12,2143	20	34	0	0
8	F6WP98_HORSE/24-138	12,2143	20	34	0	0
9	F6XCW2_XENTR/25-140	12,2143	20	34	0	0
10	F6ZRG6_CIOIN/40-158	12,2143	20	34	0	0
11	G1PDT0_MYOLU/24-138	12,2143	20	34	0	0
12	G1SER5_RABIT/24-138	12,2143	20	34	0	0
13	G3I6R7_CRIGR/24-137	12,2143	20	34	0	0
14	G3PGI7_GASAC/26-140	12,2143	20	34	0	0
15	G3TYR4_LOXAF/24-138	12,2143	20	34	0	0
16	H0ZGV5_TAEGU/24-138	12,2143	20	34	0	0
17	H2UXE3_TAKRU/26-139	12,2143	20	34	0	0



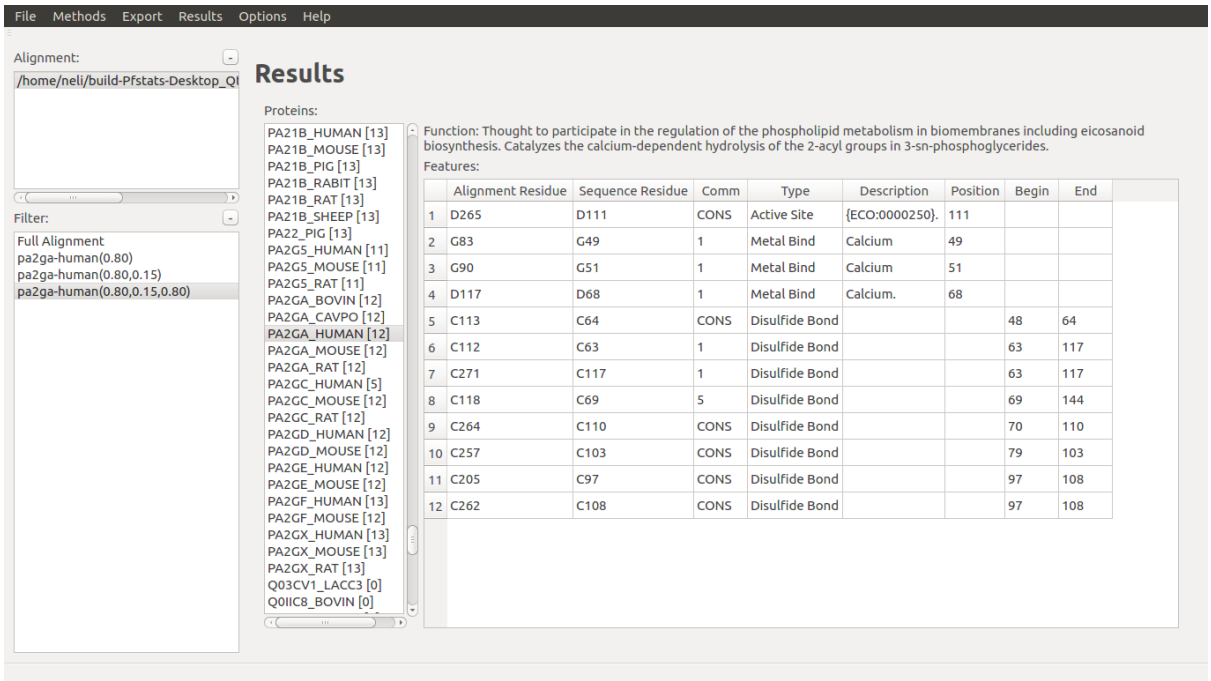
(a) Visualização por *sticks*.

(b) Visualização por *cartoon*.

Figura A.12: Visualização de comunidades de resíduos na estrutura. Resíduos pertencentes a mesma comunidade estão identificados pela mesma cor, enquanto resíduos de nenhuma comunidade estão representados na cor azul.



(a) Dados organizados por comunidades.



(b) Dados organizados por proteínas.

Figura A.13: Relação de resíduos agrupados com informações de anotação no uniprot.

Apêndice B

Lista de Alfabetos

Tabela B.1: Alfabeto T2 [Betts & Russell, 2007] (*O alfabeto Wang2 [Wang & Wang, 1999] adquiriu exatamente os dois mesmos grupos que T2.)

Rótulo	Aminoácidos
Hidrofílicos (P)	AGTSNQDEHRKP
Hidrofóbicos (H)	CMFILVWY

Tabela B.2: Alfabeto T5 [Betts & Russell, 2007]

Rótulo	Aminoácidos
Alifáticos (I)	IVL
Aromáticos (F)	FYWH
Carregados (K)	KRDE
Pequenos (G)	GACS
Outros (T)	TMQNP

Tabela B.3: Alfabeto T6 [Betts & Russell, 2007]

Rótulo	Aminoácidos
Alifáticos (I)	IVL
Aromáticos (F)	FYWH
Positivamente carregados (K)	KR
Negativamente carregados (D)	DE
Pequenos (G)	GACS
Outros (T)	TMQNP

Tabela B.4: Alfabeto 3IMGT [Pommié et al., 2004]

Rótulo	Aminoácidos
Hidrofílicos (I)	IVLFCMAW
Neutros (G)	GTSYPM
Hidrofóbicos (D)	DNEQKR

Tabela B.5: Alfabeto 5IMGT [Pommié et al., 2004]

Rótulo	Aminoácidos
60-90Å^3 (G)	GAS
108-117Å^3 (C)	CDPNT
138-154Å^3 (E)	EVQH
162-174Å^3 (M)	MILKR
189-228Å^3 (F)	FYW

Tabela B.6: Alfabeto 11IMGT [Pommié et al., 2004]

Rótulo	Aminoácidos
Alifáticos (A)	AVIL
Fenilalanina (F)	F
Enxofres (C)	CM
Glicina (G)	G
Hidroxílas (S)	ST
Triptófano (W)	W
Tirosina (Y)	Y
Prolina (P)	P
Ácidos (N)	NQ
Básicos (H)	HKR

Tabela B.7: Alfabeto Murphy15 [Murphy et al., 2000]

Rótulo	Aminoácidos
Hidrofóbicos granes (L)	LVIM
Cisteína (C)	C
Alanina (A)	A
Glicina (G)	G
Serina (S)	S
Treonina (T)	T
Triptófano (W)	W
Hidrofóbicos aromáticos (F)	FY
Prolina (P)	P
Glutamato (E)	E
Aspartato (D)	D
Asparagina (N)	N
Glutamina (Q)	Q
Cadeias longas positivamente carregadas (K)	KR
Histidina (H)	H

Tabela B.8: Alfabeto Murphy10 [Murphy et al., 2000]

Rótulo	Aminoácidos
Hidrofóbicos granes (L)	LVIM
Cisteína (C)	C
Alanina (A)	A
Glicina (G)	G
Polares (S)	ST
Hidrofóbicos aromáticos (F)	FYW
Prolina (P)	P
Polares/Carregados (E)	EDNQ
Cadeias longas positivamente carregadas (K)	KR
Histidina (H)	H

Tabela B.9: Alfabeto Murphy8 [Murphy et al., 2000]

Rótulo	Aminoácidos
Hidrofóbicos (L)	LVIMC
Pequenos (A)	AG
Polares (S)	ST
Hidrofóbicos aromáticos (F)	FYW
Prolina (P)	P
Polares/Carregados (E)	EDNQ
Cadeias longas positivamente carregadas (K)	KR
Histidina (H)	H

Tabela B.10: Alfabeto Murphy4 [Murphy et al., 2000]

Rótulo	Aminoácidos
Hidrofóbicos (L)	LVIMC
Pequenos (A)	AGSTP
Hidrofóbicos aromáticos (F)	FYW
Hidrofílicos (E)	EDNQKRH

Tabela B.11: Alfabeto Murphy2 [Murphy et al., 2000]

Rótulo	Aminoácidos
Hidrofóbicos (L)	LVIMCAGSTPFYW
Hidrofílicoss (E)	EDNQKRH

Tabela B.12: Alfabeto Wang5 [Wang & Wang, 1999]

Rótulo	Aminoácidos
I	CMFILVWY
A	ATH
G	GP
E	DE
K	SNQRK

Tabela B.13: Alfabeto Wang5v [Wang & Wang, 1999]

Rótulo	Aminoácidos
I	CMFI
L	LVWY
A	ATGS
E	NQDE
K	HPRK

Tabela B.14: Alfabeto Wang3 [Wang & Wang, 1999]

Rótulo	Aminoácidos
I	CMFILVWY
A	ATHGPR
E	DESNQK

Tabela B.15: Alfabeto Li10 [Li et al., 2003]

Rótulo	Aminoácidos
C	C
Y	FYW
L	ML
V	IV
G	G
P	P
S	ATS
N	NH
E	QED
RK	RK

Tabela B.16: Alfabeto Li5 [Li et al., 2003]

Rótulo	Aminoácidos
Y	CFYW
I	MLIV
G	G
S	PATS
N	NHQEDRK

Tabela B.17: Alfabeto Li4 [Li et al., 2003]

Rótulo	Aminoácidos
Y	CFYW
I	MLIV
S	GPATS
E	NHQEDRK

Tabela B.18: Alfabeto Li3 [Li et al., 2003]

Rótulo	Aminoácidos
I	CFYWMLIV
S	GPATS
E	NHQEDRK

Anexo A

Exemplos de redes complexas

Figura A.1: Rede de conexões sociais no Facebook em 2010 (http://hipertextual.com/2010/12/gran-mapa-mundial-de-la-amistad-facebook).

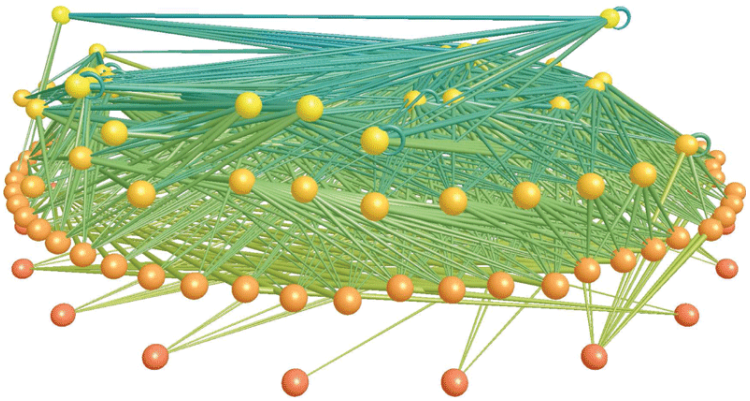


Figura A.2: Cadeia alimentar do lago Little Rock em Wisconsin [Williams & Martinez, 2000].

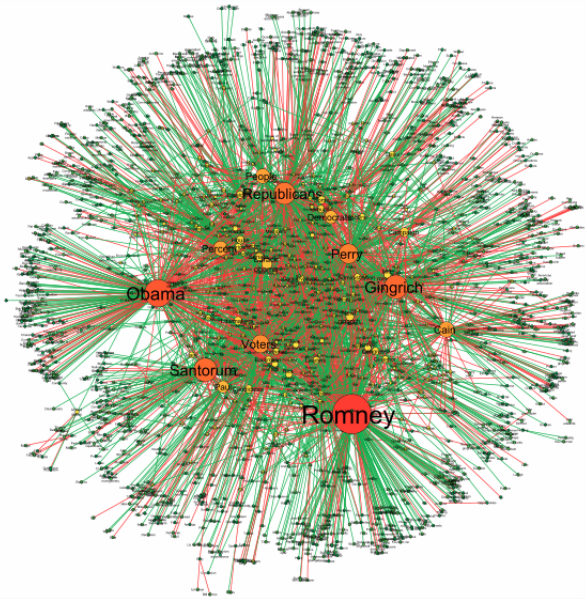


Figura A.3: Rede de palavras chaves ligadas a eleição presidencial dos EUA em 2012 [Sudhahar et al., 2015].

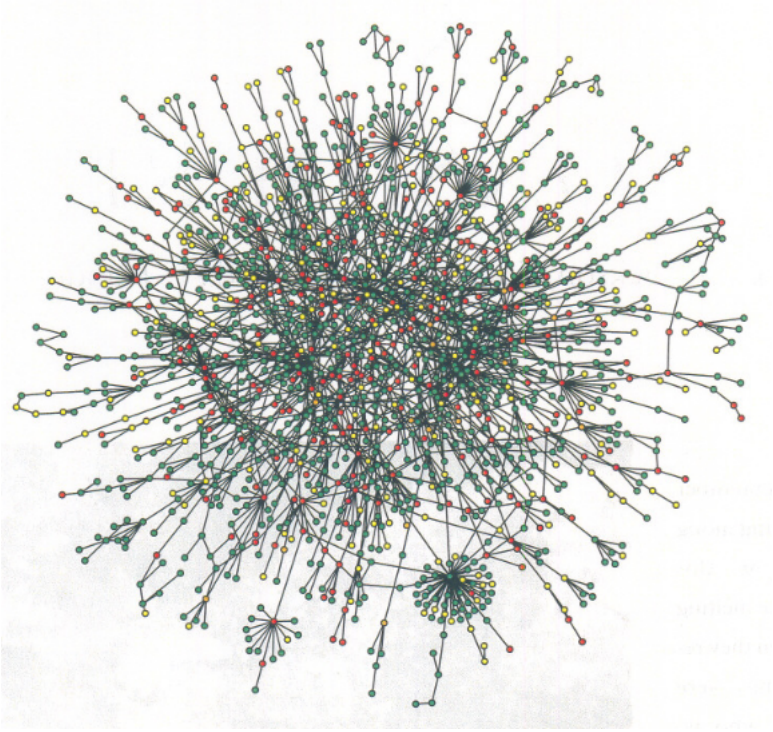


Figura A.4: Rede de interações proteína-proteína em levedura [Barabási & Bonabeau, 2003].