MODELAGEM E DECOMPOSIÇÃO DE REDES DE COEVOLUÇÃO DE AMINOÁCIDOS: APLICAÇÕES EM DETERMINAÇÃO DE ESPECIFICIDADE E ANOTAÇÃO DE PROTEÍNAS

NÉLI JOSÉ DA FONSECA JÚNIOR

MODELAGEM E DECOMPOSIÇÃO DE REDES DE COEVOLUÇÃO DE AMINOÁCIDOS: APLICAÇÕES EM DETERMINAÇÃO DE ESPECIFICIDADE E ANOTAÇÃO DE PROTEÍNAS

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

Orientador: Lucas Bleicher

Belo Horizonte

Agosto de 2020

Fonseca Júnior, Néli José da

043 Modelagem e Decomposição de Redes de Coevolução de Aminoácidos: Aplicações em Determinação de Especificidade e Anotação de Proteínas / Néli José da Fonseca Júnior. — Belo Horizonte, 2020
xxvii, 112 f. : il. ; 29,5cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Lucas Bleicher

1. Biologia computacional. 2. Coevolução. 3. Redes Reguladoras de Genes. 4 Aprendizado de Máquina. 5. Software. I. Lucas Bleicher. II. Universidade Federal de Minas Gerais. III. Título

CDU 573:004

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

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE

NELI JOSE DA FONSECA JUNIOR

Às quatorze horas do dia **30 de setembro de 2020**, reuniu-se, de forma remota, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Modelagem e decomposição de redes de coevolução de aminoácidos: aplicações"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Lucas Bleicher**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**Indicação**
Dr. Lucas Bleicher	UFMG	Aprovado
Dra. Laila Alves Nahum	Fiocruz Minas	Aprovado
Dr. Richard Charles Garratt	IFSC-USP	Aprovado
Dr. José Miguel Ortega	UFMG	Aprovado
Dr. José Ribamar dos Santos Ferreira Jr.	EACH-USP	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

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

Belo Horizonte, 30 de setembro de 2020.

Documento assinado eletronicamente por **José Ribamar dos Santos Ferreira Júnior**, **Usuário Externo**, em 30/09/2020, às 18:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Lucas Bleicher**, **Professor do Magistério Superior**, em 30/09/2020, às 18:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 30/09/2020, às 18:51, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **RICHARD CHARLES GARRATT**, **Usuário Externo**, em 30/09/2020, às 18:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Laila Alves Nahum**, **Usuário Externo**, em 02/10/2020, às 08:16, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0276940** e o código CRC **ECFBDA57**.

Referência: Processo nº 23072.230242/2020-30 SEI nº 0276940

Dedicuum cest laborae a quelquis personatum que ajudorat a facirelo.

Agradecimentos

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

Um agradecimento especial ao meu orientador e amigo, Prof. Lucas Bleicher, sem o qual este trabalho não seria possível. Aos colegas e colaboradores Lucas Carrijo, Marcelo Querino, Dhiego Souto e Natan Pedersolli. E a todos os membros do nosso grupo de pesquisa.

Aos meus companheiros de república: Keth, Sergio, Robério, Ana, Syd, Júlio, Lorena, Pedro^2, Siriema, Davi, Hugo, Leo^2, Marcelo e Nádia, que acompanharam todo o percurso, desde o começo, ajudando nos estudos, nas escolhas e nos momentos de diversão.

Aos professores da Universidade Federal de Ouro Preto: Tiago Garcia, Joubert Lima, Fernando Sica e Eduardo Luz, principais responsáveis pela minha formação acadêmica inicial.

Aos membros do colegiado do Programa de pós-graduação em Bioinformática da UFMG, especialmente Gloria, Raquel, Francisco Lobo e Miguel, responsáveis pelo enorme conhecimento adquiridos durante meu período de atuação como representante discente.

Aos colegas da organização do III Curso de Verão em Bioinformática da UFMG: Heron Hilário, Alessandra Lima, Stellamaris Soares, Ana Paula Abreu, Nayara Toledo, Raphael Tavares, Rodrigo Kato e Gabriel.

Aos colegas do ISCB Regional Student Group Brazil, por todo conhecimento adquirido durante anos de organização de iniciativas que contribuem para a difusão da Bioinformática no Brasil, incluindo a Primeira liga nacional de bioinformática. Um agradecimento especial para Nilson Coimbra, Raquel Riyuzo, Elvira Horácio, Juliana Assis, Liliane Conteville, Lucas Carvalho, Maira Neves, Sheila Nagamatsu, Antony Brayan, Deyvid Amgarten, Fernanda Caroline, Fernando Rossi, Marcus braga e Paulo Burke.

Agradeço também aos parceiros e colaboradores de São Paulo: Prof. José Riba-

mar, Janaina Paulela, Rafaela Maria, Vittoria Camandona e Lodair Junior.

Aos colegas de bebedeiras no Cabral, X-Meetings, SBBf, SBBq e outros congressos, fundamental para passar por todo esse sofrimento: Lucas, Marcelo, Ovelha, Jesus, Rayson, Felipe, Nilson e Carlos.

Aos colegas de trabalho no Instituto Europeu de Bioinformatica: Gerard, Ardan, Andrei, Sanja, Osman, Sriram, Cesare, Preeti, Paul, Zhe, Andrii e Amudha.

E finalmente um indispensável agradecimento a minha Marmotinha, Thaina Miranda, peça fundamental para que eu tenha chegado até o final sem nenhum surto psicótico, me agraciando com muito amor, carinho, conselhos e broncas.

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

Resumo

Estudos de evolução molecular computacional são geralmente conduzidos a partir de alinhamentos múltiplo de sequências homólogas, no qual sequências possivelmente originadas por um ancestral comum são alinhadas de forma que aminoácidos equivalentes ocupem a mesma posição. Padrões de conservação de resíduos em um alinhamento, ou em um subconjunto de suas sequências, podem ser informativos por sugerirem posições sob seleção e restrição evolutiva. A maioria dos métodos propostos para identificação de determinantes de especificidade são focados em posições, logo, acabam ignorando os padrões de determinante para uma subfamília, porém variável no alinhamento como um todo. Além disto, boa parte deles também requerem algum tipo de conhecimento a priori das famílias analisadas, como lista de subfamílias ou árvores filogenéticas. Neste trabalho foi desenvolvido uma metodologia baseado em ciências das redes, com objetivo de identificar grupos de resíduos funcionalmente relacionados. A metodologia foi inicialmente validada a partir de conjunto de dados artificiais e posteriormente aplicada a quatro famílias de proteínas reais. Em todos os casos foram obtidos grupos de resíduos determinantes de especificidade para diversas subclasses funcionais. Estes dados foram posteriormente utilizados como estimadores para uma máquina de suporte de vetores (SVM) que foi capaz de classificar corretamente até mesmo subclasses, a quais nenhum resíduo específico foi identificado. A classificação foi também aplicada às GPCRs órfãs gerando novas hipóteses a respeito das classes funcionais destas sequências. Um sistema web foi desenvolvido com o objetivo de promover e facilitar as analises utilizando as metodologias propostas neste projeto. Além disto, foi desenvolvido um banco de dados de sítios determinantes de especificidades contendo analises previamente calculadas com conjunto de dados obtidos pelo Pfam. Este banco, além de também produzir uma serie de relatórios dinâmicos e intuitivos, possui também uma REST API que permite que estes dados sejam acessados programaticamente.

Palavras-chave: Analises de coevolução, Bioinformática Funcional, Ciências das Redes, Aprendizagem de Máquina, Engenharia de Software.

Abstract

Computational molecular evolution analyses are usually performed by using multiple sequences alignments of homologous sequences, in which sequences likely originated from a common ancestors are aligned in a such way that equivalent amino acids are set in the same column. Conserved residues in a multiple sequence alignment can be extremely enlightening by suggesting positions under evolutionary selection and constraint. Most of the methods proposed to coevolution and specificity determinant sites are focused in finding positions, therefore they may ignore sites that are specific for a subfamily but variable in the whole alignment; or requires prior knowledge about the subject families, such as list of subfamilies or phylogenetic trees. This project presents a network-based methodology, commonly apllied to social and ecological systems, with the goal to identify clusters of functionally related residues. The method was first validated using artificial datasets and then applied to four real protein families: C-type Lysozyme/Alpha-lactoalbumin, HIUase/Transthyretin, Amidases and the class A G protein-coupled receptors. Patterns of specificity determinant sets for many functional subclasses were successfully extracted from all these families. These networks were then used as features for a support vector machine (SVM) that was able to correctly classify even subfamilies without detected specificty determinant residues. This machine was also applied to the orphan GPCRs generating novel hypothesis about these proteins. We developed a web application with the aim of promote and facilitate the studies performed by the methodology proposed in the project, this system is able to generate a series of data visualization and cross-references with external archives. Finally, we created a database for specificity determinant sites including precalculated analysis with datasets extracted from Pfam. This database, despite generating many intuitional and dynamic reports, it also has a REST API allowing programmatically access to its data.

Keywords: Coevolution analysis, Functional bioinformatics, Network science, Machine learning, software engineering.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADRP	Retinite pigmentosa autossômica dominante
AMS	Alinhamento múltiplo de sequências
API	Interface de programação de aplicações
BCM	Monocromacia de cone azul
BE	Equação de Bonacich
BHN	Normalizacao de Borgatti e Halgin
BLAST	Ferramenta básica de busca por alinhamento local
CEvADA	Arquivo de dados de análises de coevolução
CONAN	Analisador de redes de covariação
DF	Filtro de disparidade
ET	Traço evolutivo
FCDV	Fração dos vértices corretamente detectados
GATA	Glu-tRNA amidotransferases
GPCR	Receptores acoplados a proteína G
HIUase	5-hidroxi-isourato hidrolase
HMLF	Filtro de probabilidade marginal Hairball
HMM	Modelos ocultos de Markov
JC	Coeficiente de Jaccard
LALBA	Alfa-lactoalbumina
LYSC	Lisozimas de tipo C
PC	Coeficiente de correlação de Pearson
PDB	Banco de dados de proteínas
Pfam	Banco de dados de famílias de proteínas
REST	Transferência representacional de estado
SACA3	Proteínas associadas a membrana do acrossomo do espermatozoide
SCA	Analise por acoplamento estatistico
SCOP	Classificação estrutural de proteínas
SDS	Sítios determinante de especificidade
SVM	Máquina de suporte de vetores

Sumário

Capítulo 1

Introdução

A aplicação de técnicas computacionais para elucidação de problemas biológicos, área conhecida como bioinformática, obteve um crescimento em escala geométrica nas últimas décadas. Tal crescimento se deve a recentes avanços tecnológicos e metodológicos que permitiram uma evolução na capacidade de armazenamento e processamento de dados [Cook et al., 2019]. Porém, as origens deste campo datam da metade do século passado. Um dos primeiros registros conhecidos do termo bioinformática consta no trabalho de Hogeweg & Hesper [1978], ao definirem o termo como *"o estudo de processos informáticos aplicados a sistemas bióticos"*. O próprio autor porém, cita como sua primeira utilização do termo, um artigo ainda anterior publicado em uma pequena revista holandesa em 1970 [Hesper & Hogeweg, 1970; Hogeweg, 2011]. Mas talvez, o mais antigo trabalho a se encaixar na atual definição de bioinformática date de 1952, quando Bennett & Kendrew publicaram o programa de computador que foi utilizado na determinação da primeira estrutura em alta resolução de uma proteína, a mioglobina, resolvida em 1958 [Bennett & Kendrew, 1952; Kendrew et al., 1958].

Na década de 60, Linus Pauling e Emile Zuckerland, através de estudos com um pequeno conjunto de sequências de hemoglobina de diferentes espécies, perceberam que as sequências biológicas evoluem a uma taxa mensurável e relativamente constantes, e que portanto padrões evolutivos poderiam ser extraídos a partir de um conjunto de sequências homólogas [Zuckerkandl & Pauling, 1962, 1965]. Em 1968, Kimura et al. propôs a teoria neutra da evolução molecular, segundo a qual, em nível molecular, a maior parte da variabilidade genética dos organismos não é fruto de seleção natural, mas sim por derivas genéticas aleatórias, sendo portanto seletivamente neutra. Ou seja, a maioria dos aminoácidos de uma proteína poderiam passar por mutações aleatórias sem nenhuma alteração em sua função, estando apenas alguns poucos sítios sob uma restrição evolucionária mais rigorosa [Kimura et al., 1968]. Esta teoria foi

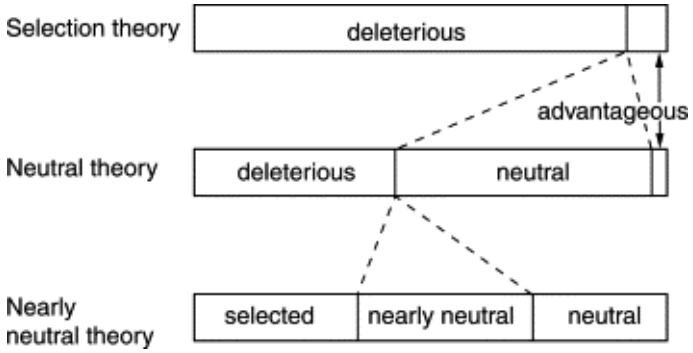


Figura 1.1: Comparação entre as taxas de mutações de acordo com as teorias da seleção natural, neutra e quase neutra. [Ohta, 2001].

estendida no começo da década de 70, apos Ohta observar que a taxa de mutações não poderia ser divida apenas em neutras ou deletérias, mas sim influenciadas tanto pela seleção, quanto por deriva genética [Ohta, 1973] (Figura 1.1). Estes trabalhos foram fundamentais para o surgimento e avanço da biologia molecular evolutiva, uma vez que possibilitou a utilização de múltiplas sequências, seja de proteínas ou DNA, para extração de informações referentes a história evolutiva das mesmas.

Margaret Dayhoff, uma das pioneiras na bioinformática, responsável pela criação da codificação de aminoácidos por únicos caracteres [Eck & Dayhoff, 1966]; primeiro banco de dados de sequências de proteínas, "*Atlas of Protein Sequence and Structure*" [Dayhoff, 1965], e da matriz de substituição de aminoácidos PAM [Dayhoff et al., 1978]; foi também responsável pelo primeiro algorítmo para extrair informação a partir de pareamento de sequências, o MAXLAP [Dayhoff & Ledley, 1962] (Figura 1.2). Porem, o uso de sequências pareadas nos estudos de evolução molecular obteve um grande salto principalmente após os trabalhos de Needleman & Wunsch [1970] e Smith & Waterman [1981], que desenvolveram algoritmos baseado em programação dinâmica para obter respectivamente alinhamentos ótimo local e global entre pares de sequências. Estes algoritmos permitiram o surgimento de métodos como o BLAST, aumentando a velocidade de buscas por sequências de acordo com identidade em ordens de magnitude [Altschul et al., 1990].

Até meados da década de 70, análises evolutivas de sequências eram realizadas por pareamento, ou seja, sequências alinhadas par a par. As primeiras tentativas de alinhar múltiplas sequências requeriam como parâmetro o fornecimento de uma árvore filogenética [Sankoff et al., 1976; Waterman & Perlwitz, 1984; Hogeweg & Hesper, 1984]. A utilização de múltiplas sequências alinhadas obteve um grande salto a partir de 1987, quando foi publicado o primeiro método utilizando a heurística progressiva [Feng & Doolittle, 1987] para alinhar mais de duas sequências. A partir daí, o uso de sequências

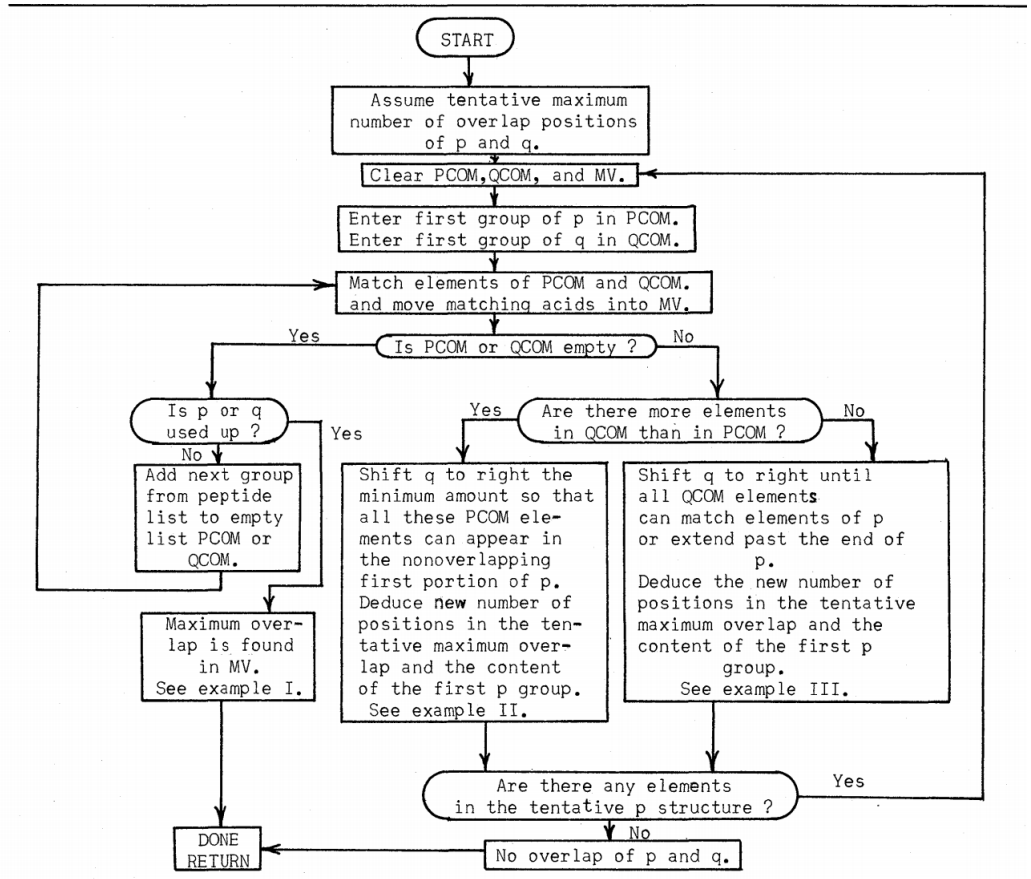


Figura 1.2: Fluxograma do algorítmo MAXLAP, um dos pioneiros a utilizar o conceito se sequências alinhadas. O programa buscava encontrar a sobreposição possível entre sequências de dois peptídeos. [Dayhoff & Ledley, 1962].

alinhadas se tornou tão comum, que o alinhamento múltiplo de sequências (AMS), hoje, é considerado um modelo indispensável na bioinformática. Sua importância é tanta que um estudo publicado na Nature em 2014 o classificou como um dos modelos mais utilizados hoje na Biologia [Van Noorden et al., 2014]. Este mesmo estudo ainda incluiu o ClustalW [Thompson et al., 1994] na décima colocação entre os artigos científicos mais citados de todos os tempos (atualmente com 62.840 citações segundo o Google Scholar).

Ao analisar alinhamentos de famílias de proteínas, como esperado pelas teorias de neutralidade molecular de Kimura e Ohto, é comum observar que a maioria das colunas apresentam uma alta variabilidade de aminoácidos, provavelmente relacionada a substituições neutras ou quase neutras. Em contraste, algumas poucas colunas apresentam aminoácidos que foram estritamente conservados durante a evolução, possuindo uma variabilidade extremamente baixa. A hipótese de que posições estritamente conservadas em sequências homologas pudessem ser utilizadas como estimadores de importância

funcional, antecede o uso de alinhamentos, sendo proposta ainda na década de 60 por Zuckerkandl & Pauling [1965]. Porém este assunto ganhou realmente atenção apos o surgimento dos primeiros algorítmos para geração de alinhamentos múltiplo de sequência, no final dos anos 80. Zvelebil et al. [1987] observaram que resíduos conservados em um alinhamento múltiplo de sequências ortólogas poderiam ser tanto utilizado como estimadores de contatos e predição de estrutura secundarias, quanto para detectar sítios ativos. Krah et al. [1998] conseguiram determinar com exatidão os resíduos envolvidos no sitio ativo das laminarinases analisando apenas a conservação de aminoácidos a partir de um pequeno conjunto de sequências ortólogas alinhadas e validando posteriormente utilizando técnicas de mutagênese sítio-dirigida. Atualmente, a conservação de aminoácidos em AMS já e considerada um dos principais sinais de importância funcional ou estrutural [Choi et al., 2012; Pazos & Bang, 2006]. E importante ressaltar que sinais de conservação em AMS ocorrem não somente em relação à especificidade de aminoácidos, mas também em relação a propriedades físico-químicas e estruturais que necessitam serem mantidas para que a proteína conserve sua atividade e estabilidade. Este tipo de padrão é comumente denominado de posições marginalmente conservadas [Chakrabarti et al., 2007]. Uma demonstração de sinais de conservação observados em um alinhamento múltiplo de sequências pode ser visto na figura 2.8.

O processo de duplicação gênica seguido de divergência permite o surgimento de proteínas com atividade diferente de seus ancestrais. Isso ocorre pelo fato de que após a duplicação, uma das cópias pode perder suas restrições evolutivas, uma vez que a produção de uma proteína com aquela atividade será compensada pela outra cópia do gene. Sendo assim, mutações anteriormente proibitivas passam a ocorrer sem que haja prejuízo para o organismo, podendo levar a um processo de neofuncionalização. Portanto, famílias de proteínas podem conter múltiplas subclasses funcionais, e a identificação de posições localmente conservadas, podem trazer tanto informação a respeito de resíduos que possuam importância funcional ou estrutural a uma determinada subclasse. Na Figura 2.8 é possível observar um exemplo de alinhamento múltiplo de sequências contendo todos os tipos de padrões citados anteriormente.

Com o avanço dos algorítmos voltados pra análise de conservação de aminoácidos, geração de alinhamentos múltiplos de sequências e do crescimento exponencial do número de sequências disponíveis a partir dos anos 90, criou-se a necessidade de analisar conservação local em subfamílias de proteínas. Uma das primeiras abordagens implementadas com esse intuito foi proposta por Livingstone & Barton [1993]. O algorítmo consistia em calcular a conservação de uma forma hierárquica, comparando as sequências de cada subgrupo predefinidos pelo usuário. Este método já utilizava um alfabeto expandido de aminoácidos para levar em consideração propriedades físico-

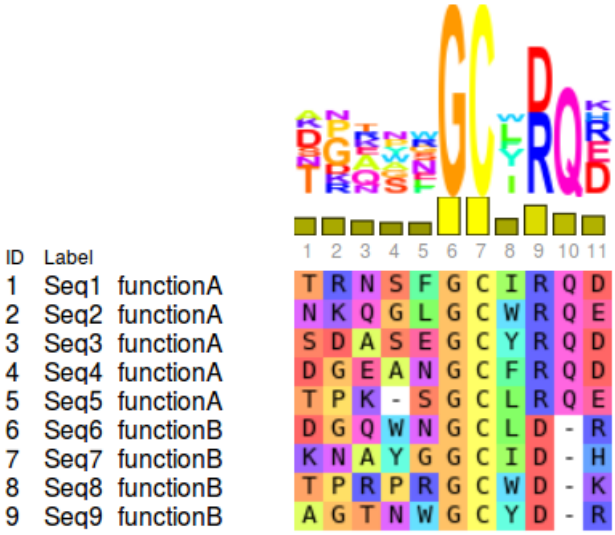


Figura 1.3: Alinhamento Múltiplo de Sequências contendo duas subclasses funcionais definidas pelas funções A e B. As colunas 1 a 5 representam posições de alta variabilidade. Colunas 6 e 7 indicam posições extremamente conservadas. A coluna 8 representa uma posição marginalmente conservadas por aminoácidos hidrofóbicos. As colunas 9 e 10 indicam posições localmente conservadas para as subclasses de função A e B. Finalmente a coluna 11 se trata de uma posição localmente marginalmente conservada, sendo composta por aminoácidos negativos na subclasse de função A, e positivos na subclasse de função B. Esta imagem foi gerada com a ferramenta MSAViewer [Yachdav et al., 2016].

químicas e estruturais. Outro algorítmo que fez bastante sucesso nos primórdios foi o traço evolutivo (ET), que utilizava uma arvore filogenética guia para quantificar a variação de conservação entre os ramos [Lichtarge et al., 1996]. Estes grupos de resíduos que definiam subfamílias de proteínas em alinhamentos, foram denominados sítios determinantes de especificidade (SDS).

Paralelamente, surgia ainda no começo dos anos 90 uma outra classe de algorítmos baseados alinhamentos múltiplos de sequências, as análises de correlação de resíduos (também chamada de coevolução ou covariação). Estes métodos consistem em observar a coocorrência de resíduos no AMS e foram inicialmente propostos para predição de contatos estruturais, uma vez que duas posições extremamente correlacionadas poderiam indicar um sinal de proximidade estrutural [Göbel et al., 1994; Atchley et al., 1999].

Em 1999, Lockless & Ranganathan publicaram o método *Statistical coupling analysis* (SCA), que consistia em calcular correlação entre posições de um AMS a partir de uma distribuição energética dos resíduos. Os autores também observaram que resíduos correlacionados poderiam também realçar grupos funcionais, principalmente

motivos relacionados aos sítios de ligação de uma proteína [Lockless & Ranganathan, 1999]. Bachega et al. utilizaram uma adaptação do SCA calculando as correlações em nível de resíduos (aminoácido-posição) foramam capazes de predizer sítios determinantes de especificidade sem a necessidade de predefinição das subclasses, arvores filogenéticas guias ou outros dados adicionais. Os autores utilizaram o método para encontrar grupos de resíduos que definiam a especificidade das superóxido dismutases de ferro e de manganês [Bachega et al., 2009]. Paralelamente, Halabi et al. também utilizaram o algorítimo SCA para encontrar sítios determinantes de especificidade na família das serino proteases, no que eles denominaram *protein sectors* [Halabi et al., 2009].

Métodos baseado em coocorrência de resíduos voltados para análise de especificidade em famílias de proteínas permitem que o usuário realize estudos exploratórios, sendo possível a detecção de padrões relativos a subfamílias ainda não caracterizadas. Alguns exemplos de aplicações deste tipo de abordagem incluem: descoberta de uma nova classe de receptores nucleares de nematódeos caracterizada por um motivo especifico do P-Box [Afonso et al., 2013]; caracterização de uma subfamília de proteína cuja função ainda não era conhecida Coitinho et al. [2019]; caracterização dos resíduos envolvidos no processo de ativação do zimogênio nas serina proteases da família das tripsinas [Querino Lima Afonso et al., 2020]; descrição dos resíduos envolvidos em mudanças conformacionais e capacidades proteolíticas no domínio metil transferase das proteínas NS2B, NS3 e NS5 de flavivírus [da Fonseca Jr et al., 2017]; anotação de sequências [Pedruzzi et al., 2014; da Fonseca Jr et al., 2019]; e aplicações no design de fármacos e planejamento de estudos de mutagêneses [Rios-Anjos et al., 2017; Suhadolnik et al., 2017; Coitinho et al., 2019; Barwinska-Sendra et al., 2020].

Existe uma serie de dificuldades encontradas na detecção de SDS, talvez uma das principais se trate da escalaridade. Chakraborty & Chakrabarti [2014] realizaram um estudo comparativo dos principais métodos para detecção de SDS disponíveis na literatura e apenas 4 dos 12 métodos avaliados foram plausíveis de serem utilizados em larga escala. Além disto, os testes foram realizados em alinhamentos com número extremamente reduzido de sequências. O maior AMS utilizado possuía 180 sequências, número muito distante dos alinhamentos disponíveis no Pfam [El-Gebali et al., 2019], que podem passar de 1 milhão de sequências. Além disto, métodos que necessitam de informações adicionais, como arquivos de estrutura, árvores filogenéticas ou predefinição das subfamílias, apesar de gerarem resultados acurados, limitam o escopo das análises a famílias de proteínas que possuam estes dados disponíveis e em amostragem suficiente. Algorítmos que predizem posições determinantes de especificidade, e não resíduos, tendem a falhar na detecção de SDS de tipo 1, quando o aminoácido é con-

servado dentro de uma única subclasse, mas é variável nas outras. Em contrapartida, algorítmos baseados em taxa de conservação tendem a falhar na detecção de SDS de tipo 2, afinal estas posições são relativamente conservadas, variando apenas a letra respectiva a cada conjunto. E finalmente, algorítmos que não levam em consideração as propriedades físico-químicas e estruturais dos aminoácidos, falham na detecção de SDS de tipo 3 [Chakraborty & Chakrabarti, 2014].

Como observado em Schnoes et al. [2009], a taxa de erro em bancos de dados de anotação automática de proteínas pode passar de 80%. Estes erros geralmente ocorrem dentro de superfamílias, ou seja, o classificador acerta a família da proteína, mas erra na escolha da subfamília. Uma provável hipótese para isto é o fato das anotações serem geralmente realizadas com base na similaridade global entre as sequências. Logo, tendo em mãos um detector de SDSs confiável e escalável pode permitir que anotações mais precisas sejam realizadas.

1.1 Objetivos Gerais

Este trabalho tem como objetivo o desenvolvimento de um novo algoritmo escalar para detecção de resíduos determinantes de especificidade, utilizando técnicas de teoria dos grafos, similaridade de sequências e análises de conservação e coevolução. O algorítmo proposto tem como objetivo ser uma continuação do método de Bleicher et al. [2011] com o intuito de corrigir algumas das falhas comentadas em Chakraborty & Chakrabarti [2014], como a detecção de SDS tipo 3 e uma melhor normalização dos escores para evitar que padrões relativos a subfamílias de baixa frequência no alinhamento sejam dissolvidos através dos filtros. Além disto, também foram construídos aplicações *web* e *desktop* com o objetivo de facilitar o uso da metodologia proposta bem como a interpretações dos resultados obtidos, uma vez que ferramentas gráficas para análises de SDS e de coevolução são bastante escassas. Finalmente, foi desenvolvido um banco de dados de sítios determinantes de especificidade a partir do Pfam. Este banco permitirá o acesso aos resultados em tempo de execução, além de acesso programático a partir de um identificador de família ou de sequência.

1.2 Objetivos Específicos

- Desenvolver um algoritmo baseado em teoria dos grafos, para detectar SDSs a partir de alinhamentos múltiplos de sequências.
- Validar a metodologia utilizando dados reais e artificiais.

- Desenvolver um *script* de fácil manuseio para o uso da metodologia proposta.
- Desenvolver uma aplicação *web* que permita o usuário utilizar a metodologia proposta e forneça um relatório rico em visualizações e cruzamento de dados. de cada uma dos algoritmos propostos neste trabalho.
- Desenvolver um banco de sítios determinantes de especificidades pré calculados que cubra a maior parte das entradas do Pfam.

Capítulo 2

Fundamentação Teórica

2.1 Proteínas

As proteínas são as macromoléculas biológicas mais encontradas na natureza e são responsáveis pela grande maioria dos processos em que ocorrem uma célula como defesa imunológica contra organismos invasores (anticorpos), catálises de reações químicas (enzimas), transmissão de sinais para o controle de processos biológicos (hormônios), transporte de átomos e moléculas pequenas (proteínas transportadoras), provimento de estrutura e suporte para as células (proteínas estruturais), dentre outros. Estas macromoléculas são compostas por combinações de dezenas a milhares de aminoácidos ligados de modo covalente em uma sequência linear característica [Nelson & Cox, 2018].

Existem 20 tipos de aminoácidos principais, denominados aminoácidos primários (este número pode variar um pouco, caso incluso aminoácidos raros na natureza, como a selenocisteína e a pirrolisina). Portanto proteínas com funções completamente diferentes, como um anticorpo, uma enzima ou a protéina estrutural do chifre de um rinoceronte são simplesmente resultado de diferentes combinações e repetições deste mesmo conjunto de 20 aminoácidos. Cada um destes aminoácidos são compostos por um grupo amina ligado a um grupo carboxila através de um átomo de carbono denominado carbono α, e se diferem um dos outros através de suas cadeias laterais. Diferentes aminoácidos possuem diferentes características fisioquímicas e estruturais, como carga elétrica, afinidade por água, presença de anel aromático, tamanho, entre outros (Figura 2.1). O aminoácido presente em uma cadeia polipeptídica é denominado resíduo de aminoácido, refletindo a perda de uma molécula de água quando ligado covalentemente a outro aminoácido.

A função de uma proteína esta diretamente relacionada a sua conformação espacial, ou seja a forma como a combinação de uma sequência de aminoácidos com

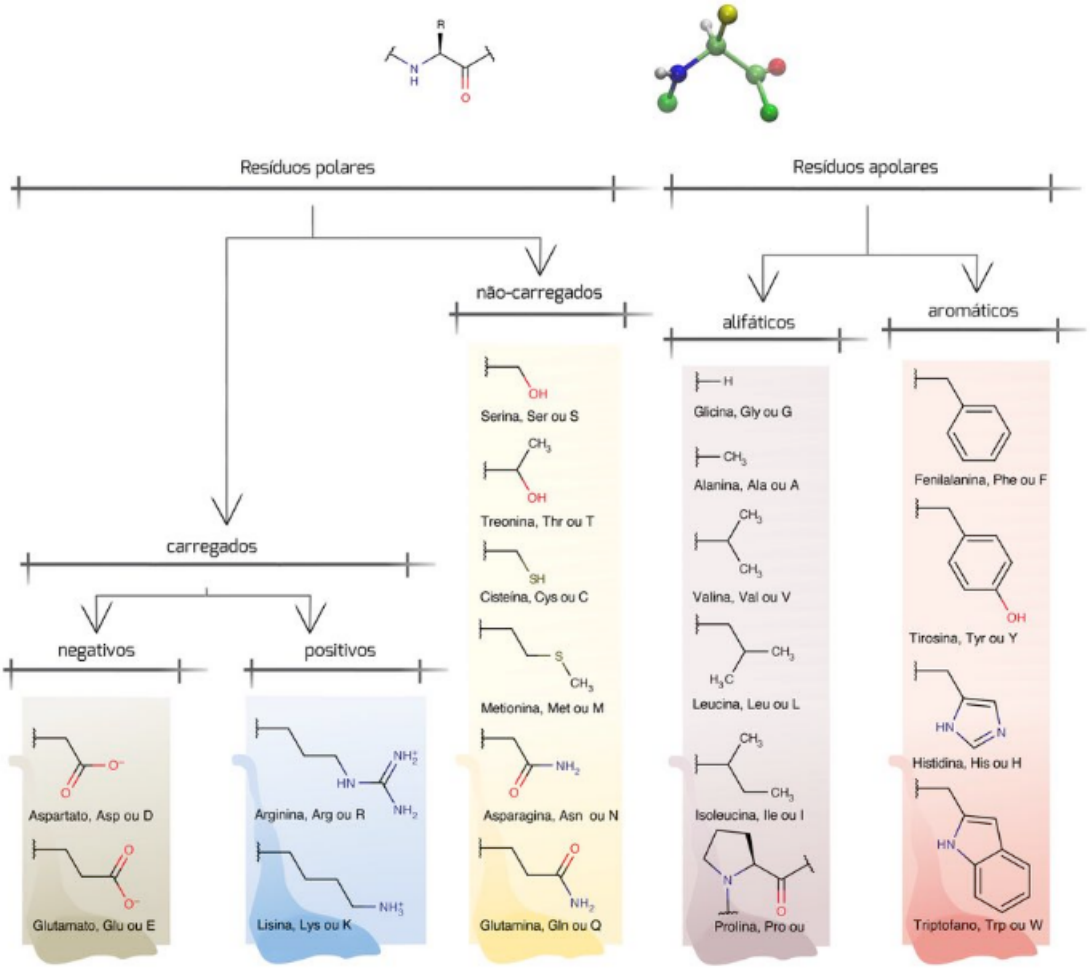


Figura 2.1: Estrutura da cadeia lateral de cada um dos 20 aminoácidos primários agrupados de acordo com suas características fisicoquímicas e estruturais [Ferreira, 2005].

diversas propriedades interagem entre si formando um enovelamento tridimensional. Essas estruturas podem possuir diferentes padrões de flexibilidade, unidades mais rígidas podem possuir um papel estrutural no esqueleto celular ou em tecidos conectivos; já as regiões mais flexíveis tendem a ter um papel importante na maquinaria celular, atuando como dobradiças, molas e alavancas [Berg et al., 2002]. A figura 2.2 mostra um exemplo de alteração conformacional na estrutura da proteína após a interação com um átomo de ferro.

A estrutura de uma proteína pode ser organizada em quatro níveis hierárquicos: estrutura primária, secundária, terciária e quaternária. A informação contida em um nível inferior é importante ou necessária para as representações em níveis superiores, apesar de não ser o único fator. Por exemplo, normalmente é considerado que a informação contida na sequência de aminoácidos (estrutura primária) é determinante para

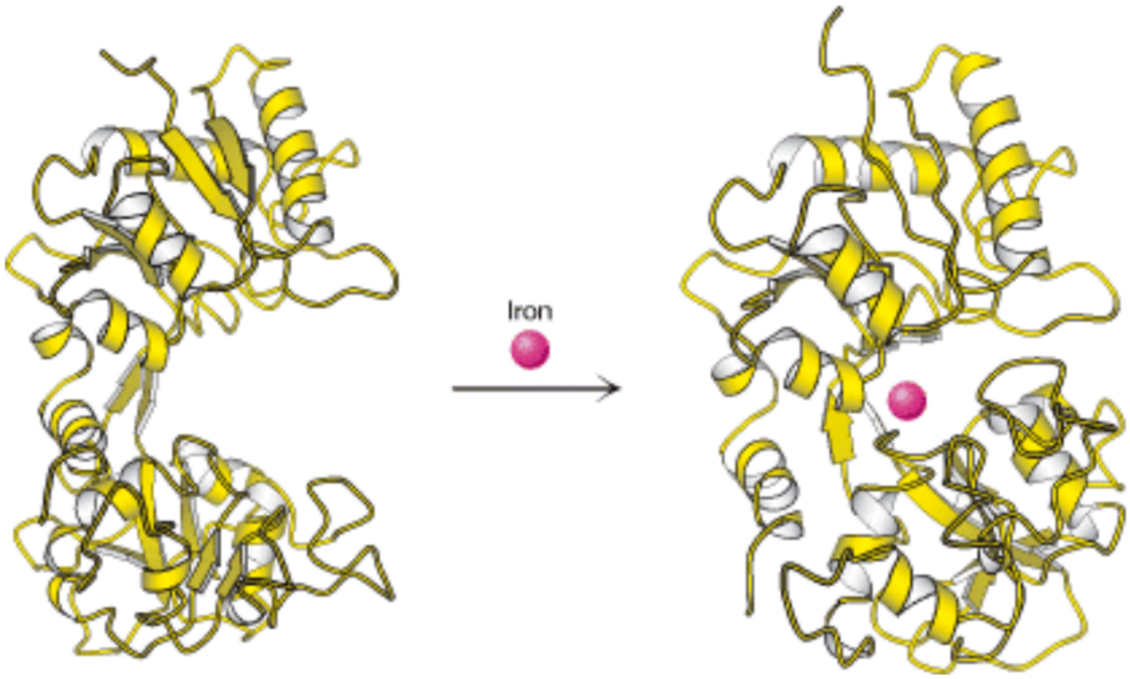


Figura 2.2: Ao interagir com uma molécula de ferro, a proteína lactoferrina sofre mudanças conformacionais que permitem que outras moléculas sejam capazes de diferenciar sua forma livre e ligada ao ferro [Berg et al., 2002].

sua estrutura secundária, porém não a única determinante [Ferreira, 2005].

A estrutura primária de uma proteína refere-se a sua sequência básica de aminoácidos conectados através de ligações peptídicas e iniciando por um grupo amino (N-terminal) e terminando por um grupo carboxila (C-terminal). A sequência de uma proteína é comumente representada de forma unidimensional, onde cada aminoácido é identificado por um código de uma ou três letras, como pode ser observado na figura 2.1. Apesar da aparente simplicidade, uma vez que a única dimensão de informação consiste da ordem de aparecimento dos resíduos, estes dados permitem uma série de análises, principalmente quando usada em conjunto com outras sequências relacionadas. Além disto, a estrutura primária de uma proteína é experimentalmente mais fácil de ser obtida do que sua estrutura tridimensional. A nível de comparação, no dia 18 de Junho de 2020 haviam aproximadamente 185 milhões de sequências de proteínas depositadas no UniprotKb [Boutet et al., 2016], das quais 562 mil manualmente curadas e revisadas; em contrapartida apenas 165 mil estruturas depositadas no PDB [Armstrong et al., 2020].

As estruturas secundárias consistem de padrões conformacionais, originados por interações entre aminoácidos vizinhos e moléculas do solvente, que tendem a se repetir em cadeias polipeptídicas. Estes padrões foram propostos inicialmente por Linus

Pauling e Robert Corey em 1951 ao definirem duas estruturas periódicas que foram chamadas de alfa hélice e folha beta [Pauling et al., 1951]. Posteriormente outras estruturas secundárias foram identificadas como as voltas, alças, beta barril, entre outras [Berg et al., 2002]. Diferentes combinações de sequências de aminoácidos podem originar uma mesma estrutura secundária. A figura 2.3 ilustra alguns exemplos de estruturas secundárias.

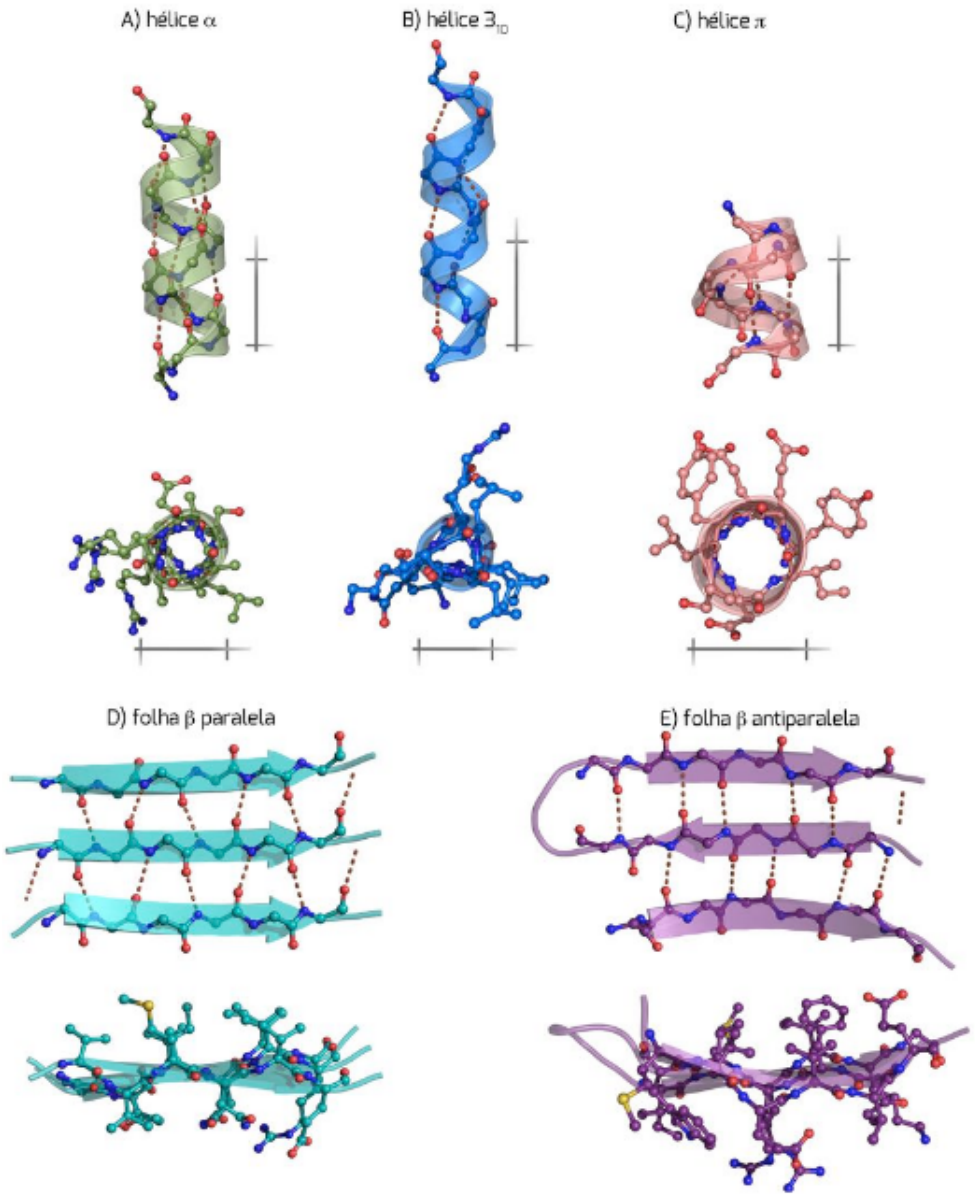
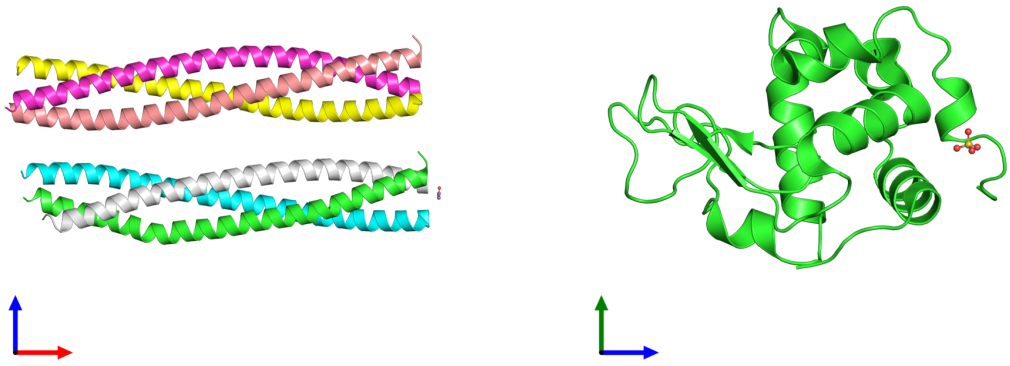


Figura 2.3: Exemplos das estruturas de algumas alfa hélices e folhas betas Ferreira [2005].

A estrutura terciária representa a conformação tridimensional de todos os átomos de uma cadeia polipeptídica, ou seja, ela descreve como os elementos das estruturas

secundárias se organizam no espaço. A organização destes átomos se estabelece através de uma serie de interações entre partes da própria cadeia e entre outras moléculas do solvente, formando o enovelamento da proteína. As proteínas podem ser classificadas, de acordo com sua estrutura terciária, entre fibrosa ou globular (Figura 2.4). As proteínas fibrosas se organizam em longos filamentos, geralmente hidrofóbicos e formados por repetições de um único tipo de estrutura secundária; enquanto as proteínas globulares possuem a cadeia polipeptídica geralmente enovelada em um formato esférico e composta por uma combinação de estruturas secundárias [Nelson & Cox, 2018]. Ha também uma variação em relação ao papel biológico de cada grupo. Proteínas fibrosas costumam ter função estrutural, como garantir forma e proteção aos vertebrados, já as proteínas globulares tendem a ter papeis na maquinaria biológica, como enzimas, anticorpos e proteínas reguladoras [Nelson & Cox, 2018]. O numero de conformações estruturais que uma proteína globular pode assumir tem se mostrado extremamente limitado, atualmente, mesmo com mais de 165 mil estruturas depositadas no PDB, existem apenas 1378 enovelamentos depositados no SCOP [Andreeva et al., 2020].



a) PDB: 2XDJ b) PDB: 2GV0

Figura 2.4: Exemplos de estruturas de uma proteína fibrosa e globular. A) Estrutura do domínio N-terminal da proteína coordenadora da divisão celular CpoB de E. coli. B) Lisozima de tipo C da tartaruga de carapaça mole chinesa.

Algumas proteínas possuem mais de uma cadeia polipeptídica ou interagem com outras macromoléculas biológicas formando complexos proteicos (Figura 2.5). O arranjo destes complexos no espaço tridimensional é denominado estrutura quaternária. Complexos proteicos podem agir sinergicamente para constituir novos papeis impossíveis de serem realizados pelas subunidades isoladas [Berg et al., 2002].

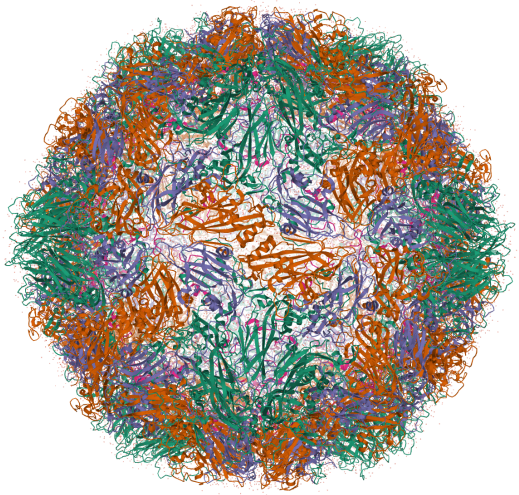


Figura 2.5: Estrutura do capsídeo do rinovírus B14 composto por 60 cópias de cada uma de suas 4 subunidades (PDB: 1k5m)

2.2 Famílias de proteínas

A família de uma proteína consiste de um grupo de proteínas que compartilham uma origem evolutiva, fato que comumente resulta em uma alta similaridade global entre as sequências, estrutura e função. Famílias de proteínas são constantemente organizadas de uma forma hierárquica, utilizando termos como superfamília para agrupar proteínas evolutivamente mais distantes, algumas vezes até mesmo não detectáveis por similaridade de sequências, apenas por homologia estrutural [Dayhoff et al., 1975]; e subfamília agrupando proteínas cada vez mais próximas e que compartilhem uma mesma função. Um exemplo desta organização hierárquica pode ser visto na figura 2.6, a superfamília dos Receptores acoplados a proteína G pode ser dividida em 6 famílias: classe A, B1, B2, C, F e *Taste 2*; de acordo com a similaridade entre as sequências. Cada uma destas famílias é então dividida em mais dois níveis hierárquicos de acordo com o tipo de ligante e função.

Proteínas de uma mesma família podem ser obtidas utilizando ferramentas de busca por sequências baseado em similaridade, como o BLAST [Altschul et al., 1990], HMMER [Eddy et al., 1995], OrthoMCL [Li et al., 2003a] e Orthofinder [Emms & Kelly, 2015]. Existem também uma séries de bancos de dados públicos que permitem o acesso a classificações de famílias de proteínas em termos de sequência e estrutura,

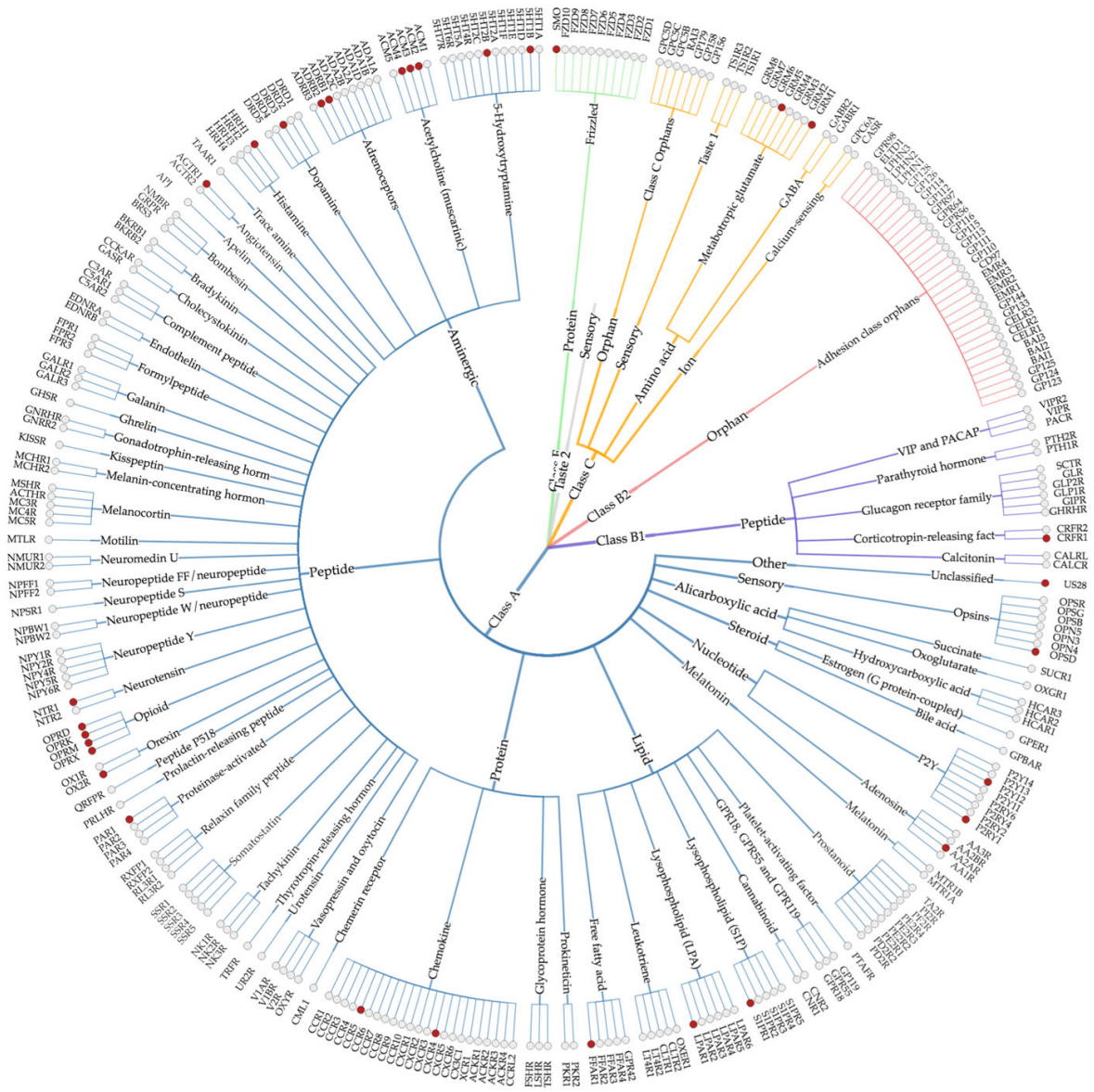


Figura 2.6: Superfamília dos Receptores acoplados a proteína G [Munk et al., 2016]

como o Pfam [El-Gebali et al., 2019], PROSITE [Hulo et al., 2006], InterPro [Mitchell et al., 2019], SUPERFAMILY [Gough et al., 2001], SCOP [Andreeva et al., 2020] e CATH [Sillitoe et al., 2019].

2.3 Domínios proteicos

Domínios são subunidades de proteínas com capacidade de se enovelar e evoluir de forma independente, consequentemente podem também possuir função ou interação especifica. Proteínas podem ter múltiplos domínios funcionais, como no caso da enzima

piruvato cinase, que possui um domínio de ligação a nucleotídeos, um domínio de interação com o substrato e um domínio regulatório [George & Heringa, 2002]. Além disto, domínios podem estar presentes em proteínas com funções distintas, como no caso do domínio SH3, uma região de aproximadamente 60 aminoácidos que está presente em uma grande variedade de proteínas variadas [Musacchio et al., 1992]. A sequência de ocorrência dos domínios de uma proteína é denominado arquitetura (Figura 2.7).

Figura 2.7: Variedade de arquiteturas distintas que possuem o domínio SH3, denotado por um quadrado verde. A figura mostra apenas algumas poucas dezenas de um total de 1.812 arquiteturas encontradas no Pfam.

2.4 Alinhamentos múltiplos de sequências

O Alinhamento múltiplo de sequências (AMS) é um modelo clássico da biologia computacional para visualizar e extrair padrões evolutivos a partir de um conjunto de 3 ou mais sequências biológicas. O modelo constitui-se de um conjunto de sequências, seja proteína, RNA ou DNA, alinhadas em um formato matricial de maneira em que as respectivas posições de cada sequência assumam a mesma coluna na matriz. Durante

o processo de construção do alinhamento, o algorítmo buscará minimizar o número de inserções (geralmente denotado pelos caracteres -"ou ".") ao mesmo tempo que maximiza o número de posições equivalentes. Ao final deste processo, todas as sequências terão o mesmo comprimento. Caso as sequências possuam uma relação homológica, é possível dizer que os padrões de variabilidade de aminoácidos em cada coluna representam uma manifestação de substituições sob restrições impostas pela função [Dima & Thirumalai, 2006]. Contudo, um AMS nos traz uma história evolutiva de acordo com eventos de pressão evolutiva, mutações, recombinação e deriva genética [Valdar, 2002]. Segundo Ferreira [2005], se duas sequências distintas puderem ser alinhadas com um grau considerável de identidade, é possível assumir que elas compartilharam um ancestral em comum em algum momento do tempo. AMS possuem uma grande variedade de aplicações, que vão desde reconstrução de árvores filogenéticas até predição de estruturas terciárias. A figura 2.8 ilustra alguns exemplos de aplicações.

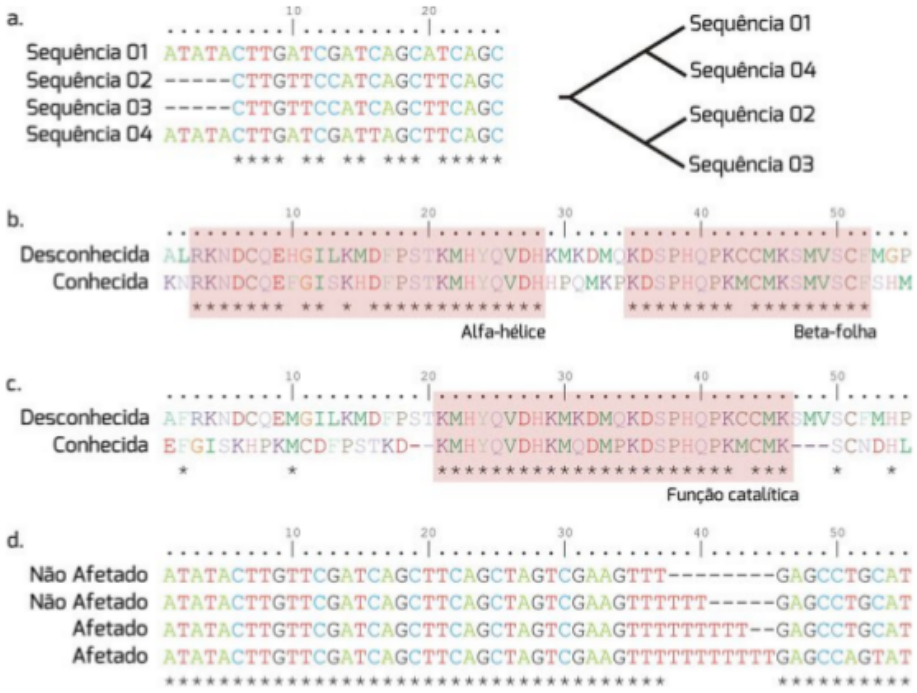


Figura 2.8: Exemplos de aplicações de alinhamentos de sequências. a) Reconstrução de uma arvore filogenética a partir de um AMS de nucleotídeos. b) predição de estruturas secundarias a partir da comparação com um homologo cuja estrutura é conhecida. c) Predição de resíduos funcionais por comparação com um homologo conhecido. d) predição de relação de mutações relacionadas a doenças utilizando alinhamento de sequências de pacientes e grupo controle.

2.5 Sítios determinantes de especificidade

Sítios determinantes de especificidade (SDSs - *specificity determinant sites*) consistem de grupos de resíduos extremamente conservados em uma subfamília de proteínas, porém muito pouco frequente nas outras. Chakraborty & Chakrabarti [2014] definiram três tipos de sítios de determinantes de especificidade. O tipo 1 é referente a divergência funcional, quando as subfamílias possuem diferentes restrições evolutivas, ou seja, os grupos determinantes de especificidade de cada subclasse podem ser compostos por aminoácidos em diferentes posições. Este tipo de divergência é comum entre subfamílias de proteínas com funções distintas. O tipo 2 ocorre quando a posição é conservada em mais de uma subfamília, porém o aminoácido que define a especificidade varia conforme cada subclasse. Este tipo de especificidade ocorre com maior frequência em enzimas, estando geralmente associado a especificidade em relação ao ligante. E finalmente, o tipo 3 é relativo aos resíduos marginalmente conservados, isto é, a especificidade está relacionada a alguma propriedade físico-química ou estrutural [Chakraborty & Chakrabarti, 2014].

2.6 Grafos

A aplicação de modelagem de redes para descrever e observar padrões entre entidades é extremamente utilizada nos dias de hoje nos mais diversos contextos e áreas. Porém, sua origem data do século XVIII, quando Euler elaborou um modelo matemático para resolver um problema clássico conhecido como "Sete Pontes de Königsberg" [Euler, 1736]. O problema era baseado na antiga cidade de Königsberg, atual Kaliningrado na Rússia. A cidade é cortada por um rio formando um complexo de quatro territórios que eram conectados por sete pontes, como pode-se observar na Figura 2.9A. O desafio consistia em obter um caminho o qual fosse possível percorrer cada uma das sete pontes, sem que houvesse nenhuma repetição. Euler modelou o problema representando cada território como um nó e cada ponte como uma aresta conectando um par de territórios. Além disso, ele observou que caso houvesse um caminho que resolvesse este problema, os nós com número ímpares de conexões deveriam ser os territórios de partida ou de chegada. Como todos os nós possuíam um número ímpar de conexões, tal caminho seria impossível. Esta resolução entrou para história por ser o primeiro problema matemático a ser resolvido por teoria dos grafos [Barabási, 2016].

No século XIX, a teoria dos grafos recebeu uma maior atenção, diversos trabalhos seminais foram publicados, como: os estudos de ciclos em poliedros por Kirkman et al. [1856] e Hamilton [1856], que levaram ao surgimento do conceito de caminho

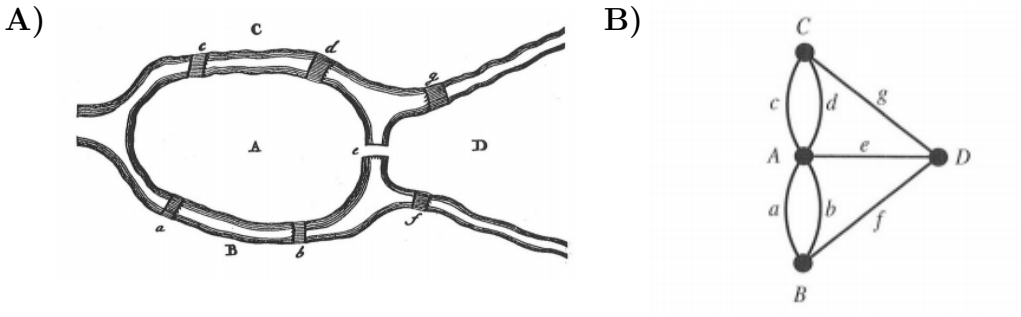


Figura 2.9: Ilustrado em A) está o diagrama de Euler, com os territórios rotulados por caracteres maiúsculos e as pontes por caracteres minúsculos. Em B) a representação em forma de grafo [Hopkins & Wilson, 2004].

hamiltoniano; a conceituação de árvores, como sendo um grafo conectado sem ciclos e aplicação no cálculo de correntes envolvendo circuitos elétricos [Kirchhoff, 1847] e as primeiras aplicações de teoria dos grafos à sistemas biológicos, para representar e enumerar moléculas [Cayley, 1874; Sylvester, 1878; Rouvray, 1989; Gupta et al., 2010].

Um grafo pode ser definido como um par de conjuntos $G = (V, A)$, no qual V representa uma lista de vértices, e A uma lista de arestas formada por pares de V. As arestas de um grafo podem ser direcionadas, indicando uma relação unilateral entre seus pares de vértices, ou não-direcionadas, representando uma relação bilateral. As arestas também podem ser ou não ponderadas, possibilitando a utilização de pesos para quantificar a intensidade das relações entre os pares de vértices. Arestas também podem ser discretizada por tipos, neste caso o grafo é denominado multigrafo.

Existem três formas básicas de representar um grafo. A forma mais comum é a representação visual, como na figura 2.9B. Esta representação consiste em retratar os nós como pontos ou círculos conectados por arestas que geralmente são denotadas por setas ou linhas, caso o grafo seja direcionado ou não. No caso dos multigrafos, os tipos de arestas são comumente diferenciados por cores ou linhas tracejadas. No caso de grafos ponderados, uma possibilidade é discretizar os pesos de acordo com a espessura das arestas. A segunda forma de representar um grafo é a partir de uma lista de arestas. Este formato não permite extrair muita informação, porém é bastante utilizado para leitura e armazenamento das redes. O terceiro formato consiste de sua matriz de adjacência. Uma matriz $A(G) = [a_{ij}]$ de tamanho NxN, sendo N o número de nós da rede. Caso o grafo não seja ponderado, $A(G)$ será uma matriz binária, com $a_{ij} = 1$ quando houver uma aresta entre os nós i e j e $a_{ij} = 0$ quando não houver conexões. Se houver a ponderação de arestas, $a_{ij} = w_{ij}$, sendo w_{ij} o peso da aresta que liga os nós i e j. Está representação é bastante útil, pois facilita o cálculo de diversas métricas. Por exemplo, para se calcular o grau, número de vizinhos, de um nó em um

grafo não ponderado, basta somar todos os valores da coluna do respectivo nó.

2.7 Ciência das redes

Em 1959, Erdös & Rényi publicaram a teoria dos grafos aleatórios, propondo que sistemas complexos poderiam ser efetivamente aproximados por um grafo, cujo os nós estariam conectados de forma aleatória [Erdös & Rényi, 1959]. Erdös & Rényi observaram inclusive o surgimento de diferentes padrões, como: conectividade, ciclos, árvores, subgrafos completos e componentes conexos, de acordo com a probabilidade do conexão dos nós utilizada na rede. O estudo de sistemas complexos através de sua modelagem por grafos ficou conhecido com ciência das redes.

Por mais de 40 anos a ciência tratou as redes complexas (sistemas complexos representados por grafos) como sendo completamente aleatórias [Barabási & Bonabeau, 2003]. Em 1998, Barabási & Albert desenvolveram um projeto com o objetivo de mapear a Internet. Eles modelaram cada página como um nó, e as arestas representavam *links* entre pares de páginas. Os autores esperavam obter uma rede aleatória, uma vez que as pessoas seguem exclusivamente seus próprios interesses ao decidir quais páginas vincular e dado a enorme quantidade e variedade de páginas disponíveis. Porém, o que foi observado, era que mais de 80% das páginas possuíam menos de 4 conexões, enquanto apenas 0,01% possuíam mais de 1.000. Os autores então constataram que a distribuição de graus da rede seguia uma lei da potência, ou seja, a probabilidade de um nó ter k conexões, $P(k)$, diminui de acordo com que o valor de k aumente, seguindo a equação $P(k) \sim k^{-\gamma}$, onde γ representa o expoente livre da escala e determinante de $P(k)$. Este modelo ficou conhecido como redes livre de escala [Barabási & Albert, 1999] e mostrou-se ser uma aproximação muito mais verossímil do que o antigo modelo das redes aleatórias, sendo posteriormente confirmado em uma gama de sistemas reais, como: redes de co-citações em revistas acadêmicas [Eom & Fortunato, 2011], estrutura física da internet [Percacci & Vespignani, 2003], transporte aéreo americano [Guimera et al., 2005], redes de interações amorosas [Liljeros et al., 2001], propagação de epidemias [Pastor-Satorras & Vespignani, 2001], redes metabólicas [Ma & Zeng, 2003], redes de atividade cerebral [Hanson et al., 2016], redes de co-expressão genica [Gibson et al., 2013], interação proteína-proteína [Jeong et al., 2001], entre muitas outras. Apesar de tudo, o modelo de Erdös & Rényi ainda é bastante utilizado na literatura como um modelo nulo para validação estatística. A figura 2.10 compara os modelos de rede aleatória e livre de escala aplicado ao transporte aéreo americano.

Sistemas reais, além de não serem aleatórios e possuírem *hubs*, nós cujo número de

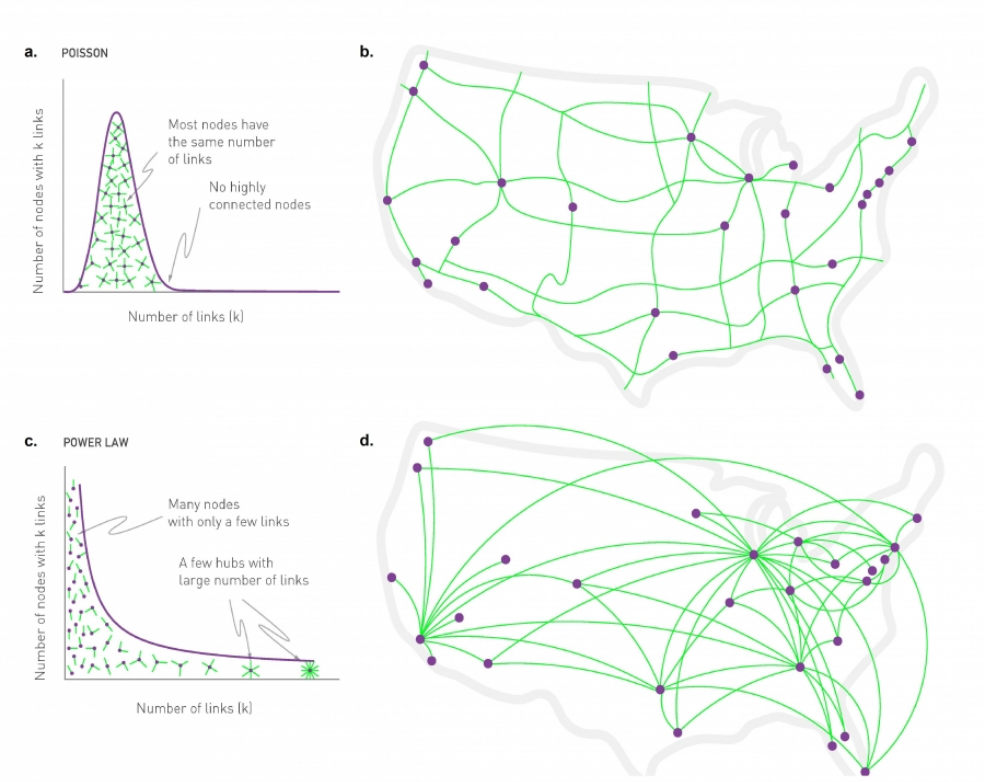


Figura 2.10: Sistema de transporte aéreo americano ilustrado como uma rede aleatória e como uma rede livre de escala. Na rede aleatória, a maioria dos nós possuem o mesmo número de conexões, portanto sua distribuição de graus pode ser aproximada por uma distribuição de Poisson. Já na rede livre de escala, a conectividade dos nós segue a lei da potência, portanto sua distribuição de graus pode ser aproximada por uma distribuição exponencial. [Barabási, 2016].

conexões é muito acima da média, também costumam possuir padrões de comunidades (Figura 2.11). Em ciência das redes, uma comunidade se refere a um conjunto de nós que possuem um número muito maior de conexões entre si, do que em relação aos outros nós da rede. Existem centenas de algoritmos para detecção de comunidades em redes, utilizando-se das mais variadas heurísticas e focando em diferentes propósitos [Fortunato & Hric, 2016]. As principais métricas para se estudar comunidades em redes são o coeficiente de agrupamento, geralmente referido ao seu nome em inglês: *clustering coefficient* e a modularidade. O coeficiente de agrupamento de um nó consiste em quantificar o quanto seus vizinhos estão conectados entre si, essa métrica também pode ser expandida para um nível global da rede através do cálculo do número de trios de nós fechados sobre o número total de trio de nós conectados (número de trios abertos mais número de trios fechados). A modularidade é uma métrica para avaliar a conectividade das partições de uma rede. Existem diversas formas descritas para

o cálculo da modularidade, porém basicamente ele leva em consideração a fração de arestas que se encontram dentro de uma partição e o número esperado de arestas dentro da partição dado uma rede aleatória [Fortunato & Hric, 2016].

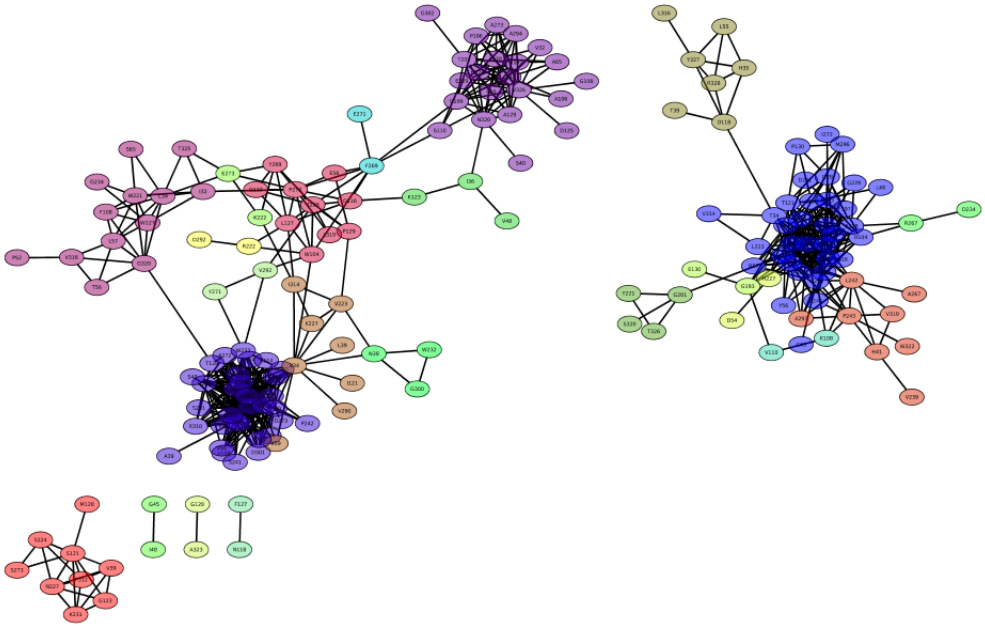


Figura 2.11: Padrão de comunidades em uma rede de coocorrência de resíduos em proteínas da família das HIUases e Transtirretinas. Imagem gerada com auxílio da ferramenta Cytoscape [Shannon et al., 2003].

2.8 Redes bipartidas

Grafos bipartidos, também chamado de redes bipartidas ou redes de afiliação, quando representam um sistema complexo, (Figura 2.12), consistem em uma classe específica de grafos, no qual seus vértices podem ser divididos em dois conjuntos disjuntos e independentes U e V, de forma que todas as arestas do grafo conectem um nó do conjunto U a um nó do conjunto V [Neal, 2014]. Grafos bipartidos podem ser representados por uma matriz de biadjacência. Este tipo de matriz funciona de forma similar às matrizes de adjacência, porém cada eixo da matriz será representado por um dos conjuntos U e V. Modelos de redes bipartidas são amplamente utilizados na modelagem de sistemas de co-ocorrência, nos mais diversos contextos: compostos químicos co-ocorrentes em alimentos [Ahn et al., 2011], genes co-associados a doenças [Barabási et al., 2011], co-ocorrência de micróbios em ecossistemas [Connor et al., 2017], atores que co-atuaram em filmes de Hollywood [Watts & Strogatz, 1998], cidades

co-hospedando subsidiárias multinacionais [Taylor, 2001], pessoas que participam dos mesmos grupos sociais [Neal & Neal, 2013], similaridade entre organismos baseado na co-ocorrência de proteínas ortólogas em seus genomas [Tumminello et al., 2011], entre muitos outros.

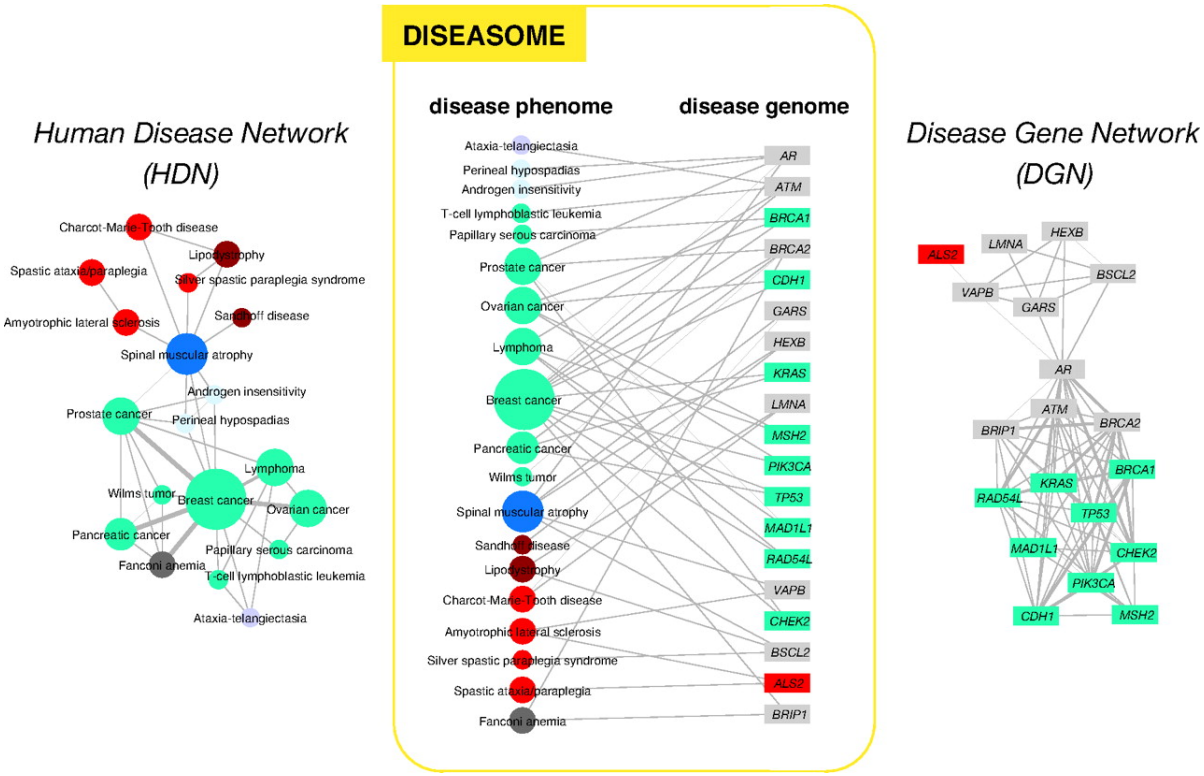


Figura 2.12: No centro é possível observar um trecho da rede bipartida de doenças e genes associados gerada por Goh et al. [2007]. Na esquerda está ilustrado a rede projetada para o conjunto de doenças (HDN), já na direita a projeção gerada utilizando o conjunto de genes (DGN). É possível observar que alguns genes tendem a estarem coassociados à oncologias, formando uma comunidade com alto coeficiente de agrupamento.

Mapeamento de co-ocorrências em redes bipartidas são geralmente realizadas através de análises em sua projeção monopartida (Figura 2.12). A projeção monopartida, também chamada de rede de co-afiliação, consiste em uma rede contendo apenas os nós de um dos conjuntos U e V da rede original. Estes nós são conectados desde que compartilhem pelo menos um vizinho no grafo bipartido e geralmente é utilizado a ponderação de arestas de acordo com o número de vizinhos compartilhados [Saracco et al., 2017]. Porém, a geração de projeções tende a produzir redes extremamente densas e não bastaria a simples aplicação de um *threshold*. De acordo com Neal, a aplicação de cortes simples em redes projetadas possui três principais deficiências: o viés de arbitrariedade, ou seja, a utilização de um valor simplesmente arbitrário no corte;

o viés estrutural, Watts demonstrou que a aplicação de um *threshold* incondicional irá sempre produzir redes com alto coeficiente de agrupamento, não pelas características estruturais da rede, mas por um viés gerado pela remoção de arestas; e finalmente, o viés de não-escalaridade, uma que vez que os pesos das arestas da projeção são diretamente correlacionados ao seus respectivos graus na rede bipartida. Os grupos que co-ocorrem em quantidades menores seriam simplesmente descartados. Existem na literatura diversas abordagens para normalizar e selecionar arestas estatisticamente relevantes, com o intuito de contornar este problema de esparsificação da rede [Serrano et al., 2009; Borgatti & Halgin, 2011; Tumminello et al., 2011; Neal, 2014; Dianati, 2016; Saracco et al., 2017].

Capítulo 3

Metodologia

A primeira etapa da metodologia consiste em descrever todas as etapas do algorítmo proposto neste trabalho para detecção de sítios determinantes de especificidades [da Fonseca Jr et al., 2019]. Conforme representado no fluxograma presente na figura 3.1, a metodologia proposta pode ser dividida em três principais etapas: pré-processamento de dados, com o objetivo de remover possíveis vieses e ressaltar as informações presentes no alinhamento de entrada; modelagem da rede, que consiste em transformar um alinhamento múltiplo de sequências em uma rede de coevolução de resíduos; e finalmente a detecção de comunidades, onde a rede de correlação e clusterizada e os possíveis SDS são identificados. Cada uma destas etapas sera explicada detalhadamente no decorrer deste capítulo.

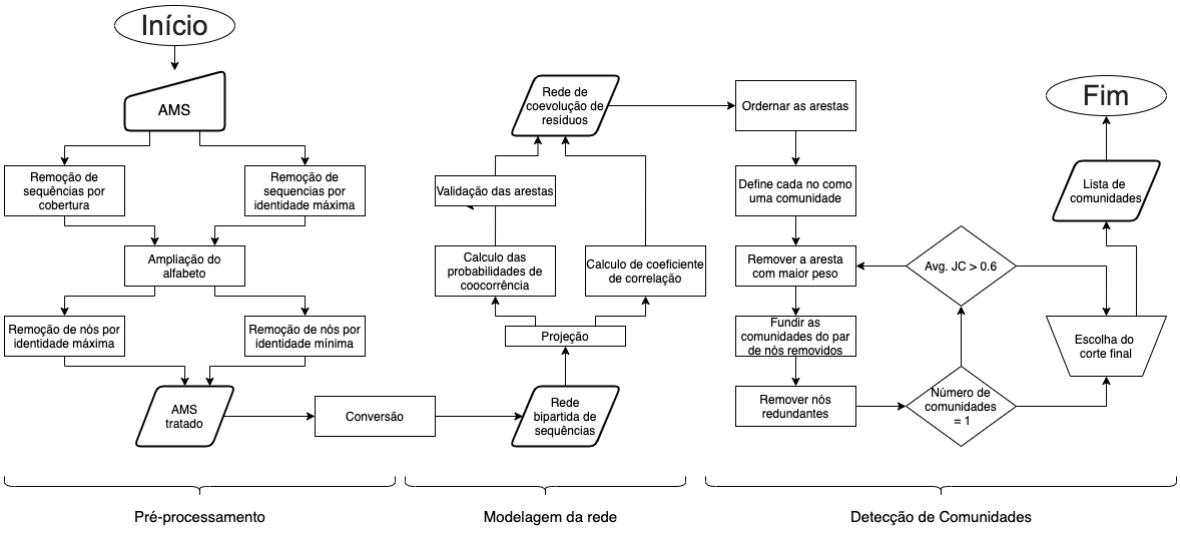


Figura 3.1: Fluxograma representando todas as etapas do algorítmo proposto para detecção de sítios determinantes de especificidade

3.1 Pré-processamento

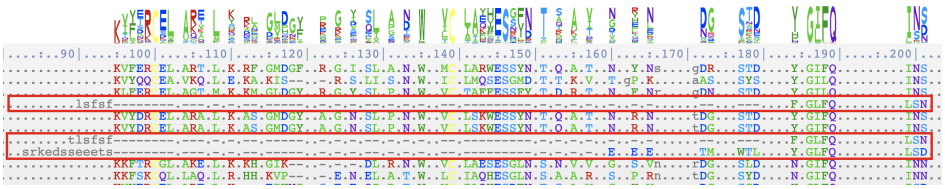
A etapa de processamento é fundamental neste tipo de metodologia, afinal a qualidade do alinhamento de entrada está diretamente relacionada a acurácia dos resultados obtidos. O alinhamento deve ser constituído por sequências potencialmente homólogas e em número expressivo para que se tenha amostragem suficiente para validação estatística. Problemas possivelmente presentes no alinhamento, como a presença de fragmentos de sequências, regiões mal alinhadas, sequências redundantes, baixa amostragem e até mesmo complexidade podem ser aliviados ao aplicar determinados filtros.

A primeira fase do pré-processamento é constituída pela aplicação de dois filtros em nível de sequência: remoção de sequências por cobertura e por identidade máxima.

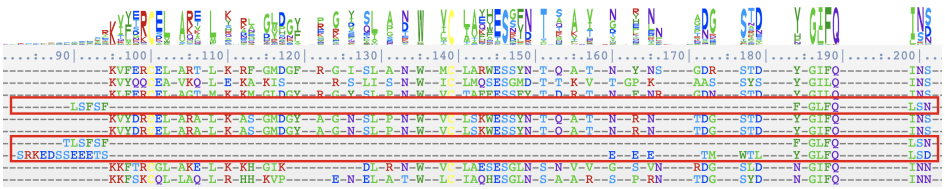
3.1.1 Filtro por cobertura

O filtro de cobertura de sequências tem como objetivo a remoção de fragmentos que possam estar presentes no alinhamento. Este filtro pode ser executado de duas maneiras, através do perfil HMM ou por comprimento médio de sequências.

Alguns alinhamentos incluem dados adicionais relativos ao perfil HMM utilizado em sua construção. Nestes casos, aminoácidos em posições válidas do alinhamento são representados por caractere maiúsculo e aminoácidos que geraram inserções no modelo são representados por caractere minúsculo. De forma semelhante, os *gaps* também são diferenciados: traços ('-') são usados para denotar a ausência de aminoácido em uma posição que o perfil HMM esperava encontrar; e pontos ('.') são usados para preencher as posições de inserção ou deleção (*indels*).



a) Remoção baseada no perfil HMM



b) Remoção baseada no comprimento médio

Figura 3.2: Exemplo de remoção de fragmentos de sequências através do perfil HMM (a) e do comprimento médio das sequências (b).

Caso o alinhamento de entrada possua dados do perfil HMM, sequências serão removidas caso o número de aminoácidos em posições válidas do alinhamento dividido pelo total de posições válidas seja menor que o valor de corte predefinido. Caso esse tipo de informação não esteja disponível, as sequências serão descartadas se o seu comprimento for menor que o comprimento médio das sequências no alinhamento multiplicada pelo valor de corte.

3.1.2 Filtro por identidade máxima

A remoção de sequências por identidade máxima tem o objetivo de remover falsos sinais de correlação causados pelo acúmulo de sequências com alta identidade no AMS, como por exemplo uma grande quantidade de ortólogos muito próximos. Este tipo de viés pode ter diversas origens, como a filogenética, uma vez que sequências evolutivamente mais recentes podem ter uma identidade global muito alta, simplesmente pelo fato de ainda não terem acumulado novas mutações. Outro possível motivo é o interesse acadêmico, tendo em vista que proteínas de maior interesse acadêmico ou farmacológico são sequenciadas com maior frequência, e portanto, presentes em uma maior proporção no bancos de dados. Neste trabalho, esta etapa é realizada com o auxilio da ferramenta CD-HIT [Fu et al., 2012].

3.1.3 Ampliação de alfabeto

É conhecido que algumas substituições de aminoácidos, mesmo em regiões de sitio funcional, podem não afetar o enovelamento ou a atividade de uma proteína, isto porque o tipo de aminoácido substituído manteve as propriedades básicas para o desempenho da atividade ou para compor a cavidade estrutural. Logo faz sentido tentar buscar por padrões de coocorrência de resíduos não somente em função de aminoácidos, mas também em razão de propriedades físico-químicas e estruturais.

O PFstats, software embrião deste projeto, utilizava reduções de alfabetos para tentar detectar este tipo de padrão [Fonseca-Júnior et al., 2018]. Porém, apesar do baixo custo computacional, esta abordagem falha em detectar correlações entre grupo de níveis hierárquicos diferentes, como por exemplo uma correlação entre uma alanina na posição X com um resíduo hidrofóbico na posição Y. Pensando nisto, neste trabalho foi aplicado uma metodologia oposta, a ampliação do alfabeto de aminoácidos. Sendo assim, é possível expandir o alfabeto de aminoácidos a serem analisados, o que permite realizar analises mais profundas a um custo de aumento na complexidade computacional e consequentemente tempo de execução. Logo o método buscará por correlações

não apenas entre resíduos de aminoácidos, mas também incluindo resíduos formados por grupos de aminoácidos representando propriedades. O alfabeto expandido, que pode ser visto na tabela 3.1 foi construído a partir de uma fusão de diversas representações de alfabetos reduzidos presentes na literatura [Wang & Wang, 1999; Murphy et al., 2000; Li et al., 2003b; Betts & Russell, 2003; Pommié et al., 2004].

Grupo	**Alfabeto de Aminoácidos**
Amida	N e Q
Alifático	G, A, V, L e Y
Básico	H, K e R
Hidroxila	S, T e Y
Enxofre	C e M
Não-Polar	F, G, V, L, A, I, P, M e W
Polar	Y, S, N, T, Q e C
Hidrofóbico	L, I, F, W, V e M
Hidrofílico	R, K, N, Q, P e D
Pos. Carregado	K e R
Neg. Carregado	D e E
Muito Pequeno	G, A e S
Pequeno	C, D, N, P e T
Médio1	E, V, Q e H
Médio2	M, I, L, K e R
Aromático/Grande	F, Y e W
ND	N e D
QE	Q e E

Tabela 3.1: Tabela de representação de super nós.

3.1.4 Filtro de resíduos

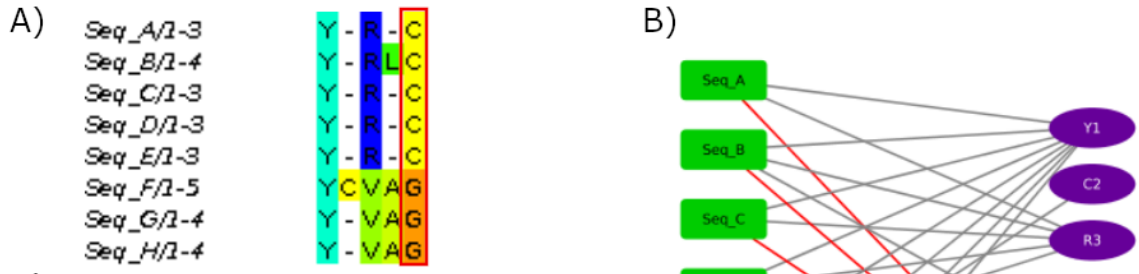
A última etapa do pré-processamento consiste em ignorar do cálculo, todos os resíduos de acordo com um limiar de frequência máxima e mínima. Esta etapa é executada após a ampliação do alfabeto pelo fato de também ser aplicada aos resíduos adicionais. Portanto, caso hipoteticamente uma asparagina na posição X esteja abaixo do limiar, porém ao incluir as glutaminas, a frequência atinja o limiar, o método irá descartar o resíduo Asn$X$, porem o resíduo Amida$X$ será mantido. Estes filtros são aplicados na matriz de distâncias, logo mesmo após remover um resíduo, sua frequência ainda é levada em consideração nos resíduos hierarquicamente superiores.

Filtros de resíduos são completamente opcionais, os grupos extremamente conservados são removidos por um motivo simplesmente computacional, uma vez que sua presença não impacta na qualidade dos resultados. Manter estes nós na rede aumenta

exponencialmente o número de arestas, sendo que resíduos conservados podem ser calculados por métricas muito mais baratas. Já a filtragem por frequência mínima é aplicada com o objetivo de remover possíveis ruídos. Nós com a frequência extremamente baixa podem gerar falsos positivos simplesmente pela falta de amostragem suficiente.

3.2 Modelagem da Rede

A modelagem das redes utilizadas neste trabalho parte da observação de um alinhamento múltiplo de sequências como uma rede bipartida, no qual o conjunto de nós U é composto pelos identificadores das sequências presentes no alinhamento, e o conjunto de nós V é formado por todos os possíveis resíduos, isto é, aminoácido seguido por sua posição no AMS (figura 3.3A). Também é incluso nesta rede os nós referentes aos resíduos adicionais obtidos na etapa de expansão do alfabeto. A rede bipartida por si só não é tao informativa, mas a sua versão matricial, chamada de matriz de biadjacência (figura 3.3C), transforma cada resíduo em um vetor binário, facilitando o calculo de uma serie de métricas de coocorrência.

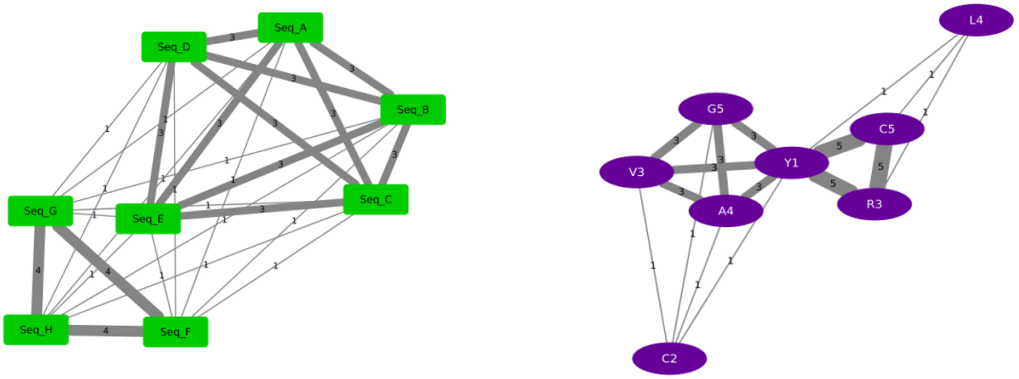


C)

	Y1	C2	R3	V3	L4	A4	C5	G5
Seq_A	1	0	1	0	0	0	1	0
Seq_B	1	0	1	0	1	0	1	0
Seq_C	1	0	1	0	0	0	1	0
Seq_D	1	0	1	0	0	0	1	0
Seq_E	1	0	1	0	0	0	1	0
Seq_F	1	1	0	1	0	1	0	1
Seq_G	1	0	0	1	0	1	0	1
Seq_H	1	0	0	1	0	1	0	1

Figura 3.3: Três formas de se representar um mesmo alinhamento múltiplo de sequências. A) forma tradicional matricial, B) Grafo bipartido de rótulos e resíduos e C) matriz de biadjacência. Em vermelho esta destacado como é representado uma única posição do alinhamento em cada uma das três formas.

O grafo bipartido gerado pelo alinhamento múltiplo de sequências pode ser projetado em duas novas redes. Ao conectar todos os nós do conjunto U (rótulos das sequências) que compartilhem ao menos um nó de V (resíduos), será obtido a projeção de sequências, uma rede onde pares de sequências são conectadas de acordo com a tendência em possuir os mesmos padrões de resíduos (figura 3.4a). Ao realizar o mesmo procedimento no conjunto oposto, será obtido uma rede de resíduos de aminoácidos conectados de acordo com a tendência de coocorrerem nas mesmas sequências (figura 3.4b). Em ambos os casos, também é realizado uma ponderação das arestas conforme o numero de nós no conjunto oposto compartilhados pelo par de nós do conjunto a ser projetado, no caso da projeção de V, cada par de resíduos receberá um peso equivalente ao numero de sequências distintas que possuem ambos os resíduos. Neste trabalho será abordado apenas análises envolvendo a projeção de V.



a) Projeção de sequências (U)

b) Projeção de resíduos (V)

Figura 3.4: Duas possíveis projeções do grafo bipartido de sequências alinhadas: a) rede de sequências que tendem a ter os mesmos resíduos e b) rede de coocorrência de resíduos no alinhamento.

3.3 Validação de Arestas

Apesar da figura 3.4 ilustrar dois exemplos pequenos e de alta simplicidade, no mundo real projeções tendem a gerar redes extremamente densas (figura 3.5), afinal uma única coocorrência já é suficiente para que seja adicionado uma aresta entre pares de resíduos. Logo, antes de analisada, essa rede precisa ser filtrada para remover as correlações fracas. Porém, conforme já abordado, redes projetadas não são normalizadas e consequentemente a aplicação de simples cortes lineares tendem a não ser eficazes. Imagine o caso de uma rede de coocorrência de resíduos gerada a partir de um alinhamento múltiplos de sequências compostos por três principais subfamílias: A, presente

em 80% das sequências, B, presente em 15% das sequências, e C, presente em apenas 5% das sequências. Mesmo que os resíduos determinantes de especificidade de C sejam 100% conservados, um simples corte de $0.05 \cdot N$ seria suficiente para remover todos os sinais de C e da mesma forma, um corte de $0.15 \cdot N$ removeria todos os sinais de B e C. Logo, as arestas de uma rede projetada precisam ser previamente validadas por métodos probabilísticos e/ou normalizadas através de coeficientes de correlação para somente então ser plausível de esparsificação.

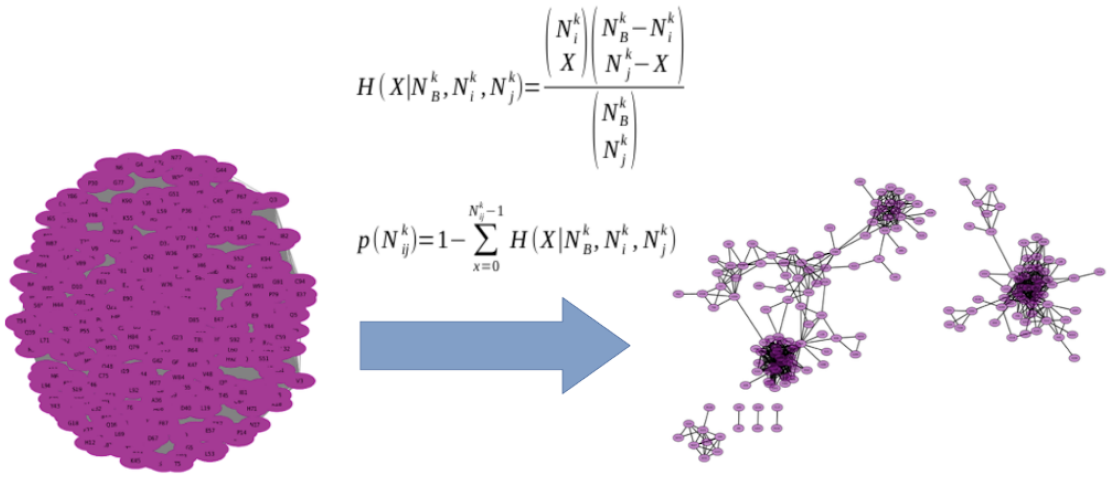


Figura 3.5: Exemplo da aplicação do teste de Tumminello et al. [2011] para transformar uma projeção quase completa em uma rede livre de escala.

Neste trabalho foram avaliados sete abordagens para normalização e validação de arestas, detalhadas abaixo.

3.3.1 Filtro de Disparidade (DF)

O método foi originalmente proposto por Serrano et al. [2009] com o objetivo de selecionar arestas estatisticamente relevantes em redes naturalmente ponderadas, porém já foi aplicado em projeções monopartidas extraídas de redes bipartidas [Ahn et al., 2011]. Para cada aresta, é computado a probabilidade do peso observado ser maior do que o valor esperado, dado um modelo nulo, no qual os pesos sejam aleatoriamente redistribuídos por todas as arestas do grafo. Na equação 3.1, O_{ij} representa o peso observado entre os nós i e j, D_i consiste do grau (número de conexões) do nó i, e finalmente p_{ij} a proporção de O_{ij} em relação a força do nó i. A força de um nó consiste na soma dos pesos de todos os nós acendentes.

$$Pr(O_{ij} > Null_{ij} = 1 - (D_i - 1) \int_0^{p_{ij}} (1 - x)^{D-2} \tag{3.1}$$

3.3.2 Normalizações de Borgatti & Halgin

Borgatti & Halgin apresentam em seu livro uma série de métricas, baseadas em tabelas na contingência, para normalizar pesos de arestas de projeções monopartidas, com o intuito de enfatizar relações de co-ocorrência. Na equação 3.2 (BHN), é proposto a normalização dos pesos em um intervalo de 0 a 1, no qual 1 representa o número máximo de sobreposições, dado o número de sequências compartilhadas pelos nós i e j. Na equação 3.3, é utilizado o coeficiente de Jaccard (JC) [Jaccard, 1912] como forma de quantificar a sobreposição entre pares de nós. Outra métrica proposta pelos autores, é a utilização do coeficiente de correlação de Pearson, equação 3.4, entre as colunas i e j da matriz de biadjacência. Por fim, também é proposto a utilização da equação de Bonacich [Bonacich, 1972], equação 3.5, propondo a normalização dos escores das arestas baseado no cálculo da probabilidade da sobreposição observada entre os nós i e j exceder a sobreposição esperada ao acaso.

$$a^*_{ij} = \frac{a}{Min(a+b, a+c)} \tag{3.2}$$

$$a^*_{ij} = \frac{a}{a+b+c} \tag{3.3}$$

$$c_{ij} = \frac{\frac{1}{m}\sum_k x_{ik}x_{jk} - u_i u_j}{S_i S_j} \tag{3.4}$$

$$P'_{ij} = \frac{a - \sqrt{adbc}}{ad - bc} \quad \text{for} \quad ad <> bc \tag{3.5}$$

3.3.3 Abordagem de Tumminello

A abordagem de Tumminello et al. [2011] consiste em validar cada aresta de uma projeção monopartida através de uma hipótese nula de coocorrência aleatória de vizinhos em comuns, levando em consideração a heterogeneidade dos elementos de ambos os conjuntos da rede bipartida. Portanto, seja N_B^k, o número de nós do conjunto B com grau k; N_i^k e N_j^k, os respectivos graus dos nós i e j na rede bipartida, é possível calcular a probabilidade dos nós i e j compartilharem X sequências, através de uma distribuição hipergeométrica, denotada na equação 3.6. Sendo assim, pode-se associar um p-valor de acordo com o número de sequências que os nós i e j compartilham

(equação 3.7).

$$H(X \mid N_B^k, N_i^k, N_j^k) = \frac{\binom{N_i^k}{X}\binom{N_B^k - N_i^k}{N_j^k - X}}{\binom{N_B^k}{N_j^k}} \tag{3.6}$$

$$p(N_{ij}^k) = 1 - \sum_{x=0}^{N_{ij}^k - 1} H(X \mid N_B^k, N_i^k, N_j^k) \tag{3.7}$$

3.3.4 Filtro de probabilidade marginal *Hairball* (HMLF)

Assim, como a abordagem de Serrano et al. [2009], o filtro *Hairball* [Dianati, 2016], foi originalmente proposto para validar arestas em qualquer tipo de rede ponderada, não sendo especifico para grafos bipartidos. Dado S, a força total da rede, isto é, a soma dos pesos de todas as arestas, pode-se dizer que para cada conexão, a escolha dos dois nós incidentes teria uma probabilidade proporcional a força dos nós. Portanto, a probabilidade de uma aresta com peso w, de um total S, conectar os nós i e j, pode ser calculada através de uma distribuição binomial. O modelo nulo é então definido através da equação 3.8, e o p-value associado através da equação 3.9, sendo, S_i e S_j, as respectivas forças dos nós i e j.

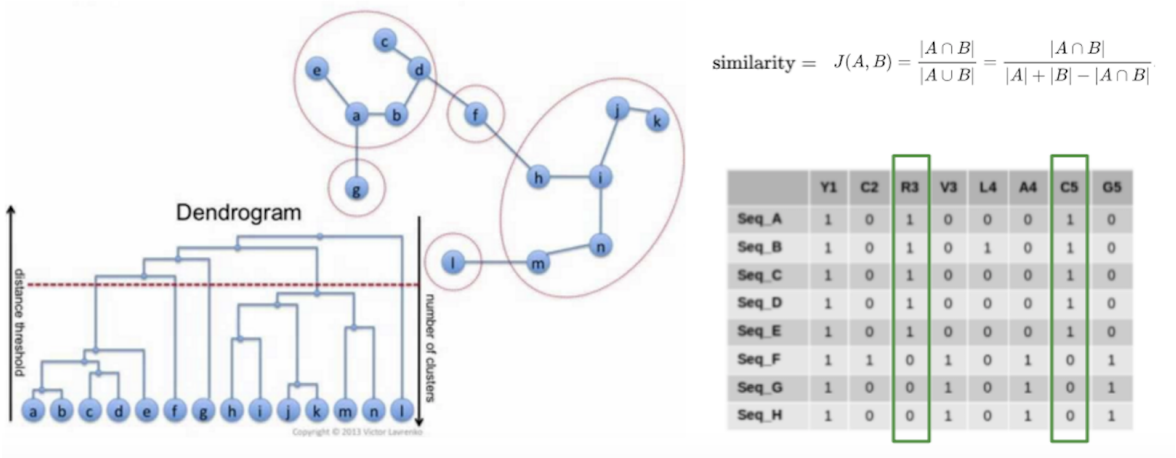
$$Pr(\sigma_{ij} = w \mid S_i, S_j, S) = \binom{S}{w} p^w (1-p)^{S-w} \quad \text{sendo} \quad p = \frac{S_i S_j}{2S^2} \tag{3.8}$$

$$P_{ij}^{'} = 1 - \sum_{w >= w_{ij}} Pr(\sigma_{ij} = w \mid S_i, S_j, S) \tag{3.9}$$

3.4 Detecção de Comunidades

A maioria dos algoritmos de detecção de comunidades em redes tem como principio básico o particionamento em função da maximização da modularidade, isto é, os nós do grafo são includidos em uma das comunidades de modo a maximizar a distribuição de graus interna em cada *cluster* em detrimento da distribuição de graus externos (entre *clusters*) [Fortunato & Hric, 2016]. Porém neste trabalho, a preocupação maior está em obter conjuntos de nós que realmente possuam tendência de coocorrer independente da estrutura de conectividade da rede. Portanto, foi desenvolvido um algoritmo para detecção de comunidades em redes projetadas baseado nas distâncias entre as colunas na matriz de biadjacência.

O algoritmo consiste de um *clustering* hierárquico aglomerativo (figura 3.6). No estágio inicial do algoritmo, cada nó da rede é atribuído a uma comunidade própria. Nas etapas posteriores, é calculado a distância entre cada par de comunidade. Os vetores representantes de cada comunidade são formados pela média das colunas de seus nós na matriz de biadjacência. Os pares de comunidade que apresentam a menor distância são fundidos. O algoritmo finaliza quando não há nenhum par de comunidades cuja distância é menor do que um dado valor (foi utilizado 0.4 durante a validação e aumentado para 0.5 no CONAN). Durante a validação foi utilizado a distância do coseno como métrica de inserção, porém após os resultados, durante a implementação do CONAN, essa métrica foi substituída pelo coeficiente de Jaccard.



	Y1	C2	R3	V3	L4	A4	C5	G5
Seq_A	1	0	1	0	0	0	1	0
Seq_B	1	0	1	0	1	0	1	0
Seq_C	1	0	1	0	0	0	1	0
Seq_D	1	0	1	0	0	0	1	0
Seq_E	1	0	1	0	0	0	1	0
Seq_F	1	1	0	1	0	1	0	1
Seq_G	1	0	0	1	0	1	0	1
Seq_H	1	0	0	1	0	1	0	1

Figura 3.6: *Clustering hierárquico aglomerativo utilizando o coeficiente de Jaccard entre as colunas da matriz de biadjacência como métrica de distância* [Lavrenko, 2014].

Ao expandir o alfabeto de aminoácidos, conexões sinônimas passam a existir na rede. Por exemplo, como ilustrado na figura 3.7, caso haja uma correlação entre uma alanina na posição 10 e um triptofano na posição 20, a expansão do alfabeto irá gerar dois conjuntos de quatro nós cada conectados uns aos outros. O que a primeira vista poderia parecer uma comunidade relativa a grupos determinantes de especificidade, se trata na verdade de um viés causado por uma única correlação. É importante ressaltar que nós sinônimos dentro da rede não são problemáticos, eles só se tornam redundantes quando agrupados em uma mesma comunidade. Portanto, a fim de selecionar uma única aresta representativa, a cada passo da deteccao de comunidades é executado um filtro mantendo em cada vizinhança um único nó correspondente a cada posição. O nó selecionado é aquele que tenha a menor distância média em relação aos seus vizinhos. Caso mais de um nó compartilhem um mesmo valor de distância, será mantido o que representar um menor subconjunto (ver tabela 3.1).

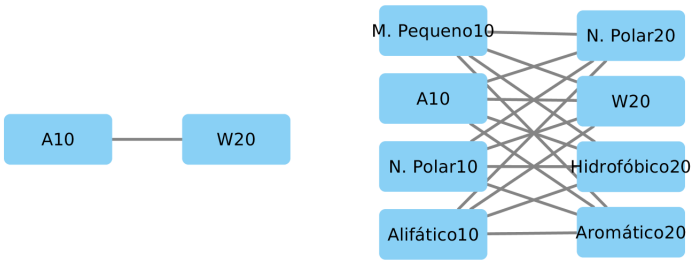


Figura 3.7: Exemplo de conectividade expandida ao incluir nós marginais na rede.

3.5 Conjunto de Dados de Validação

3.5.1 *Benchmark* Artificial

Foi desenvolvido um algoritmo estocástico com o intuito de gerar AMSs aleatórios com padrões de conservação local e designação de subclasses funcionais propositalmente inseridas, sendo assim possível de mapear os SDSs e quantificar a eficácia dos diversos métodos de validação de arestas considerados neste trabalho. Dois tipos de SDS foram considerados neste algorítmo. O primeiro é definido após as sequências serem divididas em subfamílias, padrões gerados de forma aleatória são então inseridos em cada grupo, simulando assim a existência de resíduos determinantes de especificidade entre eles. O segundo tipo de padrão é aleatoriamente distribuído no alinhamento, simulando relações de coevolução, contatos estruturais ou qualquer outro tipo de correlação entre resíduos. Após a definição dos padrões de conservação, é aplicado uma probabilidade de 0.9 para que cada sequência o mantenha, caso contrário, será substituído por outro aminoácido. O restante das posições do alinhamento são preenchidas com aminoácidos conforme uma distribuição de Dirichlet gerada para cada coluna, consequentemente, todas as sequências de cada alinhamento possuirão identidades relativamente altas entre si, simulando uma relação de homologia.

O objetivo deste algoritmo não é de simular proteínas realmente funcionais, mas sim fornecer um *benchmark* para quantificar a eficácia de cada método em mapear as relações de co-ocorrência do alinhamento. Portanto não foi utilizado nenhuma matriz de substituição, e todos os 20 aminoácidos possuem a mesma probabilidade de ocorrência.

A acurácia dos métodos de validação de arestas foi calculada em relação a manter somente os nós e arestas relativos aos padrões previamente inseridos. Para isto, os cortes escolhidos para cada rede foram otimizados com o objetivo de maximizar

duas métricas: O F1 *score* das arestas retidas e uma adaptação da fração de vértices corretamente detectados de Newman (FCDV) [Girvan & Newman, 2002]. A métrica, originalmente proposta para quantificar a eficiência de métodos de detecção de comunidades baseados em particionamento da rede, considera um nó como corretamente detectado caso ele esteja no mesmo grupo que ao menos metade dos nós esperados em sua comunidade real, e se a comunidade detectada consiste em uma fusão de duas ou mais comunidades reais, todos os seus nós são considerados incorretamente detectados. A taxa é finalmente calculada através da divisão do número de nós detectados corretamente pelo número total de nós. O FCDV foi adaptado para avaliar a eficácia dos algoritmos de seleção de arestas. Em uma seleção perfeita, a rede será composta apenas por componentes conexos determinando as comunidades. Logo a taxa foi adaptada para levar em consideração os falsos positivos e falsos negativos, através da divisão do número de nós corretamente detectados por $N_a \cup N_b$, onde N_a representa o número de nós mantidos na rede e N_b, o número de nós que deveriam ter sido mantidos na rede.

3.5.2 *Benchmark* Real

Como forma de validação da metodologia utilizando dados reais, foram conduzidos estudos de casos com a família dos receptores acoplados a proteína G (classe A) (Pfam: PF00001) e com a família das HIUases e Transtirretinas (Pfam: PF00576). Ambos os alinhamentos foram obtidos a partir do Pfam e continham respectivamente 42.500 e 1.955 sequências. O alinhamento PF00576 foi filtrado com parâmetros 0.8 de cobertura e 0.9 de identidade máxima, já para PF00001, pelo fato da família possuir uma grande variedade de subclasses funcionais, foi utilizado um parâmetro de cobertura um pouco menos rigoroso (0.7), assim sequências com uma variação um pouco maior de comprimento ainda serão mantidas no alinhamento. Porém, pelo fato do AMS possuir um número extremamente alto de sequências, o parâmetro escolhido para identidade máxima foi um pouco mais restritivo (0.8). A validação de arestas foi realizada com a abordagem de Tumminello.

A seleção automatizada do valor de corte foi realizada conforme os trabalhos de Borate et al. [2009] e Perkins & Langston [2009], que indicaram a maximização do número de comunidades como uma das formas mais efetivas para automatizar a escolha do corte em redes de correlação. Sendo assim, após normalizar os pesos de todas as arestas da rede, estas são removidas em passos únicos, de forma decrescente. Após cada remoção é realizado a detecção de comunidades. Ao final da rotina, a rede selecionada será aquela que gerou o maior número de comunidades. Este tipo de abordagem parte do pressuposto de que em um cenário inicial, com uma rede extremamente co-

nectada, haverá um único grande componente conexo. Uma vez que as comunidades são altamente conectadas entre si e pouco conectadas com os demais nós, e que após a normalização, é esperado que suas arestas sejam mais fortes do que as pontes (devido a probabilidade de ocorrência ao acaso), ao remover os nós mais fracos, a pontes serão quebradas e as estruturas de comunidade começarão a surgir, fazendo com que o número de agrupamentos cresça. Após atingir, o que seria o corte ideal, as arestas pertencentes a comunidades começarão a serem removidas, porém devido a alta conectividade de seus nós, a probabilidade de divisão acaba sendo baixa, fazendo com que os grupos acabem sumindo por inteiro (Figura 3.8).

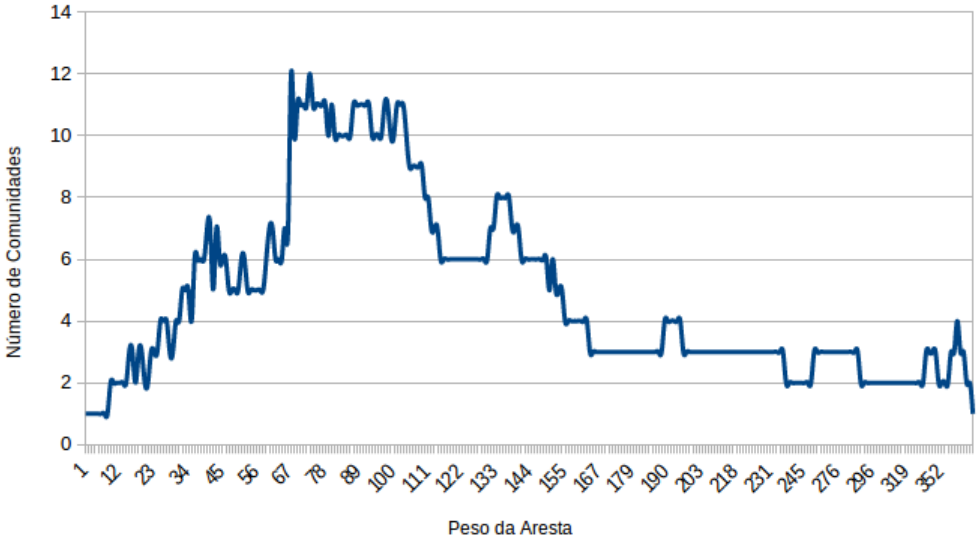


Figura 3.8: Representação gráfica do número de comunidades por corte na rede após a normalização das arestas. É possível observar um crescimento no número de agrupamentos até por volta do peso de 65 a 75. O número de comunidades então começa a reduzir até a remoção de todas as arestas. A rede utilizada foi referente a família das HIUases/Transtirretinas.

Os resultados obtidos foram primeiramente inspecionados manualmente, através de revisões bibliográficas, correlações com bancos de dados e análise estrutural. Posteriormente foi realizado uma validação baseada em classificação de sequência. A hipótese de que este método seja capaz de detectar grupos de resíduos que realmente determinam a especificidade de uma subfamília pode ser colocada à prova usando um classificador. Ou seja, caso uma máquina seja capaz de corretamente identificar as sequências usando apenas os grupos de resíduos detectados pela metodologia proposta como fonte de informação, teremos um forte indicio de eficácia.

Para realizar a validação por classificação, cada sequência do alinhamento é representada por um vetor: v, de tamanho N, sendo este o número de comunidades

detectadas nas redes. Cada posição v_i é composta pela média dos resíduos da comunidade i presentes na sequência. Estes dados são utilizados como características para alimentar uma máquina de suporte de vetores (SVM, do inglês *Suport Vector Machine*). A máquina é treinada apenas com sequências extraídas do *Swiss-Prot*. Por se tratar de um conjunto com um número geralmente limitado de sequências e não normalizado, em relação ao conjunto de todas as sequências do alinhamento, a abordagem escolhida para validação foi a LOOCV (*Leave-one-out Cross Validation*). Esta abordagem consiste em a cada passo, separar uma única sequência para a etapa de validação, enquanto todas as outras são utilizadas no treinamento. Este processo foi repetido até que todas as sequências do conjunto de treinamento tenham sido utilizadas na validação.

Um algoritmo *Random Forest* foi utilizado com o objetivo de quantificar a importância de cada estimador. O algoritmo *Random Forests* consiste em avaliar múltiplas árvores de decisões. O funcionamento de uma árvore de decisão é baseado em definir múltiplas divisões no conjunto de dados que, matematicamente, melhor se associam as classes. Logo, este tipo de algoritmo é comumente utilizado para seleção de características relevantes. Neste trabalho, estes dados de associação gerado por *Random Forest* foram utilizados com o intuito de realizar uma seleção frontal de estimadores. Portanto, a cada passo, o estimador com a maior pontuação de importância é adicionado ao modelo, este modelo é então validado, e ao final, o conjunto de características que obteve a maior pontuação F1 é selecionado.

3.6 Ferramentas

Neste trabalho foram desenvolvidos duas ferramentas que fazem a aplicação da metodologia proposta: o CONAN (Co-variation Network Analyzer) [Fonseca et al., 2020], um servidor web para executar e interpretar os resultados de análises de coevolução; e o CEvADA (Co-Evolution Analysis Data Archive) que se trata de um banco de dados de grupos determinantes de especificidades preditos pela metodologia proposta.

3.6.1 CONAN

O CONAN foi desenvolvido utilizando uma arquitetura "produtor-consumidor" conforme esquematizado na figura 3.9. Esta arquitetura permite que as requisições dos usuários sejam processadas de forma assíncrona, gerenciada por uma fila de tarefas, uma vez que cada processo pode consumir uma quantidade significativa de recursos computacionais.

Ao submeter uma tarefa no *website* do CONAN (*software* produtor), uma mensagem em XML será enviada para o servidor ActiveMQ (http://activemq.apache.org/) contendo a entrada e um código identificador único, simultaneamente, o código será inserido em uma tabela do banco de dados informando que a tarefa está enfileirada. Ao receber a prioridade, o ActiveMQ envia a tarefa ao *software* consumidor que é responsável por realizar todos os cálculos necessários. O consumidor atualiza no banco de dados cada etapa concluída da execução, permitindo que o usuário tenha acesso em tempo real do estado atual do *job*. Além disto, o consumidor também salva todos os dados em um diretório de acesso compartilhado com o produtor, permitindo que estes dados sejam utilizados para a construção da páginas de resultados.

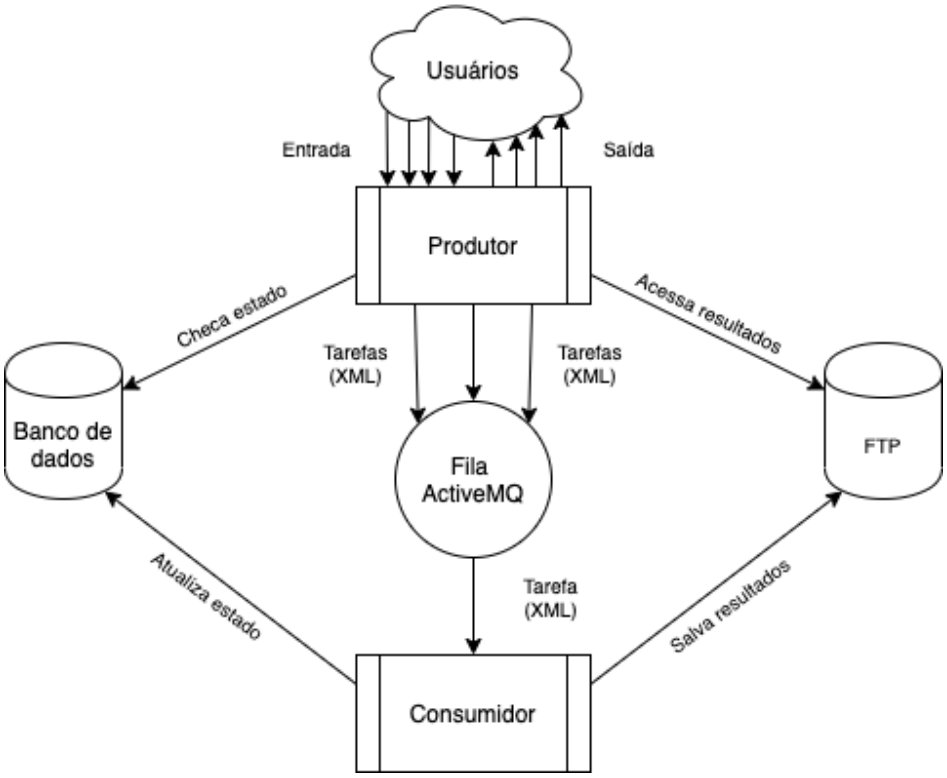


Figura 3.9: Fluxograma representando a arquitetura cliente-servidor utilizada no CONAN.

O *software* produtor foi desenvolvido em Java utilizando a especificação de interfaces *Java Server Faces*, os *frameworks* Spring (https://spring.io/) e PrimeFaces (https://www.primefaces.org/) para arquitetura de telas e gerenciamento das classes e objetos, o *framework* Hibernate (http://hibernate.org/) foi utilizado para gerenciar toda a comunicação com o banco de dados, a biblioteca JAXB [Ort & Mehta, 2003] para conversão de objetos java para XML, a biblioteca BioJava [Holland et al., 2008] para facilitar o cálculo, modelagem e acesso a diversos dados biológicos, além das bi-

bliotecas de visualização de dados: visNetwork [Almende et al., 2016], MSAViewer [Yachdav et al., 2016] 3DMol [Rego & Koes, 2014], ProtVista [Watkins et al., 2017] e d3 [Bostock et al., 2011].

O *software* consumidor foi desenvolvido na linguagem *Python* e com auxílio das bibliotecas: Python-Levenshtein para calcular distância de Levenshtein, SciPy [Jones et al., 2014] para o cálculo de diversas métricas, Pandas [McKinney, 2011] para armazenamento e manipulação de dados e NetworkX [Hagberg et al., 2005] para modelagem e manipulação de redes. O consumidor implementa o *pipeline* descrito na figura 3.1 utilizando a validação estatística pelo teste de Tumminello et al. [2011] e a detecção de comunidades por coeficiente de Jaccard. O *software* também contém uma etapa adicional no pré-processamento no qual para cada sequência do alinhamento, é acessado e mapeado uma série de anotações posicionais através da API do UniprotKb.

3.6.2 CEvADA

Os dados que compõe o *Specificity Determinant Data Base* foram gerados com o CONAN e utilizando alinhamentos do Pfam 32.0. Os parâmetros utilizados foram: 80% de identidade máxima, 5% de frequência mínima, 95% de frequência máxima e p-value mínimo de 10^{-15} utilizando teste de Tumminello et al. [2011]. Nesta primeira fase não foi utilizado ampliação de alfabeto, devido ao alto custo computacional. Foram consideradas todas as famílias do Pfam que obtiveram entre 500 e 20.000 sequências após o pré-processamento. O banco armazena cortes das redes obtidas a partir de 0.6 de coeficiente médio de coocorrência (Jaccard) e em intervalos de 0.05.

O banco de dados foi construído a partir da arquitetura do Pfam 32.0, acrescentando apenas as novas tabelas respectivas aos dados de especificidade (figura 3.10).

A aplicação *web* foi desenvolvida na linguagem Python utilizando o *framework* Django (https://www.djangoproject.com/) e as bibliotecas de visualização de dados: visNetwork [Almende et al., 2016], MSAViewer [Yachdav et al., 2016], ProtVista [Watkins et al., 2017] e d3 [Bostock et al., 2011].

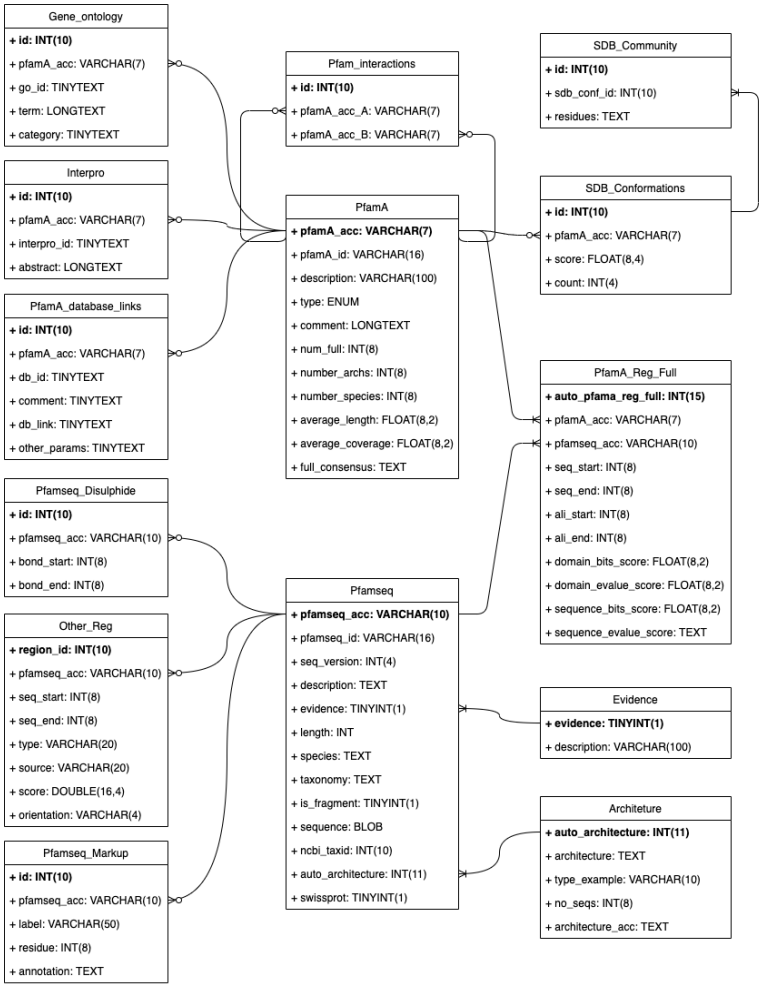


Figura 3.10: Diagrama de entidade relacionamento do CEvADA. O banco utiliza uma versão simplificada do Pfam 32.

Capítulo 4

Discussão e resultados

Foram realizadas dois tipos de análises com o objetivo de validar a metodologia proposta. A primeira foi conduzida com um conjunto de dados artificiais (descrito em métodos) com o intuito adicional de selecionar o melhor algoritmo para selecionar as arestas da rede monopartida. Na segunda etapa, foram efetuados estudos de caracterização e classificação com as famílias das Lisozimas de tipo C/Alfa-lactoalbuminas (Pfam: PF00062), Amidases (Pfam: PF01425), Transtirretinas/HIUases (Pfam: PF00576) e com a classe A dos receptores acoplados a proteína G (Pfam: PF00001). Estas análises tiveram como objetivo principal avaliar a eficácia da metodologia proposta quando aplicada a dados reais. As quatro famílias foram selecionadas por possuírem características distintas em relação à quantidade e distribuição de subfamílias e especificidade funcional.

4.1 Conjunto de Dados Simulados

Foram gerados 100 alinhamentos simulados, contendo padrões artificiais de conservação entre aminoácidos (conforme descrito na metodologia) incluindo também padrões de conservação marginal (propriedades fisioquímicas dos aminoácidos). O número de sequências nestes AMS variou entre 1000 e 2500 e o número de colunas entre 50 a 100. Para cada alinhamento foi gerado a rede bipartida correspondente e posteriormente sua projeção monopartida. Estas redes foram utilizadas para avaliar a eficácia de diversas abordagens para normalização e esparsificação de redes. Para cada método, foi observada a sua eficiência em reter na rede apenas arestas esperadas, aquelas que representam conexões entre resíduos determinantes de especificidades artificialmente

inseridos no alinhamento.

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.1}$$

Em um primeiro momento, foi aplicado um valor de corte nas arestas de cada rede com o objetivo de maximizar o *F1-score* (média harmônica da precisão e revocação, equação 4.1) dos nós retidos, considerando cada comunidade como um componente conexo.

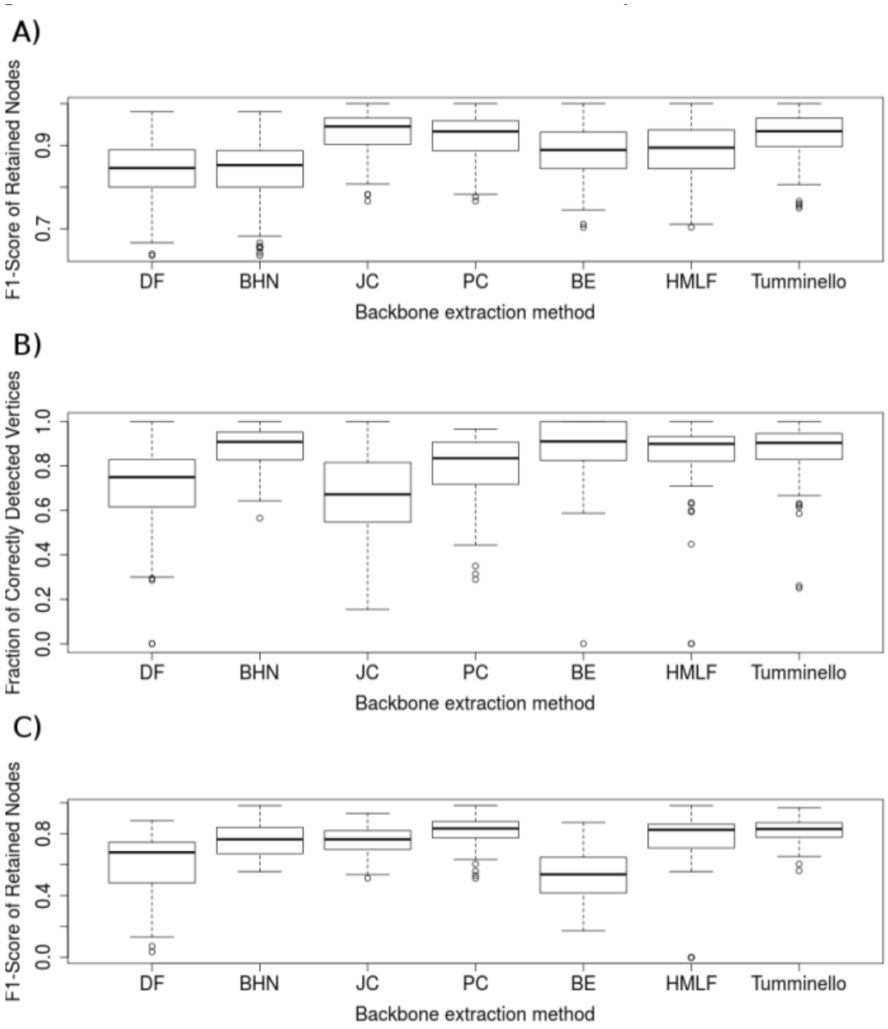


Figura 4.1: Eficácia de cada método de extração de *backbone* aplicado à detecção de covariação de resíduos. Em A) o *F1-score* dos nós corretamente retidos na rede quando o corte ótimo é conhecido. Em B) e C) respectivamente, é mostrado a fracão dos nós corretamente detectados (FCDV) e o *F1-score* dos nós corretamente mantidos na rede após a execução do *pipeline* completo [da Fonseca Jr et al., 2019].

Como pode ser visto na figura 4.1A, ao considerar o corte ótimo, todos os métodos

analisados podem ser considerados aptos para extrair o *backbone* das redes artificiais geradas neste trabalho, obtendo valores de *F1-score* superiores a 0.8. Este resultado pode ser interpretado como um primeiro indicativo de que a modelagem de rede proposta neste trabalho pode realmente ser utilizada para analisar covariação de resíduos. Porém, em um experimento real, o corte ótimo seria desconhecido. Logo, um bom extrator precisa ser capaz de ponderar grupos coocorrentes de forma balanceada e penalizar coocorrências ao acaso de forma a maximizar o intervalo de valores cortes que se possa resultar na *backbone* ideal. Sendo assim, mesmo a aplicação de valores de cortes mais distantes do ótimo resultaria em resultados próximos do esperado. Para levar em conta estes fatores, foram também realizadas análises com as mesmas redes aleatórias, porém incluindo também a seleção automatizadas do corte e a detecção de comunidades. Os resultados podem ser observados na figura 4.1. Um ponto interessante a se destacar é que a equação de Bonacich obteve o melhor resultado levando em consideração o FCDV, porém o pior *F1-score*, o que levanta indícios de que o método realizou uma excelente tarefa na remoção de ruídos, porém falhou na normalização, uma vez que os sinais de determinantes de subfamílias de baixa frequência foram perdidos. De forma inversa, porém semelhante, os coeficientes de Jaccard e de correlação de Pearson obtiveram ótimos *F1-score*, mas baixos valores de FCDV. O filtro de disparidade não obteve bons resultados em nenhum dos testes, e os dois métodos que demonstraram resultados mais consistentes foram os filtros de Tumminello et al. e o de probabilidade marginal *Hairball* [Dianati, 2016].

4.2 Conjunto de Dados Reais

A validação da metodologia proposta utilizando dados reais foi divididas em três etapas: busca por correlação entre comunidades detectadas e grupos funcionais já conhecidos, validação baseada em classificação e correlação entre resíduos detectados e proteínas ainda não caracterizadas. Cada uma destas três etapas será detalhada com mais rigor no decorrer deste capítulo, bem como uma breve descrição de cada uma das quatro famílias utilizadas.

4.2.1 Lisozimas e Alfa-lactoalbuminas

A família das lisozimas de tipo C e das alfa-lactoalbuminas, também chamadas de família dos glicolídeos hidrolase 22, são consideradas bons exemplos para *benchmarking* de métodos de detecção de sítios determinantes de especificidade, uma vez que um processo de duplicação genica seguida de divergência resultou em ao menos

duas subclasses com atividades completamente distintas [Nitta & Sugai, 1989; Davies & Henrissat, 1995]. As Lisozimas de tipo C (LYSC) se tratam de enzimas com atividades bacteriolíticas, através da hidrólise das ligações glicosídicas do peptidoglicano β-1,4 (EC: 3.2.1.17) e são distribuídas por todo reino Metazoa [Jollès & Jollès, 1984; Zhang et al., 2005]. Já as alfa-lactoalbuminas (LALBA) são proteínas de mamíferos, especificamente expressas no leite, e atuam como reguladoras ao se associarem ao β-1,4-galactosil-transferase, formando um heterodimêro funcional chamado lactose-sintetase, essencial para produção de leite [Hall & Campbell, 1986].

As alfa-lactoalbuminas não possuem o sítio ativo das lisozimas, porém todas possuem a capacidade de se ligar a íons de cálcio [Stuart et al., 1986], caracteristica que é restrita a apenas algumas poucas lisozimas, como a equina [Nitta et al., 1987]. Porém apesar da diferença funcional, ambas as subfamílias compartilham uma alta similaridade em nível de sequência e de estrutura, possuindo cerca de 35% a 40% de resíduos conservados, além de quatro pontes disulfeto [Nitta & Sugai, 1989].

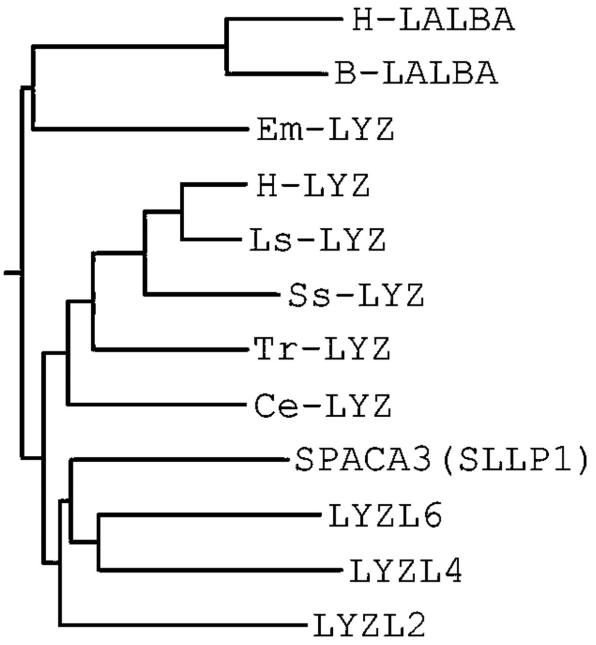


Figura 4.2: Reconstrução filogenética da família dos glicolídeos hidrolase 22] [Zhang et al., 2005].

Apesar das lisozimas de tipo C serem conhecidas desde 1922 [Fleming, 1922], a família dos glicolídeos hidrolase 22 ainda incluem várias subfamílias ainda muito pouco caracterizadas, geralmente chamadas de LLP (*Lysozyme-like proteins*). Alguma destas envolvidas na reprodução, como no caso das proteínas associadas à membrana do acrossomo do espermatozoide (SPACA3), provavelmente envolvida na adesão e fusão da membrana plasmática do óvulo pelo espermatozoide durante a fertilização [Mandal

et al., 2003; Herrero et al., 2005]. A figura 4.2 mostra uma reconstrução filogenética com as proteínas da família [Zhang et al., 2005].

4.2.2 Amidases

Amidases constituem de um grande grupo de enzimas encontradas na maioria dos organismos que possuem a atividade de hidrolisar ligações amidas (-CO-NH-) [Ko et al., 2010]. A família das amidases pode ser subdivididas em subfamílias de acordo com sua função molecular e afinidade por ligante, alguns exemplos incluem: amidases de peptídeo [Neumann & Kula, 2002], amidases de ácido graxo [McKinney & Cravatt, 2005], malonamidases [Shin et al., 2002] e a subunidade A da Glu-tRNA amidotransferases (GATA) [Kwak et al., 2002; Ko et al., 2010]. Apesar de possuírem aproximadamente 160 resíduos conservados, incluindo a tríade catalítica, Ser-Ser-Lys, os membros desta família costumam se divergir de acordo com a especificidade pelo ligante [Valiña et al., 2004].

4.2.3 Transtirretinas e HIUases

A família das Transtirretinas/HIUases é composta por um número relativamente pequeno de sequências e pode ser dividida em basicamente duas subfamílias: Hidrolase 5-hidroxi iso-hidratada (HIUase), enzima presente desde bactérias a vertebrados, envolvida no metabolismo do ácido úrico, catalisando a hidrólise do 5-hidroxiisourato na via de degradação do urato [Richardson, 2015; Cendron et al., 2011]; e a a transtirretina, uma proteína responsável pelo transporte dos hormônios tireoidianos T3 e T4 provavelmente originada durante a emergência dos vertebrados após uma duplicação no gene codificante da HIUase, como pode ser observada na árvore filogenética da figura 4.4b [Richardson, 2015; Cendron et al., 2011]. Ambas as subclasses possuem uma alta similaridade em nível de sequência e estrutura, além disto, é sabido que algumas poucas substituições na região do sítio ativo são suficientes para permitir que uma HIUase seja capaz de ligar aos hormônios tireoidianos [Zanotti et al., 2006; Romero & Arnold, 2009; Cendron et al., 2011]. A relação evolutiva entre estas duas proteínas e o provável efeito de neofuncionalização após uma duplicação gênica seguida de mutações específicas fazem com que esta família seja um perfeito caso de estudo para uma metodologia de predição de sítios determinantes de especificidade.

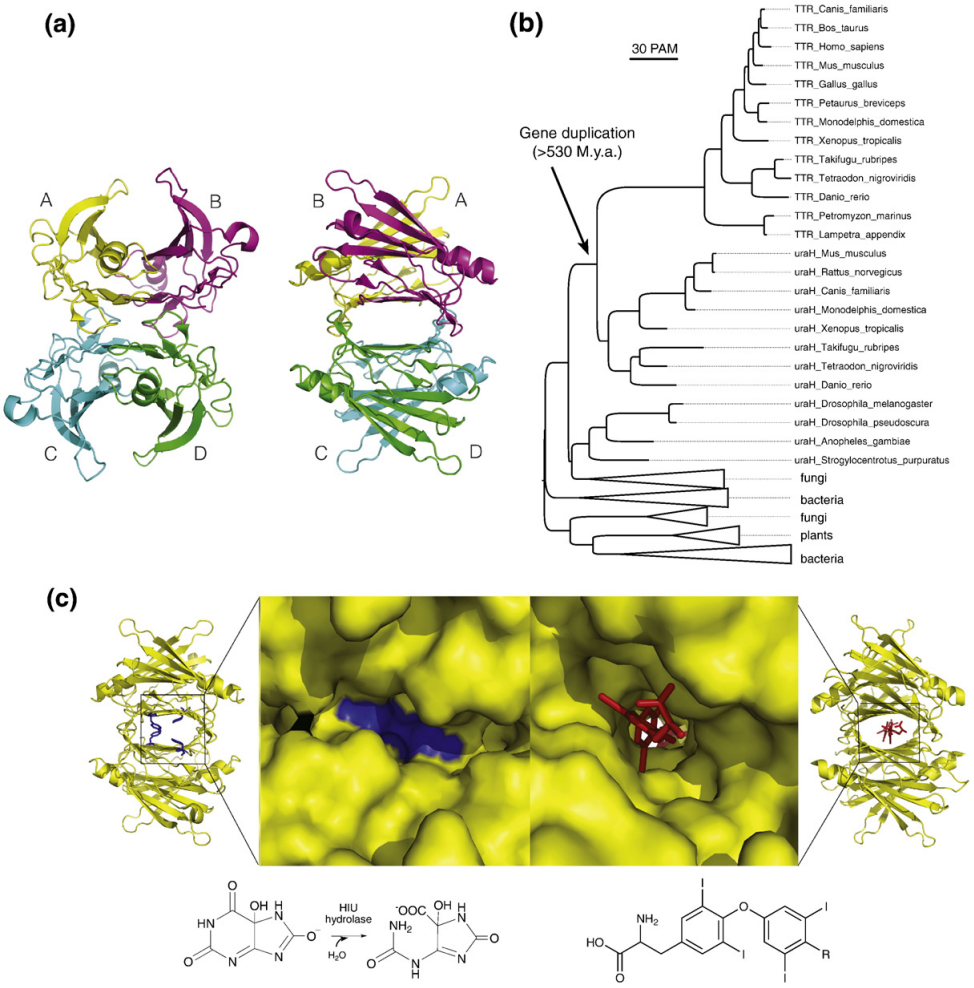


Figura 4.3: Comparação estrutural e filogenética da HIUase e Transtirretina. Em (a) a estrutura da HIUase de *Zebrafish*; em (b) árvore filogenética da família enfatizando a provável origem da Transtirretina a partir da HIUase; em (c) comparação dos sítios ativos da HIUase de *Zebrafish* e da Transtirretina humana. A imagem enfatiza a obstrução do sítio causada pela presença da Tyr116 na HIUase, resíduo não presente na transtirretina [Cendron et al., 2011].

4.2.4 Receptores acoplados à proteína G

Os receptores acoplados à proteína G (GPCRs) consistem de componentes chaves na comunicação celular, geralmente associados à detecção de sinais exógenos, como odores, sabores, luz e feromônios; ou na regulação de processos fisiológicos nos sistemas nervosos e endócrinos [Mombaerts, 2004; Munk et al., 2016]. A família também consiste de um dos principais alvos para desenvolvimento de fármacos, representando cerca de 19% dos alvos para as drogas disponíveis no mercado [Rask-Andersen et al., 2014; Munk et al., 2016]. Os GPCRs são codificadas pela maior família do genoma humano e podem ser divididas em seis principais classes (Figura 2.6) de acordo com similaridade entre

as sequências, tipos de ligantes e árvores filogenéticas [Kolakowski, 1994; Fredriksson et al., 2003; Munk et al., 2016; Møller et al., 2017].

Os GPCRs de classe A (Pfam: PF00001) constituem a maior classe da família, contendo 689 membros em humanos e atualmente 1.827 sequências depositadas no Swiss-Prot. Estas podem ser novamente subdivididas em mais nove subfamílias (aminérgica, peptídeo, proteína, lipídio, melatonina, nucleotídeo, esteroide, ácido carboxílico alifático e sensoriais) de acordo com seu tipo de função e de ligante (Figura 2.6). Em um nível mais baixo, a família ainda possui centenas de classes de proteínas com funções e ligantes específicas. Outro fator que torna esta família de proteínas um caso interessante para o estudo de determinantes de especificidade é a presença de proteínas órfãs, proteínas cuja função ou ligante ainda são desconhecidos. Muitas destas sequências possuem baixa similaridade global com outras GPCRs já conhecidas, tornando-as ótimos casos para classificação baseada em SDSs [Song et al., 2017].

4.2.5 Correlacao entre comunidades detectadas e grupos funcionais

Ao plotar a frequência média dos resíduos detectados em cada comunidade em alinhamentos formados por subgrupos funcionais de cada uma das quatro famílias analisadas, é possível observar o indício de alguns sítios determinantes de especificidades. Isto se deve pelo fato de serem altamente conservados dentro de uma subfamília e quase nulo nas outras.

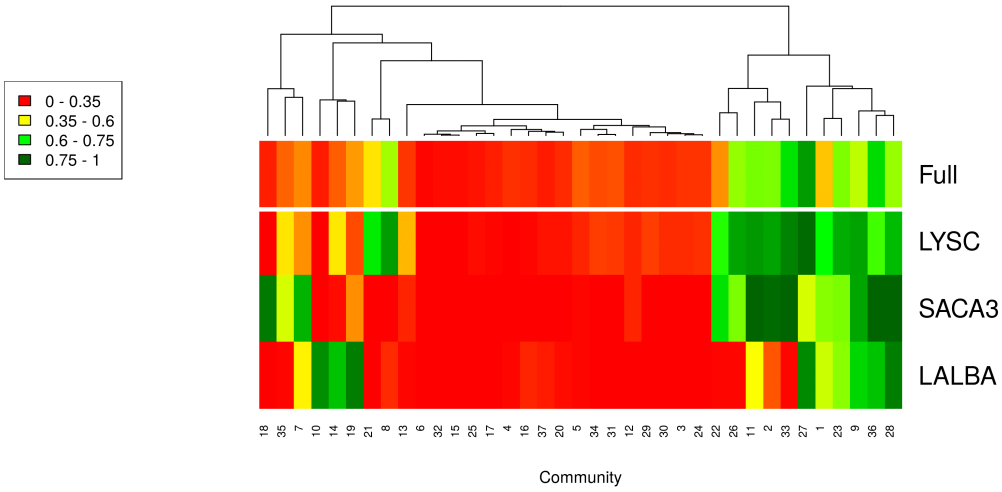


Figura 4.4: Frequência média dos resíduos detectados nos sublinhamentos compostos por sequências de LYSC, LALBA e SACA3 obtidos no Swiss-Prot.

No caso da família dos glicolídeos hidrolase 22 (LYSC/LALBA) este tipo de pa-

drão foi detectado para as três principais classes: LYSC, LALBA e SACA3. Comunidades como a 8 e 21 são altamente conservadas em Lisozimas de tipo C e praticamente não ocorrem em LALBAS e SACA3 e de forma semelhante, as comunidades 10 e 19 parecem ser especifica das alfa-lactoalbuminas. Esta hipótese fica ainda mais em evidência ao analisar os resíduos que compõe estas comunidades. A comunidade 8 inclui o par de resíduos catalíticos das lisozimas: Glu53 e Asp71 (numeração referente à Lisozima C humana), o que era esperado de se encontrar, uma vez que as LALBAs não possuem atividade bacteriolítica. Os outros 6 resíduos presentes nestas duas comunidade se localizam na região próxima da cavidade catalítica. Padrões interessantes também podem ser observados nas comunidades especificas das LALBAs. Dois resíduos da comunidade 10 possuem a mesma posição de resíduos detectados como SDPs de lisozimas: Glu44 (nas lisozimas ocorre um Asn) e Glu68 (numeração referente a alfa-llactoalbumina humana), posição referente ao Asp71 do sítio ativo das lisozimas. Já na comunidade 19, foi detectado uma tríade de aspartatos que compoe o sítio de ligacao a calcio nas LALBAs, característica chave para sua atividade. Outro ponto interessante pode ser observado na tabela A.2, alguns resíduos da comunidade 19 ocorrem em algumas lisozimas, porém a única a possuir a tríade completa é a lisozima de cavalo, justamente uma das poucas conhecidas por ter a capacidade de se ligar a íons de cálcio [Nitta et al., 1987]. Além disto, a comunidade 18 também parece ser determinante de especificidade para as proteínas associadas à membrana do acrossomo do espermatozoide.

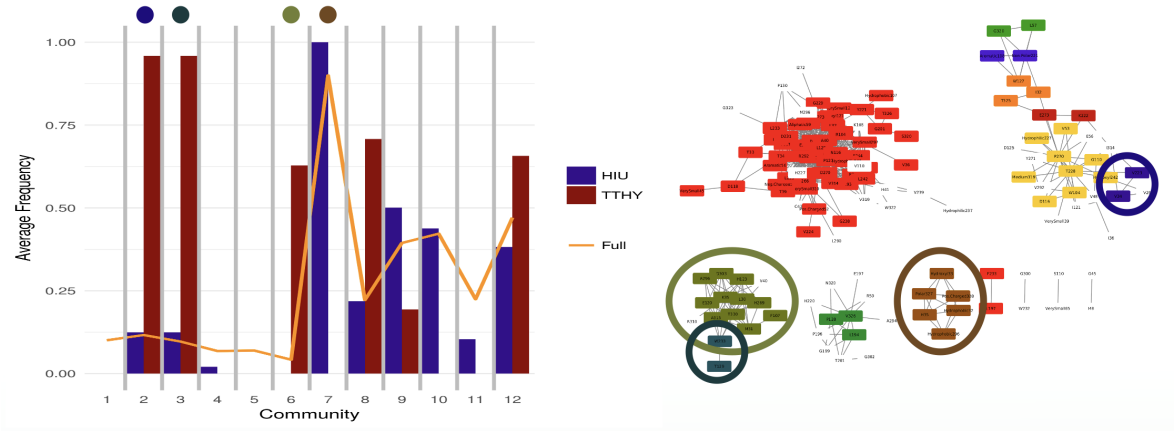


Figura 4.5: Frequência média dos resíduos de cada comunidade em cada subalinhamento composto por sequências das HIUases e Transtirretinas no Swiss-Prot e a respectiva rede de co-variação com as comunidades identificadas por coloração.

A rede de co-variação gerada para a família das HIUases e transtirretinas também separou bem os resíduos. Pode-se observar pela figura 4.5 que a maioria das comunida-

des possuem uma tendência consideravelmente maior para uma classe do que a outra. O principal indicativo de SDS obtido nesta análise consiste da comunidade 7, composta por 6 resíduos completamente conservados em HIUases e ausente nas transtirretinas (tabela A.4). Outra três comunidades podem ser consideradas possiveis determinantes de transtirretinas: 2, 3 e 6 (tabela A.5.

Ao analisar a distribuição espacial destas comunidades de resíduos detectados (figura 4.6) nas estruturas da HIUase e da transtirretina, é possível observar que em ambos os casos, os resíduos tendem a estarem localizados na região dos sítios de atividade das mesmas. Principalmente em regiões que sofreram alterações na estrutura secundária, onde na HIUase se observa presença de alças, enquanto na transtirretinas observa-se a presença de folhas beta, fortalecendo a hipótese de que estes resíduos sejam realmente determinantes da divergência funcional entre ambas as classes. Outro ponto interessante é a Tyr115 (numeracao referente a HIUase de camundongo), que aparece na comunidade 7 como resíduo polar. Esta tirosina é considerada um dos principais motivos para a divergência funcional destas duas proteínas, uma vez que a perda deste aminoácido pelas transtirretinas causou uma expansão da cavidade do sítio ativo formando uma estrutura de túnel, o que provavelmente facilitou a capacidade de interação com o hormônio tireoidiano [Cendron et al., 2011; Lee et al., 2005].

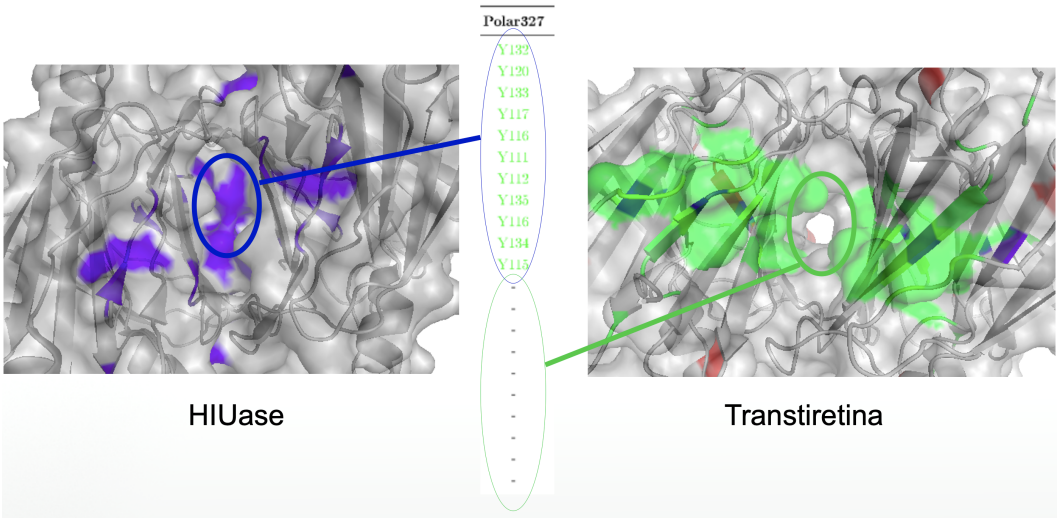


Figura 4.6: Comparacao da cavidade do sítio ativo da HIUase de peixe-zebra (PDB: 2h1x) e da transtirretina de rato (PDB: 1gke). Os resíduos da comunidade 7 estão marcados na cor azul na estrutura da HIUase e os resíduos da comunidade 6 estão marcados na cor verde na estrutura da transtirretina.

O estudo com as amidases não foi muito diferente, considerando as amidases que possuíam amostragem no Swiss-prot, foi identificado possíveis SDS para três classes: amidases de ureia, acetamidase e a GATA. Uma vez que a tríade catalítica das ami-

dases é conservada em toda a família, é esperado que os resíduos que determinem a especificidade para cada tipo de ligante estejam no entorno da tríade, na região da cavidade catalítica.

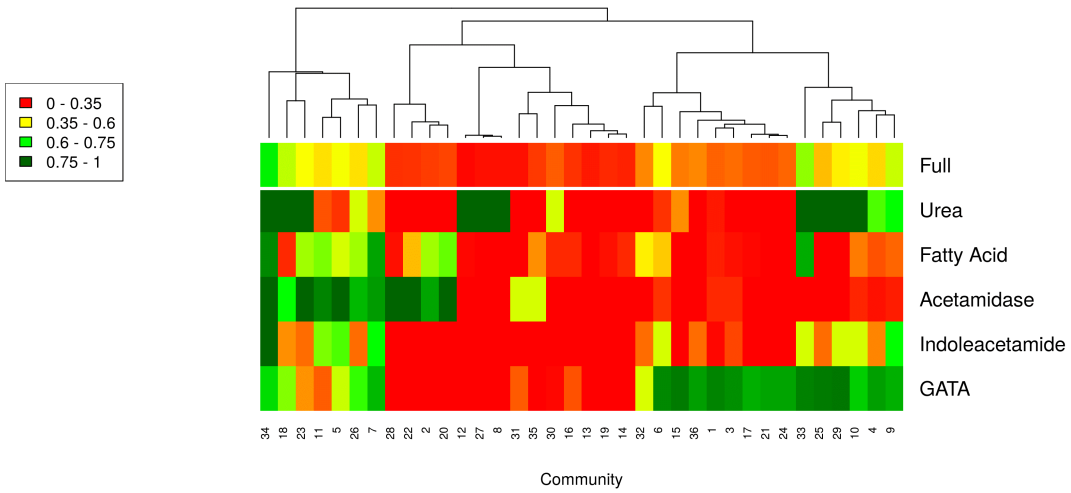


Figura 4.7: Frequência média dos resíduos de cada comunidade em cada subalinhamento composto por sequências das amidases.

Como é possível observar na figura 4.7, 7 comunidades demonstraram uma possível relação de especificidade para GATA: comunidades 1, 3, 15, 17, 21, 24 e 36 (tabela A.6 e figura 4.8). Ao mapear estas comunidades na estrutura de GATA de levedura (PDB: 4n0h), foi possível observar que os resíduos detectados nas comunidades 1 e 3 estão concentrados na entrada da cavidade catalítica, um padrão que era esperado. As comunidades 22 e 28 (tabela A.7) são aparentemente conservadas em acetamidases, já as comunidades 2 e 20 (tabela A.8), além de ocorrerem em acetamidases também ocorrem em outra classe de amidases de ácido carboxílicos, as amidases de ácido graxo. Atualmente não há nenhuma estrutura de acetamidases depositadas no PDBe, porém ao mapear os resíduos das comunidades 2 e 20 na estrutura da amidase de ácido graxo de rato (PDB: 3ppm), é possível observar que os padrões não estão tão bem definidos quanto no caso da GATA. Apesar de alguns resíduos estarem interagindo com o ligante ou próximo da interface de dimerização, há também resíduos mais afastados do sítio em hélices e alças em contato com o solvente. As comunidades 8, 12 e 27 (tabela A.9) também foram observadas estritamente conservadas em amidases de uréia, porém não há estruturas disponíveis para localizar os resíduos.

Nos estudos de caso realizados com a família das GPCRs de classe A, a principio foi procurado observar a efetividade da metodologia em detectar resíduos determinantes de especificidade em relação aos tipos gerais de ligantes: aminérgicos, melatonina, nucleotídeo, ácido carboxílico, lipídeo, sensorial, peptídeo, esteroide e proteína. Nesse

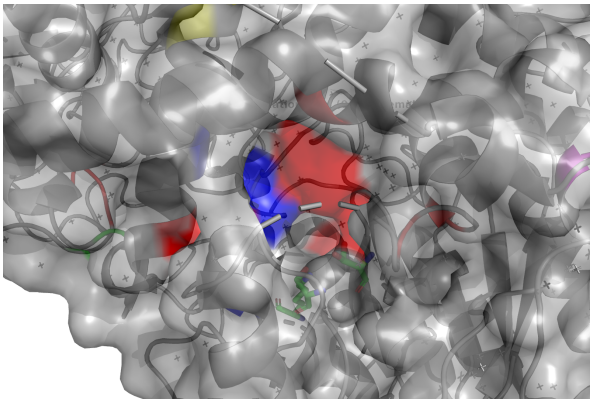


Figura 4.8: Resíduos das comunidades 1 e 3 mapeados na entrada do sítio ativo da GATA de levedura.

sentido, é possível observar pela figura 4.9 dois possíveis grupos: a comunidade 4, exclusivamente conservada em GPCR's aminérgicas e a comunidade 9, também exclusivamente conservada em GPCR's sensoriais. Os resíduos da comunidade 4 (tabela A.10) estão situados no sítio de ligação destas proteínas e muitos destes possuem anotações no UniprotKb a respeito de interação com ligantes. Em relação a comunidade (tabela A.11), mutações em resíduos já foram descritas por estarem relacionadas a doenças hereditárias oculares, como a monocromacia de cone azul e a retinose pigmentar autossômica dominante [Sung et al., 1991; Nathans et al., 1993; Yang et al., 1997; Gardner et al., 2010]. Além disto, a mutação Lys312Glu (numeracao referente à opsina sensível a ondas longas humana) desnatura o sítio de ligação com o cromóforo, impedindo que a proteína seja ativada pela luz [Li et al., 1995].

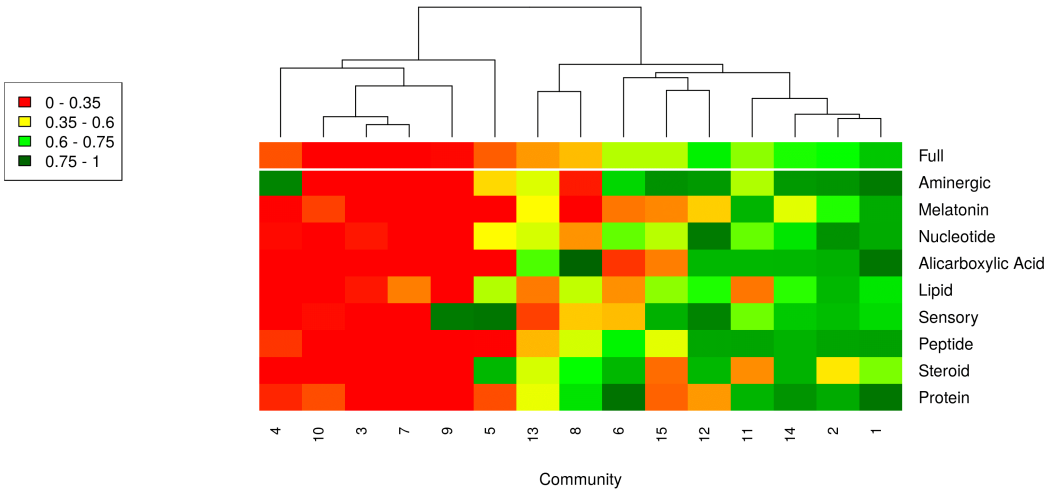


Figura 4.9: Frequência média dos resíduos de cada comunidade em cada subalinhamento composto por sequências das GPCRs de classe A.

Outras comunidades não apresentaram uma ocorrência exclusiva para uma de-

terminado tipo de ligante, porém possuem uma conservação local acima da média do alinhamento, como no caso da comunidade 12 para GPCR's que ligam a nucleotídeos, comunidade 13 para a classe dos ácidos carboxílicos alifáticos, comunidade 11 para peptídeos, comunidades 5 para esteroides.

Foi avaliada também a capacidade de identificar sítios determinantes de especificidade para as subfamílias funcionais das GPCRs, isto é, referente a cada ligante especificamente. Esta é uma tarefa não trivial, uma vez que a família possui centenas de subclasses funcionais, muitas delas com identidade global extremamente alta, outro fator dificultante está na validação, uma vez que o número de sequências de cada subclasse depositadas no Swiss-Prot é geralmente muito baixo. Apesar disto, como é possível observar na figura 4.10, o método foi capaz de detectar alguns grupos determinantes para as classes prostanoide, opsinas e hormônios glicoproteicos.

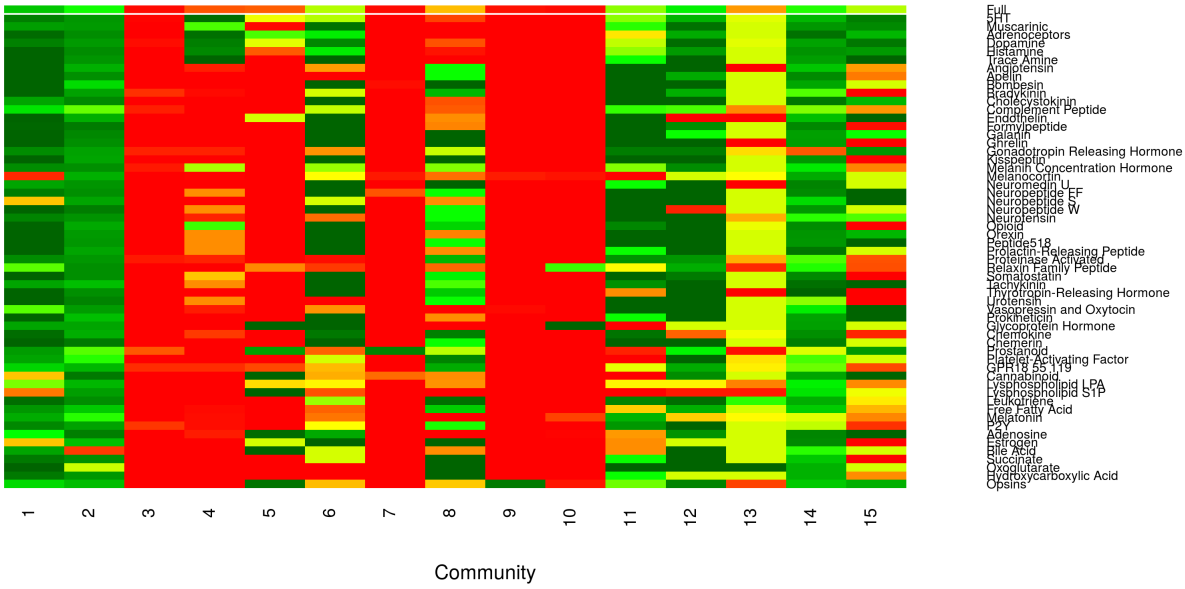


Figura 4.10: Frequência média dos resíduos de cada comunidade da rede 2 em cada subalinhamento composto pelas classes funcionais das GPCRs.

A comunidade 7, tabela A.12, é exclusivamente conservada em receptores de Prostanoide. A importância desta arginina na interação com o ligante já foi demonstrada por diversos estudos de mutagêneses envolvendo receptores de prostaglandina [Huang & Tai, 1995; Negishi et al., 1995; Chang et al., 1997; Kedzie et al., 1998]. A comunidade 9, já foi comentada anteriormente por se tratar de um determinante de especificidade para GPCRs sensoriais. Porém, é possível observar por esta figura, que este par de resíduos é literalmente exclusivo de receptores de opsinas. A comunidade 10 (tabela A.13), apesar de possuir uma frequência razoável nos receptores de relaxina, pode ser considerada uma determinante de especificidade para os receptores de hormônios glicoproteicos. A

comunidade é composta pelos resíduos Tyr633, Glu883, Asp2832 e Pro3059 e é completamente conservada nestes receptores. Mutações no Asp633 já foram associadas a doenças nas três subclasses de receptores de hormônios glicoproteicos: Hipogonadismo masculino nos receptores de hormônio folículo-estimulante (FSHR) [Huhtaniemi, 2017]; puberdade precoce masculina, testotoxicose e hipogonadismo masculino nos receptores de lutropina-coriogonadotrópico (LSHR) [Kosugi et al., 1996; Siviero-Miachon et al., 2017; Juel Mortensen et al., 2017; Huhtaniemi, 2017]; e bócio multinodular tóxico nos receptores de tirotropina (TSHR) [Tonacchera et al., 2000].

4.2.6 Validação por classificação de sequências

Além de detectar possíveis grupos de resíduos determinantes de especificidade, a metodologia proposta é capaz de identificar grupos possivelmente proibitivos para certas classes, sendo assim possível criar uma espécie de assinatura para as sequências, conforme a coocorrência dos resíduos detectados na rede. Caso estas assinaturas contenham realmente informações acerca de resíduos determinantes e proibitivos para as subclasses, é esperado que eles sejam suficientemente informativos para alimentar um classificador de sequências. Portanto foi utilizado uma máquina de suporte de vetores (SVM) alimentado pela média de frequência dos resíduos de cada comunidade para cada sequência depositada no Swiss-Prot.

E como é possível observar na tabela 4.1, o classificador conseguiu separar corretamente as sequências de LYSC, LALBA e SACA3, bem como as transtirretinas das HIUases. Já no caso das amidases, devido a falta de sequências manualmente curadas, não foi possível incluir as amidases de ureia e de indoleacetamida nesta análise. Dentre as classes restantes, o modelo errou apenas uma sequência de amidase de ácido graxo.

Grupo	**Precisão**	**Revocação**	***F1 Score***	**Suporte**
LYSC/LALBA	1,00	1,00	1,00	75
HIUase/TTR	1,00	1,00	1,00	25
Amidase	0,98	0,99	0,98	357
GPCR/General	0,87	0,87	0,85	1070
GPCR/Ligands	0,76	0,81	0,77	1049
Média/Total	0,84	0,86	0,84	2576

Tabela 4.1: Validação da classificação para as quatro famílias analisadas.

A obtenção de uma alta taxa de acerto para a as famílias da lisozimas, transtirretinas e amidases já era esperada, uma vez que foi detectado SDS para todas as subfamílias analisadas nesta etapa. O principal desafio se trata da família das GPCR's,

afinal só foi possível detectar SDS para 2 classes das 9 gerais, e para 3 das dezenas quando considerado o ligante especifico. Como é possível observar na tabela 4.2, o método foi capaz de determinar com precisão as sequências de GPCRs com ligantes aminérgicos, esteróides, peptídeos, proteícos e sensoriais. Porém não demonstrou sinais de especificidade para as classes melatonina e ácidos carboxílicos alifáticos, mesmo havendo uma comunidade 100% conservada em GPCR's de ácidos carboxílicos. Outro fato interessante é que o classificador obteve uma alta taxa de acerto para a classe das esteroides, mesmo esta não demonstrando nenhum sinal de determinantes de especificidades conforme observado na figura 4.9. Isto pode ser atribuído pelo fato desta classe possuir uma baixa frequência para as comunidades 1 e 2 (rede 2), que além de serem bastante conservadas no alinhamento global, são também conservadas em todas as subclasses analisadas, gerando um padrão proibitivo. Estas comunidades incluem resíduos importantes para a atividade geral das GPCRs, incluindo o principal motivo da família, NPxxY.

Grupo	**Precisão**	**Revocação**	***F1 Score***	**Suporte**
Ácido carboxílico	0,00	0,00	0,00	12
Aminérgico	0,98	0,97	0,98	235
Esteroide	0,91	1,00	0,95	10
Lipídeo	0,98	0,70	0,82	137
Melatonina	0,00	0,00	0,00	11
Nucleotídeo	0,94	0,44	0,60	66
Peptídeo	0,75	0,97	0,84	66
Proteína	0,94	0,84	0,89	139
Sensorial	1,00	1,00	1,00	78
Média/Total	0,87	0,87	0,85	1070

Tabela 4.2: Acurácia da classificação das sequências de GPCR, utilizando como fonte de informação os vetores de frequência média de cada sequência para cada comunidade da rede 2.

Apesar da detecção de apenas 3 comunidades com possíveis características de SDS, quando considerando as classes de acordo com os ligantes, a informação gerada por essa rede foi suficiente para que um classificador conseguisse distinguir a grande maioria das subclasses funcionais. Como é possível observar na tabela anexo A.1, o classificador obteve um *F1 Score* médio de 0.77, e apenas 10 das 52 subclasses não mostraram nenhum sinal de especificidade (Aminas de Traço, Apelina, Chemerina, Dopamina, Galanina, Lisofosfolipídeo LPA, Neuropeptídeo FF, Neuropeptídeo W, Neurotensina e Urotensina).

Como o método foi capaz de classificar corretamente a maioria das classes neste cenário, foi buscado o entender o tanto que a informação de resíduo proibitivo foi importante nesta classificação, afinal haviam apenas três SDS. Para isto, foi utilizado um *random forest* para calcular o quanto que cada *feature* (neste caso comunidade) influenciou na decisão do classificador. Como pode ser visto na figura 4.11, apenas uma única comunidade foi completamente ignorada pelo classificador, a comunidade 3, afinal era uma comunidade praticamente ausente em todas classes analisadas. Mas o interessante, é que as comunidades contendo SDS (7, 9 e 10) foram as que menos influenciaram as decisões, afinal estas apenas influenciam os resultados de uma única subclasse. Isto mostra o tanto que a informação de resíduos proibitivos (anti-correlacionados) são informativos na definição de especificidade de sequências.

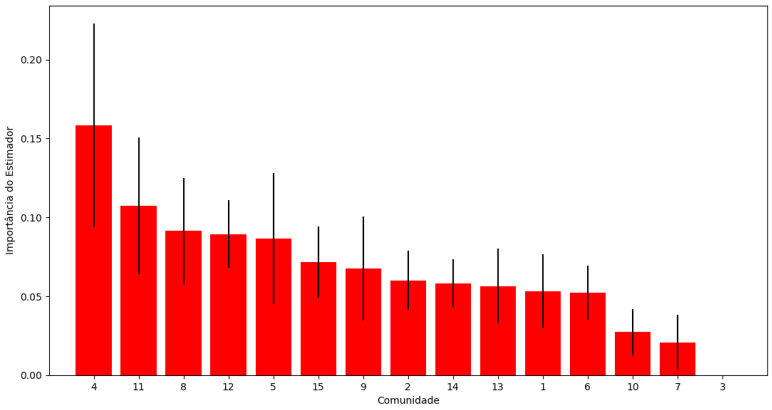


Figura 4.11: Eficácia de cada comunidade detectada na classificação de sequências. O cálculo foi realizado por um algoritmo *random forest*. As letras A e B se referem respectivamente as comunidades detectadas pelo método proposto e por um algoritmo de maximização da modularidade por *simulated annealing*.

4.2.7 Correlação entre grupos detectados e proteínas não caracterizadas

Uma das principais aplicações de métodos de detecção de SDS consiste em auxiliar processos de caracterização de proteínas. Tendo em vista que o método obteve bons resultados na classificação de sequências conhecidas e que muitas das comunidades detectadas não consistiam de SDS claros para grupos já conhecidos, foram realizados testes com proteínas órfãs ou ainda não caracterizadas.

No caso da família das lisozimas de tipo C e alfa-lactoalbuminas, foi separado 5 classes de proteínas similares a lisozimas extraídas do *Swiss-Prot*: *Lysozyme-like*

proteins 1, 2, 4, 5 e 6. Ao plotar as frequências médias de acordo com as comunidades, é possível constatar que a assinatura de todas elas se assemelham mais às LYSC e SACA3 do que as LALBA. Nenhuma das 5 classes analisadas possuem as comunidades 10, 14 e 19, determinantes de LALBA, além disto, a maioria destas classes possuem uma série de comunidades que são proibitivas em LALBA. Portanto, é improvável que estas proteínas tenham a capacidade de se interagir com íons de cálcio. Outro ponto interessante é que nenhuma das classes também possuem simultaneamente as duas comunidades detectadas como determinantes de lisozimas, nem a comunidade determinante de SACA3. Porém, a comunidade 8, que inclui o sítio ativo das lisozimas, é razoavelmente conservada nas LIZL5 e LIZL6 e de fato a atividade bacteriolítica das LIZL6 já foi demonstrada [Wei et al., 2013; Huang et al., 2017]. LIZL4 também possui uma característica única, que é a ausência de quatro comunidades conservadas nas lisozimas. Ao incluir estas 5 novas classes, mais duas possíveis SDS aparecem: comunidade 12 (composta por Glu64, Lys83, Non-Polar86 e Glu92,), conservada nas LIZL4 (e frequente em algumas LIZL6) e a comunidade 37 (composta por Glu58 e um negativamente carregado 144) conservada nas LIZL6.

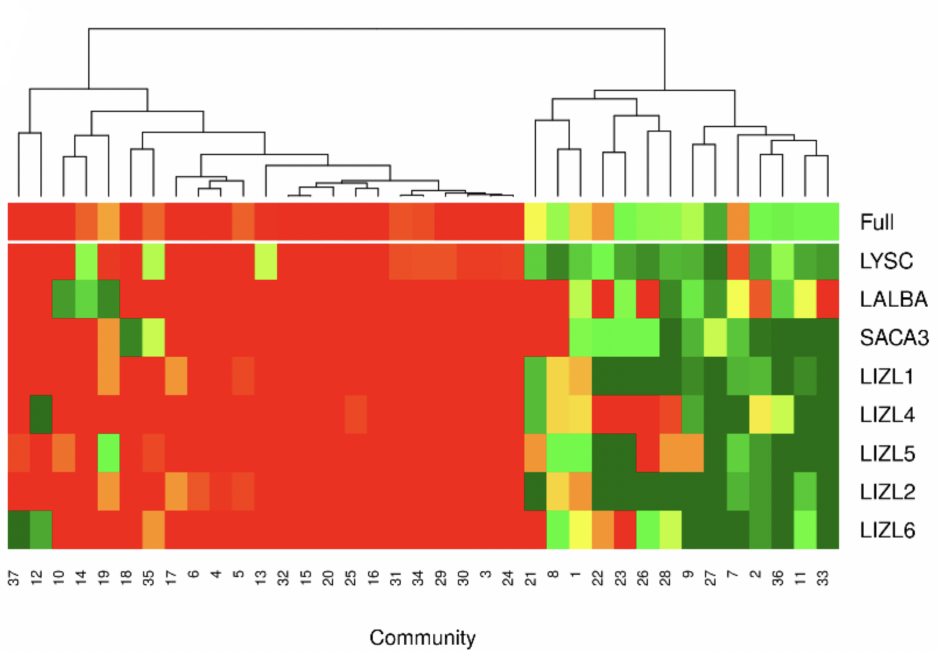


Figura 4.12: Frequência média dos resíduos detectados nos sublinhamentos compostos por sequências de subfamílias não caracterizadas da família dos glicolídeos hidrolase 22 obtidos no Swiss-Prot.

Uma característica interessante da família dos receptores acoplados à proteína G, é que ainda existem diversas sequências órfãs, proteínas cujo ligante ou até mesmo sua função são desconhecidos. Tendo isto em vista, foi realizado um teste de classificação

das 195 sequências anotadas como órfãs no GPCRdb [Munk et al., 2016]. É possível observar na figura 4.13, que os resultados da classificação em relação aos ligantes específicos obtiveram uma baixa confiabilidade. Porém tal resultado, além de esperado é interessante, pois, dado que estas proteínas possuem ligantes/função desconhecidas e distinta das anotadas, seus rótulos correspondentes não estariam disponíveis na etapa de treinamento. Já em relação as classes genéricas (aminérgicas, esteróides, lípidos, etc.), a classificação obteve resultados instigantes. Mais de 25% das sequências obtiveram probabilidade de estimação acima de 80%. Além disto, o método classificou corretamente todos os receptores de amina traço, receptores sabidamente aminérgicos, porém com ligantes ainda desconhecidos [Zucchi et al., 2006]. A tabela anexo A.2 lista o resultado da classificação de todas as sequências órfãs.

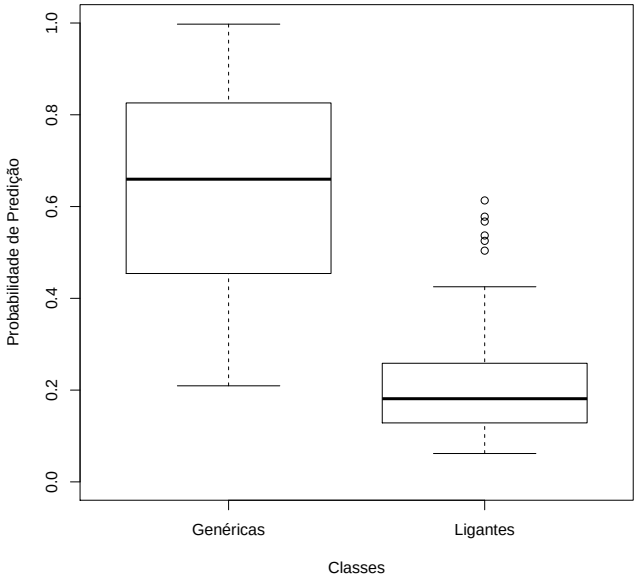


Figura 4.13: Distribuição das probabilidades de classificação entre classes genéricas e ligantes específicos das GPCRs órfãs

A GPR61 se trata de um caso interessante, pois ainda se sabe muito pouco a seu respeito. É sabido que é um receptor órfão, expresso abundantemente no cérebro [Toyooka et al., 2009]. Análises filogenéticas a classificaram como um receptor de melatonina [Bjarnadóttir et al., 2006; Gloriam et al., 2007; Civelli et al., 2013], porém um estudo recente demonstrou sua incapacidade de interação com ligantes de melatonina [Oishi et al., 2017]. Esta sequência possui algumas comunidades determinantes de especificidade para receptores aminérgicos, e sua probabilidade de predição foi de 0.75. Apesar de ainda ser um caso em aberto, um artigo recente de fato demonstrou a capa-

cidade desta proteína de interagir com o 5-(Nonyloxy)Tryptamine [Kozielewicz et al., 2019], um ligante já conhecido por interagir com receptores de serotonina [Glennon et al., 1994, 1996]. Outros acertos recentemente elucidados foram a GPR151, provável receptor de galanina [Liu et al., 2010] e a GPR19, provável receptor de adropina [Rao & Herr, 2017].

4.3 CONAN

O CONAN é a ferramenta computacional desenvolvida para facilitar o uso e a interpretação dos resultados através da metodologia proposta. A ferramenta aceita quatro tipos de entradas: alinhamento múltiplo de sequências (Selex or fasta), uma única sequência no formato fasta, um Pfam ID ou um Uniprot ID. Porém caso seja fornecido uma sequência ou Uniprot ID, o método realizará um *HmmerScan* [Finn et al., 2011] para encontrar o domínio correspondente e posteriormente adquirir o alinhamento diretamente do Pfam. Infelizmente, devido ao fato de ser um sistema *web*, o CONAN possui uma limitação no tamanho do alinhamento de entrada, porém caso o usuário deseje analisar dados maiores, existe a opção de baixar e utilizar diretamente os *scripts* consumidores utilizando o terminal. A ferramenta pode ser acessada em: http://bioinfo.icb.ufmg.br/conan/ [Fonseca et al., 2020].

As páginas de relatórios do CONAN são divididas em 8 abas: principal, rede, conservação, sequências de referências, estrutura, características, aderência e taxonomia. Boa parte das análises realizadas anteriormente neste trabalho já podem ser feitas de forma automatizada pelo CONAN, inclusive o cruzamento de dados com outros bancos de dados, como Uniprot, PDBe, GO e INTERPRO.

4.3.1 Entrada

A pagina de entrada dos resultados permite que o usuário possa refinar os parâmetros da rede e ver como isto afeta as comunidades e as subfamílias geradas por elas. Nesta página, o usuário pode selecionar um corte na rede (figura 4.14A) e para este valor selecionado, visualizar a distribuição de cada comunidade na sequência alinhada, visualizar a matriz de coocorrência de cada comunidade e para cada comunidade visualizar os gráficos de frequência dos resíduos da comunidade em subalinhamentos formados de acordo com termos do *Gene Ontology* e do INTERPRO. O valor de corte selecionado nesta página é salvo em *cache* e será utilizado durante toda navegação na aplicação.

N. Residues	Communities	Avg. Co-Occurrence
2	1	0.8901689708141322
3	1	0.8901689708141322
5	2	0.8675844854070660
6	2	0.8378225806451614
8	3	0.8099145073153534
10	4	0.7879914360420707
11	4	0.7717872347091296
13	5	0.7551781321381645
15	6	0.7438582246582518
17	7	0.7353371248950429
18	6	0.7295199514798206
20	7	0.7184705174174860
22	8	0.7088664808972969

Figura 4.14: Página principal dos relatórios gerados pelo CONAN. A) Seleção do corte da rede; B) Coeficiente de Coocorrência médio por número de comunidades pra cada corte da rede; C) Distribuição das comunidades na sequência consenso; D) Seleção da comunidade; E) Matriz de coocorrência para comunidade selecionada; F) Distribuição de termos (GO e INTERPRO) para subalinhamentos gerados pela presença dos resíduos da comunidade selecionada.

4.3.2 Rede

A aba da rede de coevolução, como o nome já diz, contêm uma visualização da rede gerada. Assim como na página principal, aqui também é possível alterar o valor de corte e ver como isso afeta a rede. Além disto, também é possível alterar parâmetros estéticos da rede, como escalar os nós de acordo com a frequência no alinhamento, colorir os nós de acordo com as comunidades detectadas e escalar as arestas de acordo com escores de coocorrência.

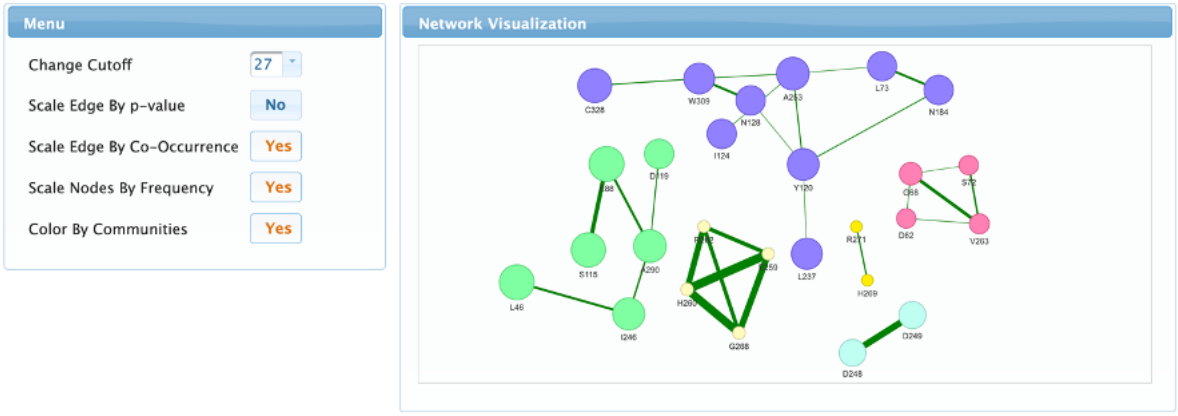


Figura 4.15: Rede de coevolução gerada pelo CONAN

4.3.3 Conservação

A aba de conservação permite que o usuário interaja com o alinhamento múltiplo de sequências original, filtrado e subalinhamentos compostos por sequências que possuem cada padrão de comunidade. A página inclui uma subaplicação incorporada pelo MSAViewer [Yachdav et al., 2016] que permite que o usuário realize uma série de interações com os alinhamentos, como: ordenar e filtrar, buscar sequências e colunas, encontrar motivos, gerar logos, aplicar diferentes escalas de coloração de resíduos, exportar o alinhamento, entre outras (figura 4.16).

4.3.4 Sequências de referência

Esta página permite que o usuário informe um conjunto de sequências presentes no alinhamento de entrada. Cada resíduo de cada comunidade detectada é então mapeado em cada sequência do conjunto informado. Este tipo de análise permite gerar automaticamente tabelas como as do apêndice deste trabalho, onde o usuário pode incluir ortólogos e paralógos e observar quais resíduos de cada comunidade são mantidos e quais são as substituições que ocorreram (figura 4.17).

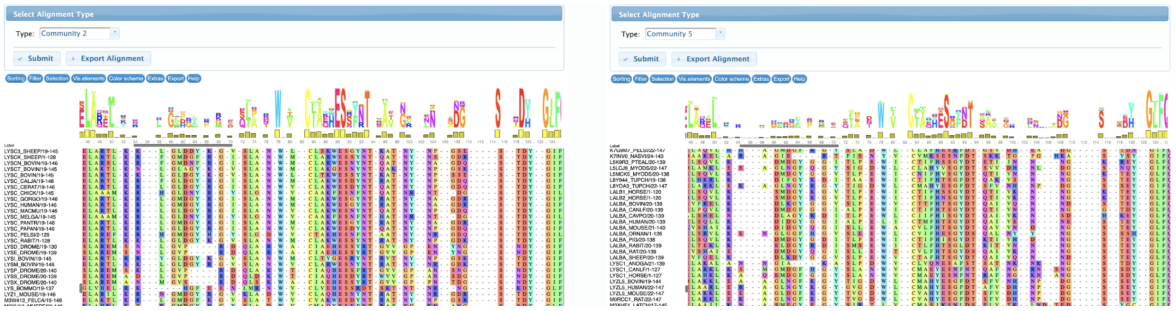


Figura 4.16: Exemplos de dois sub-alinhamentos gerados para comunidades diferentes.

Community 2

Sequence	A290	D119	L46	S115	I246	E88
LYSC_BOVIN/19-145	A125	D71	L26	S69	I107	E53
LYSC1_BOVIN/19-145	A125	D71	L26	S69	I107	E53
LYSC_RABIT/1-128	A108	D53	L8	S51	I89	E35
LYSC_CHICK/19-145	A125	D70	L26	S68	I106	E53
LALBA_RABIT/20-139	H122	E68	L27	S66	L104	T52
LALBA_RAT/20-139	Y122	E68	V27	S66	L104	T52
LALBA_PIG/20-138	Y121	E68	L27	S66	L103	I52
LALBA_SHEEP/20-139	Y122	E68	V27	S66	L104	T52

Community 5

Sequence	D249	D248
LYSC_BOVIN/19-145	A110	K109
LYSC1_BOVIN/19-145	A110	K109
LYSC_RABIT/1-128	A92	Q91
LYSC_CHICK/19-145	S109	A108
LALBA_RABIT/20-139	D107	D106
LALBA_RAT/20-139	D107	D106
LALBA_PIG/20-138	D106	D105
LALBA_SHEEP/20-139	D107	D106

Figura 4.17: Comparação entre duas comunidades detectadas utilizando um conjunto de sequências de Lisozimas e um conjunto de sequências de alfa-lactoalbuminas.

4.3.5 Estruturas

Na aba de estruturas, o usuário pode fornecer um arquivo de estrutura para ser alinhado com uma sequência do alinhamento e posteriormente mapeado com os resultados obtidos. A estrutura pode ser fornecida pelo usuário através de um arquivo no formato PDB ou adquirida automaticamente através da REST API do PDBe. Após a submissão da estrutura, são geradas cinco visualizações. Uma destas visualizações consiste na distribuição das comunidades e anotações automaticamente adquiridas do UniprotKb na sequência da estrutura fornecida (figura 4.18A). Algumas anotações que são mapeadas neste método incluem: estruturas secundárias, sítios ativos, pontes disulfeto e dados de mutações. Além disto, está pagina também permite que o usuário interaja com a estrutura fornecida utilizando escalas de cor de acordo com a frequência de cada resíduo no alinhamento ou de acordo com as comunidades detectadas (figura 4.18B e D). Nesta página também é possível gerar uma rede de contatos estruturais incluindo apenas os nós detectados pelo método, assim o usuário pode comparar de forma interativa como os contatos entre os resíduos detectados mudam de acordo com diferentes estruturas (figura 4.18C). Por fim, este método também gera uma tabela com o mapeamento das posições no alinhamento, sequência e estrutura (figura 4.18E).

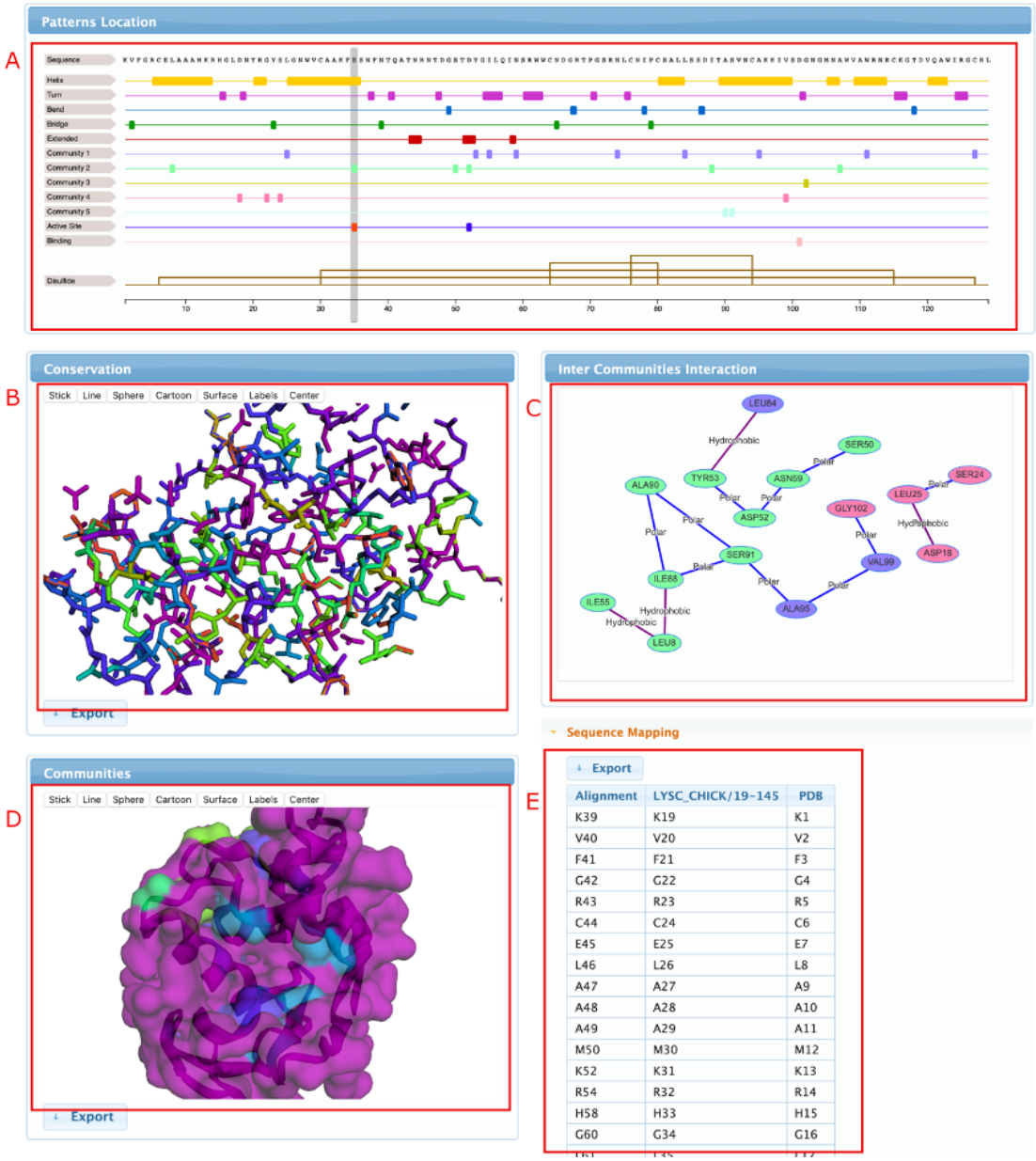


Figura 4.18: Relatório gerado pelo CONAN a partir de um arquivo de estrutura.

4.3.6 Características

A aba de características faz o cruzamento de todos os resíduos detectados em cada comunidade com anotações posicionais adquiridas automaticamente do UniprotKb. Esta função pode ser bastante importante para entender o papel biológico de um grupo de resíduos coevoluídos, pois muitas vezes o sítio determinante de especificidade de uma subfamília vai estar localizado em variações no sítio ativo. Este tipo de padrão pode ser facilmente observado nessa análise, como exemplificado na figura 4.19.

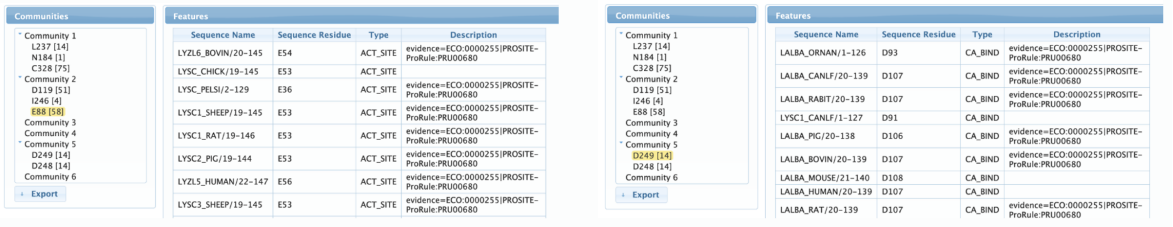


Figura 4.19: Mapeamento de anotações do UniprotKb para dois resíduos de comunidades distintas da família das lisozimas e alfa-lactoalbuminas.

4.3.7 Aderência

Uma das principais formas de se avaliar sítios determinantes de especificidade consiste em plotar gráficos de aderência dos grupos de resíduos em subconjunto de sequências agrupadas por algum critério, como os gráficos e mapas de calor presentes nas análises com dados reais neste trabalho. CONAN permite gerar estes gráficos de forma automática, utilizando sequências extraídas do Swiss-Prot (ou do TrEMBL se o usuário preferir) e com a possibilidade de agrupar de acordo com uma série de fatores, como: classificação enzimática (*E.C. number*), *Gene Ontology* (função molecular, processo biológico e localização celular), INTERPRO, gene, interação com proteínas e envolvimento em doenças (figura 4.21).

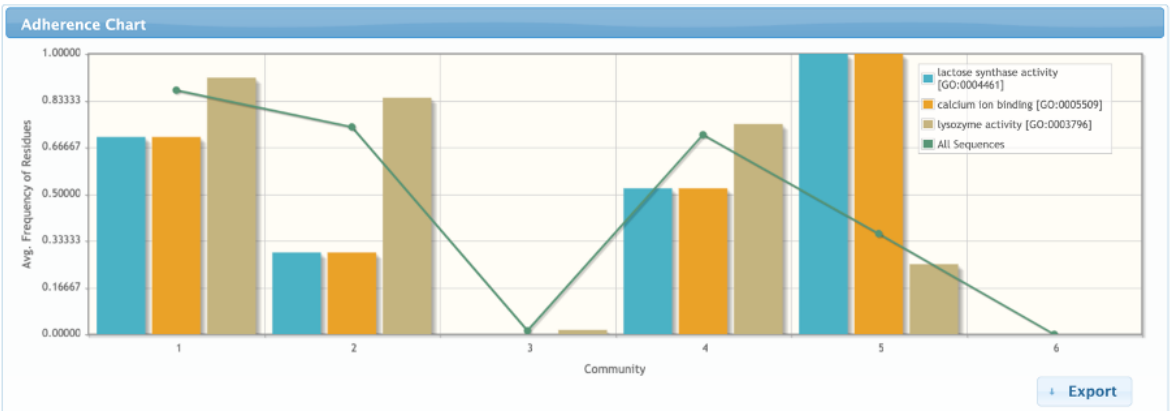


Figura 4.20: Exemplo de gráfico de aderência de comunidades gerado pelo CONAN, utilizando a função molecular (GO) como forma de agrupamento.

4.3.8 Taxonomia

A página de taxonomia permite uma análise similar à da aderência, porém agora em relação a distribuição taxonômica dos subalinhamentos baseado em comunidades. A página contém duas visualizações no estilo *sunburst*, sendo uma distribuição taxonô-

mica do alinhamento completo e a distribuição taxonômica de acordo com a comunidade selecionada. O usuário também pode navegar pelos clados e realçar um clado especifico nos três plots. Além da distribuição taxonômica, a página também contém um gráfico que faz a comparação da frequência dos resíduos da comunidade selecionada no alinhamento e em um subalinhamento baseado no clado selecionado. Este tipo de análise permite encontrar relações de especificidade taxonômica, como a descoberta de um receptor nuclear específico de nemátodos caracterizado por uma especificidade no seu p-box (a região da proteína responsável pela seletividade por um elemento responsivo) [Afonso et al., 2013]. Um exemplo prático pode ser ser visto na figura 4.21, a detecção de uma comunidade de lisozimas não caracterizadas (LLP's), sem o sítio ativo tradicional das lisozimas, e que ocorre especificamente em insetos.

Figura 4.21: Distribuição taxonômica de do alinhamento e de um subalinhamento baseado na comunidade 3. Nesta imagem é possível observar que os resíduos da comunidade selecionada ocorrem especificamente em insetos, em cerca de 10% das sequências do alinhamento e 20% das sequências de insetos no alinhamento.

4.4 CEvADA

Como já demonstrado previamente, dados de coevolução de resíduos podem auxiliar processos de anotações de sequências. Porém, o custo computacional desse tipo de análise o torna praticamente impraticável, tendo em vista que na maioria das vezes o usuário não vai querer anotar uma única sequência, mas sim um proteoma inteiro. Tendo isso em mente, desenvolvemos o CEvADA, um banco de dados de sítios determinantes de especificidades preditos por análise de coevolução de resíduos. O principal

objetivo do CEvADA é ter as correlações pré-calculadas e fornecer uma API para que *softwares* externos adquiram esses dados de forma programática e os incorporem em suas análises. O banco pode ser acessado em http://bioinfo.icb.ufmg.br/cevada.

O CEvADA foi construído em cima do Pfam, utilizando o mesmo esquema de banco de dados como base. Portanto, os dados do mesmo foram também utilizados para calcular as correlações. Atualmente o banco possui dados de aproximadamente 35% dos alinhamentos disponíveis na versão 32 do Pfam, foram incluídos todos os alinhamentos que possuem entre 500 e 20.000 sequências após o pré-processamento. A principal dificuldade em aumentar essa taxa está relacionado a falta de amostragem. Como é possível observar na figura 4.22, a maior parte dos alinhamentos do Pfam possui uma amostragem muito baixa, insuficiente para realizar qualquer tipo de análise estatística.

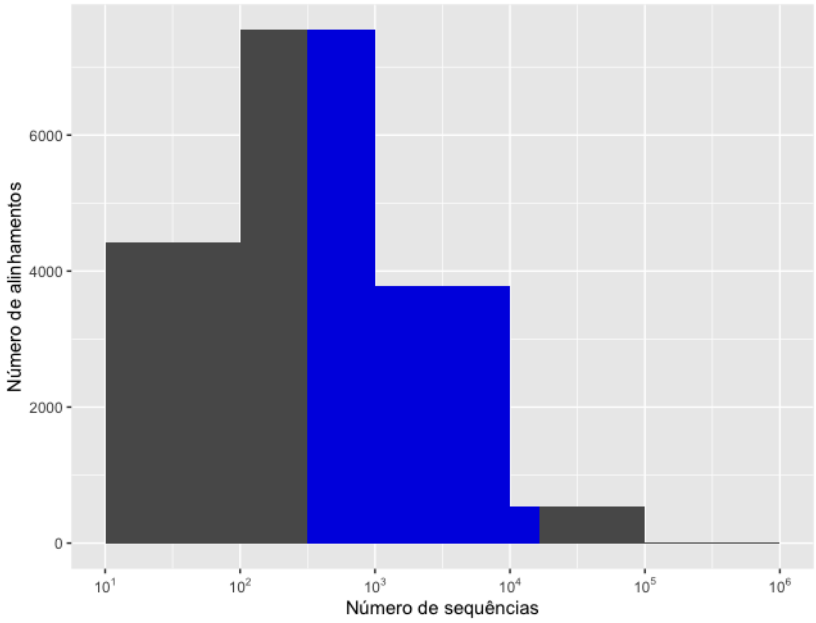


Figura 4.22: Distribuição do tamanho dos alinhamentos do Pfam. Destacado em azul estão as famílias atualmente inseridas no CEvADA

4.4.1 REST API

O sistema possui dois pontos finais de uma API REST para fazer a comunicação do banco com softwares externos e internos, sendo um retornando dados relativos a uma família de proteínas e o outro relativo a uma única proteína. Ambos os pontos fazem a comunicação via GET com saída no formato JSON. Além disto, ambos aceitam dois tipos de entradas, no caso de família é aceito o Pfam ID ou o código de acesso,

e de forma semelhante, no caso de proteínas, é aceito o UniProtKb ID ou o código de acesso.

Além de informações gerais da família ou da proteína, os pontos retornam os dados de coevolução já mapeados de acordo com o tipo de entrada, mostrando inclusive possíveis substituições. No caso de uma família de proteínas, os dados são retornados mapeados para todas as sequências do alinhamento *Full* do Pfam, no caso de uma única proteína, os dados são mapeados para todos os domínios que representem entradas no Pfam. Exemplos de saídas podem ser visto na figura 4.23 e os esquemas completos podem ser acessados no material suplementar, nas figuras A.16 e A.17

a) Sequência b) Família

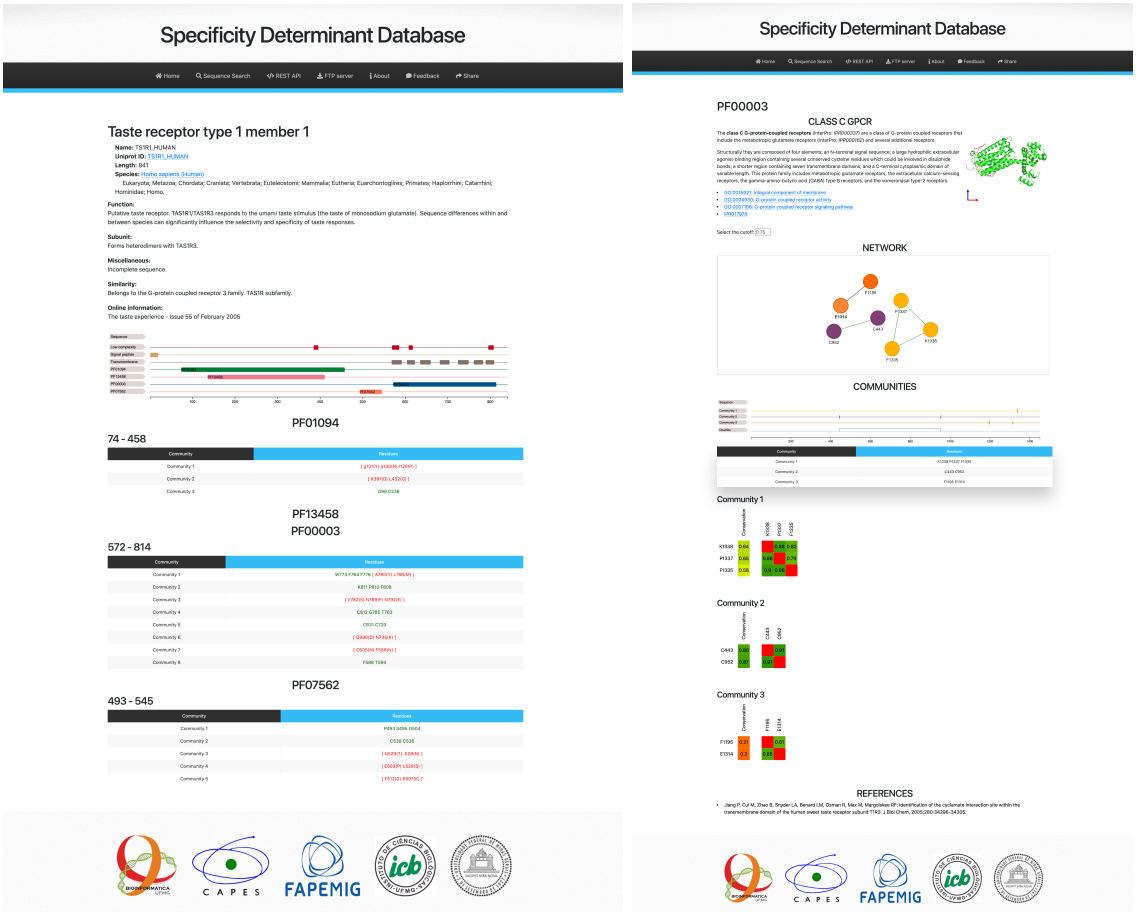
Figura 4.23: Exemplos de saída da API do CEvADA. a) ponto de saída de sequências e b) ponto de saída de família.

4.4.2 Vistas

As vistas são páginas de entradas de um banco de dados. No caso do CEvADA, o sistema web do banco possui duas vistas para facilitar o acesso aos dados através do navegador: sequência e família. Ambas as vistas são alimentadas pela própria API do CEvADA, descrita anteriormente, além de também utilizar dados de API's externas, como o UniProtKb, PDBe, Gene Ontology Resource e Wikipedia.

As páginas de entrada de sequências, como pode ser visto na figura 4.24a, contêm a descrição e anotações gerais extraídas do UniProtKb. Além disto, todas as regiões e domínios conhecidos, incluindo domínios com entrada no Pfam, são mapeados e ilustrados em uma visualização de dados adaptada do ProtVista [Watkins et al., 2017]. Para cada domínio Pfam, se ele é representado no CEvADA, é mostrado sua lista de comunidades detectadas, com os resíduos já representados utilizando a numeração da proteína da página e caso o resíduo não seja conservado nesta proteína, é mostrado o aminoácido substituído.

A página que representa entradas de famílias de proteínas, exemplificada na figura 4.24b, assim como o próprio Pfam, utiliza descrições automaticamente extraídas do Wikipedia. Também é incluso links de termos GO e INTERPRO que representem a família e as referências de artigos que o Pfam utilizou para definir a família. A página também inclui visualizações de dados que também estao presentes no CONAN, como a rede, localização das comunidades no alinhamento e as matrizes de correlação.



a) Sequência b) Família

Figura 4.24: Vistas do CEvADA

Capítulo 5

Considerações Finais

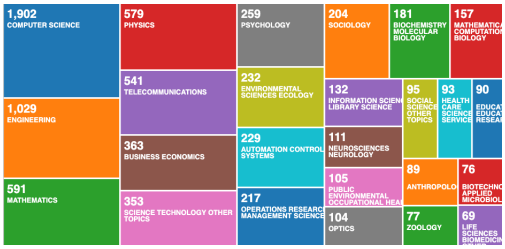
5.1 Conclusões

Foi demonstrado neste trabalho que a representação de um alinhamento múltiplo de sequências como uma rede de afiliação de resíduos pode ser efetiva para a detecção de padrões de coevolução de resíduos e detecção de sítios determinantes de especificidades. Isto é um fator interessante, afinal abre uma a possibilidade para que esta modelagem seja aplicada a diversos outros problemas biológicos. Um possível exemplo consiste em analisar a outra projeção da rede bipartida, gerando um agrupamento de sequências de acordo com a tendencia em possuir os mesmos conjuntos de resíduos.

O método apresentado aqui possui algumas características incomuns em relação à literatura, como a capacidade de realizar análises exploratórias de determinantes de especificidades e a capacidade de se trabalhar com alinhamentos com um grande número de sequências e colunas. Apesar de haver uma restrição em relação ao tamanho do alinhamento de entrada no CONAN, ele também abre uma nova possibilidade de se analisar coevolução de resíduos de uma forma simples, no próprio navegador e sem a necessidade de baixar programas e bibliotecas de terceiros. Além disto, esta restrição é em partes compensada pelo CEvADA, que já possui dados pré calculados para a maior parte de famílias plausíveis de serem analisadas (que possuem um número mínimo de sequências) do Pfam. Neste caso, a simplicidade é ainda maior, pois só resta ao usuário interpretar os resultados.

O fato de redes de afiliações serem cada vez mais utilizadas e nos mais variados contextos (como pode ser visto na figura 5.1), faz com que novos métodos e aprimoramentos neste tipo de análises sejam constantemente publicados e consequentemente podem ser utilizados para aprimorar ainda mais a eficiência e a complexidade desta abordagem.

a) Publicações anuais



b) Distribuição por áreas de concentração

Figura 5.1: Artigos relacionados a redes bipartidas no *Web of Science*

Foi demonstrado que sinais de coevolução de resíduos podem ser utilizados como estimadores para classificação funcional de sequências. Porém, o custo computacional para calcular coevolução em tempo de execução de uma anotação praticamente inviabiliza esse tipo de abordagem. Contudo, a API presente no CEvADA permite que ferramentas externas adquiram estes dados, já previamente calculados, em tempo de execução e utilize para alimentar algorítmos de aprendizagem de máquina.

A metodologia apresentada neste trabalho contribui não apenas com a biologia computacional, mas com a ciência como um todo, uma vez que esta pode ser adaptada e aplicada em diversos outros contextos relacionados a sistemas de afiliação.

5.2 Trabalhos Futuros

Como a área de ciência das redes tem evoluído muito rápido, é importante avaliar novas abordagens de validação de arestas e detecção de comunidades em redes bipartidas que podem aprimorar a metodologia proposta neste trabalho. Além disto, um dos planos futuros é incorporar a metodologia proposta na ferramenta PFstats [Fonseca-Júnior et al., 2018], para que os usuários possam analisar alinhamentos maiores no próprio computador com auxílio de uma interface gráfica e com acesso a todas as visualizações de dados geradas pelo CONAN e pelo PFstats.

Em relação ao CONAN, pretendemos adicionar suporte para que os usuários obtenham novos tipos de alinhamentos de forma automatizada, além do *Full* do Pfam, como os baseados em proteomas representativos, NCBI e metagenômica. Outro ponto a ser aprimorado é permitir o mapeamento de uma sequência a múltiplas cadeias de uma estrutura.

O CEvADA ainda está em desenvolvimento, portanto os próximos planos incluem

adicionar uma ferramenta de busca por sequências (blast) utilizando apenas sequências que possuem entradas no CEvADA. Adicionar uma máquina de busca eficiente, atualmente o CEvADA utiliza um sistema de busca simples, utilizando apenas combinação perfeita com algumas chaves do banco. O plano para o futuro é utilizar a ferramenta Solr, para indexar uma máquina de busca moderna. Criar um servidor FTP para facilitar o acesso aos dados brutos do banco de dados.

Finalmente, após a publicação do CEvADA, planejamos desenvolver uma nova ferramenta, baseada em técnicas de de aprendizagem de máquina, que utilize os dados fornecidos pela API do CEvADA para realizar predição funcional de proteínas.

Referências Bibliográficas

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

Ahn, Y.-Y.; Ahnert, S. E.; Bagrow, J. P. & Barabási, A.-L. (2011). Flavor network and the principles of food pairing. *Scientific reports*, 1:196.

Almende, B.; Thieurmel, B. & Robert, T. (2016). visnetwork: Network visualization using vis.js library. *R package version*, 1(1).

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403--410.

Andreeva, A.; Kulesha, E.; Gough, J. & Murzin, A. G. (2020). The scop database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic acids research*, 48(D1):D376--D382.

Armstrong, D. R.; Berrisford, J. M.; Conroy, M. J.; Gutmanas, A.; Anyango, S.; Choudhary, P.; Clark, A. R.; Dana, J. M.; Deshpande, M.; Dunlop, R. et al. (2020). Pdbe: improved findability of macromolecular structure data in the pdb. *Nucleic acids research*, 48(D1):D335--D343.

Atchley, W. R.; Terhalle, W. & Dress, A. (1999). Positional dependence, cliques, and predictive motifs in the bhlh protein domain. *Journal of molecular evolution*, 48(5):501--516.

Bachega, J. F. R.; Navarro, M. V. A. S.; Bleicher, L.; Bortoleto-Bugs, R. K.; Dive, D.; Hoffmann, P.; Viscogliosi, E. & Garratt, R. C. (2009). Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. *Proteins: Structure, Function, and Bioinformatics*, 77(1):26--37.

Barabási, A.-L. (2016). *Network science*. Cambridge university press.

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):60--69.

Barabási, A.-L.; Gulbahce, N. & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56.

Barwinska-Sendra, A.; Garcia, Y. M.; Sendra, K. M.; Baslé, A.; Mackenzie, E. S.; Tarrant, E.; Card, P.; Tabares, L. C.; Bicep, C.; Un, S. et al. (2020). An evolutionary path to altered cofactor specificity in a metalloenzyme. *Nature Communications*, 11(1):1--13.

Bennett, J. M. & Kendrew, J. C. (1952). The computation of fourier synthesis with a digital electronic calculating machine. *Acta Crystallographica*, 5(1):109--116.

Berg, J. M.; Tymoczko, J. L. & Stryer, L. (2002). Biochemistry, ; w. h.

Betts, M. J. & Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists*, 317:289.

Bjarnadóttir, T. K.; Gloriam, D. E.; Hellstrand, S. H.; Kristiansson, H.; Fredriksson, R. & Schiöth, H. B. (2006). Comprehensive repertoire and phylogenetic analysis of the g protein-coupled receptors in human and mouse. *Genomics*, 88(3):263--273.

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.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113--120.

Borate, B. R.; Chesler, E. J.; Langston, M. A.; Saxton, A. M. & Voy, B. H. (2009). Comparison of threshold selection methods for microarray gene co-expression matrices. *BMC research notes*, 2(1):240.

Borgatti, S. P. & Halgin, D. S. (2011). Analyzing affiliation networks. *The Sage handbook of social network analysis*, 1:417--433.

Bostock, M.; Ogievetsky, V. & Heer, J. (2011). D^3 data-driven documents. *IEEE Transactions on Visualization & Computer Graphics*, (12):2301--2309.

Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bansal, P.; Bridge, A. J.; Poux, S.; Bougueleret, L. & Xenarios, I. (2016). Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. Em *Plant Bioinformatics*, pp. 23--54. Springer.

Cayley, A. (1874). On the mathematical theory of isomers. *Philosophical Magazine, Series 5*, 47:444--446.

Cendron, L.; Ramazzina, I.; Percudani, R.; Rasore, C.; Zanotti, G. & Berni, R. (2011). Probing the evolution of hydroxyisourate hydrolase into transthyretin through active-site redesign. *Journal of molecular biology*, 409(4):504--512.

Chakrabarti, S.; Bryant, S. H. & Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids. *Journal of molecular biology*, 373(3):801--810.

Chakraborty, A. & Chakrabarti, S. (2014). A survey on prediction of specificity-determining sites in proteins. *Briefings in bioinformatics*, 16(1):71--88.

Chang, C.-S.; NEGISHI, M.; NISHIGAKI, N. & ICHIKAWA, A. (1997). Functional interaction of the carboxylic acid group of agonists and the arginine residue of the seventh transmembrane domain of prostaglandin e receptor ep3 subtype. *Biochemical Journal*, 322(2):597--601.

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.

Civelli, O.; Reinscheid, R. K.; Zhang, Y.; Wang, Z.; Fredriksson, R. & Schiöth, H. B. (2013). G protein–coupled receptor deorphanizations. *Annual review of pharmacology and toxicology*, 53:127--146.

Coitinho, J. B.; Costa, M. A.; Melo, E. M.; Morais, E. A.; de Andrade, L. G.; da Rocha, A. M.; de Magalhães, M. T.; Favaro, D. C.; Bleicher, L.; Pedroso, E. R. et al. (2019). Structural and immunological characterization of a new nucleotidyltransferase-like antigen from paracoccidioides brasiliensis. *Molecular immunology*, 112:151--162.

Connor, N.; Barberán, A. & Clauset, A. (2017). Using null models to infer microbial co-occurrence networks. *PloS one*, 12(5):e0176751.

Cook, C. E.; Lopez, R.; Stroe, O.; Cochrane, G.; Brooksbank, C.; Birney, E. & Apweiler, R. (2019). The european bioinformatics institute in 2018: tools, infrastructure and training. *Nucleic acids research*, 47(D1):D15--D22.

da Fonseca Jr, N. J.; Afonso, M. Q. L.; de Oliveira, L. C. & Bleicher, L. (2019). A new method bridging graph theory and residue co-evolutionary networks for specificity determinant positions detection. *Bioinformatics*, 35(9):1478--1485.

da Fonseca Jr, N. J.; Afonso, M. Q. L.; Pedersolli, N. G.; de Oliveira, L. C.; Andrade, D. S. & Bleicher, L. (2017). Sequence, structure and function relationships in flaviviruses as assessed by evolutive aspects of its conserved non-structural protein domains. *Biochemical and biophysical research communications*, 492(4):565--571.

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

Dayhoff, M.; Schwartz, R. & Orcutt, B. (1978). 22 a model of evolutionary change in proteins. Em *Atlas of protein sequence and structure*, volume 5, pp. 345--352. National Biomedical Research Foundation Silver Spring, MD.

Dayhoff, M. O. (1965). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.

Dayhoff, M. O. & Ledley, R. S. (1962). Comprotein: a computer program to aid primary protein structure determination. Em *Proceedings of the December 4-6, 1962, fall joint computer conference*, pp. 262--274. ACM.

Dayhoff, M. O.; McLaughlin, P. J.; Barker, W. C. & Hunt, L. T. (1975). Evolution of sequences within protein superfamilies. *Naturwissenschaften*, 62(4):154--161.

Dianati, N. (2016). Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Physical Review E*, 93(1):012304.

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.

Eck, R. V. & Dayhoff, M. O. (1966). Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*, 152(3720):363--366.

Eddy, S. R. et al. (1995). Multiple alignment using hidden markov models. Em *Ismb*, volume 3, pp. 114--120.

El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A. et al. (2019). The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427--D432.

Emms, D. M. & Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1):157.

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

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

Euler, L. (1736). Solutio problematis ad geometriam situs pertinensis. *Comm. Acad. Sci. Imper. Petropol.*, 8:128--140.

Feng, D.-F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, 25(4):351--360.

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.

Finn, R. D.; Clements, J. & Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29--W37.

Fleming, A. (1922). On a remarkable bacteriolytic element found in tissues and secretions. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 93(653):306--317.

Fonseca, N.; Afonso, M.; Carrijo, L. & Bleicher, L. (2020). Conan: A web application to detect specificity determinants and functional sites by amino acids co-variation network analysis. *Bioinformatics*.

Fonseca-Júnior, N. J.; Afonso, M. Q.; Oliveira, L. C. & Bleicher, L. (2018). Pfstats: A network-based open tool for protein family analysis. *Journal of Computational Biology*.

Fortunato, S. & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1--44.

Fredriksson, R.; Lagerström, M. C.; Lundin, L.-G. & Schiöth, H. B. (2003). The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Molecular pharmacology*, 63(6):1256--1272.

Fu, L.; Niu, B.; Zhu, Z.; Wu, S. & Li, W. (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150--3152.

Gardner, J. C.; Webb, T. R.; Kanuga, N.; Robson, A. G.; Holder, G. E.; Stockman, A.; Ripamonti, C.; Ebenezer, N. D.; Ogun, O.; Devery, S. et al. (2010). X-linked cone dystrophy caused by mutation of the red and green cone opsins. *The American Journal of Human Genetics*, 87(1):26--39.

George, R. A. & Heringa, J. (2002). An analysis of protein domain linkers: their classification and role in protein folding. *Protein Engineering, Design and Selection*, 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.

Girvan, M. & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821--7826.

Glennon, R. A.; Hong, S.-S.; Bondarev, M.; Law, H.; Dukat, M.; Rakhit, S.; Power, P.; Fan, E.; Kinneau, D.; Kamboj, R. et al. (1996). Binding of o-alkyl derivatives of serotonin at human 5-ht1dβ receptors. *Journal of medicinal chemistry*, 39(1):314--322.

Glennon, R. A.; Hong, S.-S.; Dukat, M.; Teitler, M. & Davis, K. (1994). 5-(nonyloxy) tryptamine: a novel high-affinity 5-ht1d. beta. serotonin receptor agonist. *Journal of medicinal chemistry*, 37(18):2828--2830.

Gloriam, D. E.; Fredriksson, R. & Schiöth, H. B. (2007). The g protein-coupled receptor subset of the rat genome. *BMC genomics*, 8(1):338.

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.

Goh, K.-I.; Cusick, M. E.; Valle, D.; Childs, B.; Vidal, M. & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685--8690.

Gough, J.; Karplus, K.; Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of molecular biology*, 313(4):903--919.

Guimera, R.; Mossa, S.; Turtschi, A. & Amaral, L. N. (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.

Gupta, C.; Singh, S. & Kumar, S. (2010). *Advance Discrete Structure*. IK International Publishing House Pvt. Ltd.

Hagberg, A.; Schult, D. & Swart, P. (2005). Networkx: Python software for the analysis of networks. *Mathematical Modeling and Analysis, Los Alamos National Laboratory*.

Halabi, N.; Rivoire, O.; Leibler, S. & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774--786.

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

Hamilton, W. R. (1856). Lvi. memorandum respecting a new system of roots of unity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 12(81):446--446.

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*.

Herrero, M. B.; Mandal, A.; Digilio, L. C.; Coonrod, S. A.; Maier, B. & Herr, J. C. (2005). Mouse sllp1, a sperm lysozyme-like protein involved in sperm–egg binding and fertilization. *Developmental biology*, 284(1):126--142.

Hesper, B. & Hogeweg, P. (1970). Bioinformatica: een werkconcept. kameleon 1 (6): 28–29. *Dutch.) Leiden: Leidse Biologen Club*.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS computational biology*, 7(3):e1002021.

Hogeweg, P. & Hesper, B. (1978). Interactive instruction on population interactions. *Computers in biology and medicine*, 8(4):319--327.

Hogeweg, P. & Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, 20(2):175--186.

Holland, R. C.; Down, T. A.; Pocock, M.; Prlić, A.; Huen, D.; James, K.; Foisy, S.; Dräger, A.; Yates, A.; Heuer, M. et al. (2008). Biojava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096--2097.

Hopkins, B. & Wilson, R. (2004). The truth about kõnigsberg. *The College Mathematics Journal*, 35(3):198.

Huang, C. & Tai, H.-H. (1995). Expression and site-directed mutagenesis of mouse prostaglandin e2 receptor ep3 subtype in insect cells. *Biochemical Journal*, 307(2):493--498.

Huang, P.; Li, W.; Yang, Z.; Zhang, N.; Xu, Y.; Bao, J.; Jiang, D. & Dong, X. (2017). Lyzl6, an acidic, bacteriolytic, human sperm-related protein, plays a role in fertilization. *PloS one*, 12(2):e0171452.

Huhtaniemi, I. T. (2017). Male hypogonadism resulting from mutations in the genes for the gonadotropin subunits and their receptors. Em *Male Hypogonadism*, pp. 127--152. Springer.

Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; De Castro, E.; Langendijk-Genevaux, P. S.; Pagni, M. & Sigrist, C. J. (2006). The prosite database. *Nucleic acids research*, 34(suppl_1):D227--D230.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37--50.

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

Jollès, P. & Jollès, J. (1984). What's new in lysozyme research? *Molecular and cellular biochemistry*, 63(2):165--189.

Jones, E.; Oliphant, T. & Peterson, P. (2014). {SciPy}: open source scientific tools for {Python}.

Juel Mortensen, L.; Blomberg Jensen, M.; Christiansen, P.; Rønholt, A.-M.; Jørgensen, A.; Frederiksen, H.; Nielsen, J. E.; Loya, A. C.; Grønkær Toft, B.; Skakkebæk, N. E. et al. (2017). Germ cell neoplasia in situ and preserved fertility despite suppressed gonadotropins in a patient with testotoxicosis. *The Journal of Clinical Endocrinology & Metabolism*, 102(12):4411--4416.

Kedzie, K. M.; Donello, J. E.; Krauss, H. A.; Regan, J. W. & Gil, D. W. (1998). A single amino-acid substitution in the ep2prostaglandin receptor confers responsiveness to prostacyclin analogs. *Molecular pharmacology*, 54(3):584--590.

Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H. & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662--666.

Kimura, M. et al. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624--626.

Kirchhoff, G. (1847). Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497--508.

Kirkman, T. P. et al. (1856). Xvii. on the enumeration of x-edra having triedral summits, and an (x-1)-gonal base. *Philosophical Transactions of the Royal Society of London*, 146:399--411.

Ko, H.-J.; Lee, E. W.; Bang, W.-G.; Lee, C.-K.; Kim, K. H. & Choi, I.-G. (2010). Molecular characterization of a novel bacterial aryl acylamidase belonging to the amidase signature enzyme family. *Molecules and cells*, 29(5):485--492.

Kolakowski, J. L. (1994). Gcrdb: a g-protein-coupled receptor database. *Receptors & channels*, 2(1):1--7.

Kosugi, S.; Mori, T. & Shenker, A. (1996). The role of asp578 in maintaining the inactive conformation of the human lutropin/choriogonadotropin receptor. *Journal of Biological Chemistry*, 271(50):31813--31817.

Kozielewicz, P.; Grafton, G.; Sajkowska-Kozielewicz, J. J. & Barnes, N. M. (2019). Overexpression of orphan receptor gpr61 increases camp levels upon forskolin stimulation in hek293 cells: in vitro and in silico validation of 5-(nonyloxy) tryptamine as a low-affinity inverse agonist. *Pharmacology*, 104(5-6):376--381.

Krah, M.; Misselwitz, R.; Politz, O.; Thomsen, K. K.; Welfle, H. & Borriss, R. (1998). The laminarinase from thermophilic eubacterium rhodothermus marinus: Conformation, stability, and identification of active site carboxylic residues by site-directed mutagenesis. *European journal of biochemistry*, 257(1):101--111.

Kwak, J. H.; Shin, K.; Woo, J. S.; Kim, M. K.; Kim, S. I.; Eom, S. H. & Kwang-Won, H. (2002). Expression, purification, and crystallization of glutamyl-trna gln specific amidotransferase from bacillus stearothermophilus. *Molecules and cells*, 14(3):374--381.

Lavrenko, V. (2014). Agglomerative clustering: how it works. https://www.youtube.com/watch?v=XJ3194AmH40. Accessed: 2020-06-01.

Lee, Y.; Lee, D. H.; Kho, C. W.; Lee, A. Y.; Jang, M.; Cho, S.; Lee, C. H.; Lee, J. S.; Myung, P. K.; Park, B. C. et al. (2005). Transthyretin-related proteins function to facilitate the hydrolysis of 5-hydroxyisourate, the end product of the uricase reaction. *FEBS letters*, 579(21):4769--4774.

Li, L.; Stoeckert, C. J. & Roos, D. S. (2003a). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178--2189.

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

Li, T.; Franson, W. K.; Gordon, J. W.; Berson, E. L. & Dryja, T. P. (1995). Constitutive activation of phototransduction by k296e opsin is not a cause of photoreceptor degeneration. *Proceedings of the National Academy of Sciences*, 92(8):3551--3555.

Lichtarge, O.; Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2):342--358.

Liljeros, F.; Edling, C. R.; Amaral, L. A. N.; Stanley, H. E. & Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907.

Liu, Z.; Xu, Y.; Wu, L. & Zhang, S. (2010). Evolution of galanin receptor genes: insights from the deuterostome genomes. *Journal of biomolecular structure and dynamics*, 28(1):97--106.

Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Bioinformatics*, 9(6):745--756.

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.

Mandal, A.; Klotz, K. L.; Shetty, J.; Jayes, F. L.; Wolkowicz, M. J.; Bolling, L. C.; Coonrod, S. A.; Black, M. B.; Diekman, A. B.; Haystead, T. A. et al. (2003). Sllp1, a unique, intra-acrosomal, non-bacteriolytic, c lysozyme-like protein of human spermatozoa. *Biology of reproduction*, 68(5):1525--1537.

McKinney, M. K. & Cravatt, B. F. (2005). Structure and function of fatty acid amide hydrolase. *Annu. Rev. Biochem.*, 74:411--432.

McKinney, W. (2011). Pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, pp. 1--9.

Mitchell, A. L.; Attwood, T. K.; Babbitt, P. C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S. D.; Chang, H.-Y.; El-Gebali, S.; Fraser, M. I. et al. (2019). Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research*, 47(D1):D351--D360.

Møller, T. C.; Moreno-Delgado, D.; Pin, J.-P. & Kniazeff, J. (2017). Class cg protein-coupled receptors: reviving old couples with new partners. *Biophysics reports*, 3(4-6):57--63.

Mombaerts, P. (2004). Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Reviews Neuroscience*, 5(4):263.

Munk, C.; Isberg, V.; Mordalski, S.; Harpsøe, K.; Rataj, K.; Hauser, A.; Kolb, P.; Bojarski, A.; Vriend, G. & Gloriam, D. (2016). Gpcrdb: the g protein-coupled receptor database–an introduction. *British journal of pharmacology*, 173(14):2195--2207.

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.

Musacchio, A.; Gibson, T.; Lehto, V.-P. & Saraste, M. (1992). Sh3 - an abundant protein domain in search of a function. *FEBS letters*, 307(1):55--61.

Nathans, J.; Maumenee, I. H.; Zrenner, E.; Sadowski, B.; Sharpe, L. T.; Lewis, R. A.; Hansen, E.; Rosenberg, T.; Schwartz, M.; Heckenlively, J. R. et al. (1993). Genetic heterogeneity among blue-cone monochromats. *American journal of human genetics*, 53(5):987.

Neal, J. W. & Neal, Z. P. (2013). The multiple meanings of peer groups in social cognitive mapping. *Social Development*, 22(3):580--594.

Neal, Z. (2014). The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84--97.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443--453.

Negishi, M.; Harazono, A.; Sugimoto, Y.; Hazato, A.; Kurozumi, S. & Ichikawa, A. (1995). Selective coupling of prostaglandin e receptor ep3d to multiple g proteins depending on interaction of the carboxylic acid of agonist and arginine residue of seventh transmembrane domain. *Biochemical and biophysical research communications*, 212(2):279--285.

Nelson, D. L. & Cox, M. M. (2018). *Princípios de Bioquímica de Lehninger-7*. Artmed Editora.

Neumann, S. & Kula, M.-R. (2002). Gene cloning, overexpression and biochemical characterization of the peptide amidase from stenotrophomonas maltophilia. *Applied microbiology and biotechnology*, 58(6):772--780.

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.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96--98.

Ohta, T. (2001). Nearly neutral theory. Em Brenner, S. & Miller, J. H., editores, *Encyclopedia of Genetics*, pp. 1301 – 1302. Academic Press, New York.

Oishi, A.; Karamitri, A.; Gerbier, R.; Lahuna, O.; Ahmad, R. & Jockers, R. (2017). Orphan gpr61, gpr62 and gpr135 receptors and the melatonin mt 2 receptor reciprocally modulate their signaling functions. *Scientific reports*, 7(1):8990.

Ort, E. & Mehta, B. (2003). Java architecture for xml binding (jaxb). *Sun Developer Network*.

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

Pauling, L.; Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205--211.

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

Pedruzzi, I.; Rivoire, C.; Auchincloss, A. H.; Coudert, E.; Keller, G.; De Castro, E.; Baratin, D.; Cuche, B. A.; Bougueleret, L.; Poux, S. et al. (2014). Hamap in 2015: updates to the protein family classification and annotation system. *Nucleic acids research*, 43(D1):D1064--D1070.

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.

Perkins, A. D. & Langston, M. A. (2009). Threshold selection in gene co-expression networks using spectral graph theory techniques. Em *BMC bioinformatics*, volume 10, p. S4. BioMed Central.

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.

Querino Lima Afonso, M.; Fonseca, N. J.; de Oliveira, L. C.; Lobo, F. P. & Bleicher, L. (2020). Coevolved positions represent key functional properties in the trypsin-like serine proteases protein family. *Journal of Chemical Information and Modeling*.

Rao, A. & Herr, D. R. (2017). G protein-coupled receptor gpr19 regulates e-cadherin expression and invasion of breast cancer cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1864(7):1318--1327.

Rask-Andersen, M.; Masuram, S. & Schiöth, H. B. (2014). The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology*, 54:9--26.

Rego, N. & Koes, D. (2014). 3dmol. js: molecular visualization with webgl. *Bioinformatics*, 31(8):1322--1324.

Richardson, S. J. (2015). Tweaking the structure to radically change the function: the evolution of transthyretin from 5-hydroxyisourate hydrolase to triiodothyronine distributor to thyroxine distributor. *Frontiers in endocrinology*, 5:245.

Rios-Anjos, R. M.; de Lima Camandona, V.; Bleicher, L. & Ferreira-Junior, J. R. (2017). Structural and functional mapping of rtg2p determinants involved in retrograde signaling and aging of saccharomyces cerevisiae. *PloS one*, 12(5):e0177090.

Romero, P. A. & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866.

Rouvray, D. (1989). The pioneering contributions of cayley and sylvester to the mathematical description of chemical structure. *Journal of Molecular Structure: THEOCHEM*, 185:1--14.

Sankoff, D.; Cedergren, R. J. & Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5s ribosomal rna. *Journal of Molecular Evolution*, 7(2):133--149.

Saracco, F.; Straka, M. J.; Di Clemente, R.; Gabrielli, A.; Caldarelli, G. & Squartini, T. (2017). Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, 19(5):053022.

Schnoes, A. M.; Brown, S. D.; Dodevski, I. & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605.

Serrano, M. Á.; Boguná, M. & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483--6488.

Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498--2504.

Shin, S.; Lee, T.-H.; Ha, N.-C.; Koo, H. M.; Kim, S.-y.; Lee, H.-S.; Kim, Y. S. & Oh, B.-H. (2002). Structure of malonamidase e2 reveals a novel ser-cisser-lys catalytic triad in a new serine hydrolase fold that is prevalent in nature. *The EMBO Journal*, 21(11):2509--2516.

Sillitoe, I.; Dawson, N.; Lewis, T. E.; Das, S.; Lees, J. G.; Ashford, P.; Tolulope, A.; Scholes, H. M.; Senatorov, I.; Bujan, A. et al. (2019). Cath: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic acids research*, 47(D1):D280--D284.

Siviero-Miachon, A. A.; Kizys, M. M.; Ribeiro, M. M.; Garcia, F. E.; Spinola-Castro, A. M. & Dias da Silva, M. R. (2017). Cosegregation of a novel mutation in the sixth transmembrane segment of the luteinizing/choriogonadotropin hormone receptor with two brazilian siblings with severe testotoxicosis. *Endocrine research*, 42(2):117--124.

Smith, T. F. & Waterman, M. S. (1981). Comparison of biosequences. *Advances in applied mathematics*, 2(4):482--489.

Song, J. S.; Gonzales, N. R.; Yamashita, R. A.; Bryant, S. H. & Marchler-Bauer, A. (2017). Cdd: functional insights into orphan gpcrs via subfamily domain architectures.

Stuart, D.; Acharya, K.; Walker, N.; Smith, S.; Lewis, M. & Phillips, D. (1986). α-lactalbumin possesses a novel calcium binding loop. *Nature*, 324(6092):84--87.

Suhadolnik, M. L.; Salgado, A. P.; Scholte, L. L.; Bleicher, L.; Costa, P. S.; Reis, M. P.; Dias, M. F.; Ávila, M. P.; Barbosa, F. A.; Chartone-Souza, E. et al. (2017). Novel arsenic-transforming bacteria and the diversity of their arsenic-related genes and enzymes arising from arsenic-polluted freshwater sediment. *Scientific reports*, 7(1):11231.

Sung, C.-H.; Schneider, B. G.; Agarwal, N.; Papermaster, D. S. & Nathans, J. (1991). Functional heterogeneity of mutant rhodopsins responsible for autosomal dominant retinitis pigmentosa. *Proceedings of the National Academy of Sciences*, 88(19):8840--8844.

Sylvester, J. J. (1878). On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. *American Journal of Mathematics*, 1(1):64--104.

Taylor, P. J. (2001). Specification of the world city network. *Geographical analysis*, 33(2):181--194.

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.

Tonacchera, M.; Agretti, P.; Chiovato, L.; Rosellini, V.; Ceccarini, G.; Perri, A.; Viacava, P.; Naccarato, A. G.; Miccoli, P.; Pinchera, A. et al. (2000). Activating thyrotropin receptor mutations are present in nonadenomatous hyperfunctioning nodules of toxic or autonomous multinodular goiter. *The Journal of Clinical Endocrinology & Metabolism*, 85(6):2270--2274.

Toyooka, M.; Tujii, T. & Takeda, S. (2009). The n-terminal domain of gpr61, an orphan g-protein-coupled receptor, is essential for its constitutive activity. *Journal of neuroscience research*, 87(6):1329--1333.

Tumminello, M.; Micciche, S.; Lillo, F.; Piilo, J. & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994.

Valdar, W. S. (2002). Scoring residue conservation. *Proteins: structure, function, and bioinformatics*, 48(2):227--241.

Valiña, A. L. B.; Mazumder-Shivakumar, D. & Bruice, T. C. (2004). Probing the ser-ser-lys catalytic triad mechanism of peptide amidase: computational studies of the ground state, transition state, and intermediate. *Biochemistry*, 43(50):15657--15672.

Van Noorden, R.; Maher, B. & Nuzzo, R. (2014). The top 100 papers. *Nature News*, 514(7524):550.

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

Waterman, M. S. & Perlwitz, M. D. (1984). Line geometries for sequence comparisons. *Bulletin of Mathematical Biology*, 46(4):567--577.

Watkins, X.; Garcia, L. J.; Pundir, S.; Martin, M. J. & UniProt Consortium (2017). ProtVista: visualization of protein sequence annotations. *Bioinformatics (Oxford, England)*, 33(13):2040--2041. ISSN 1367-4803.

Watts, D. J. (2004). *Six degrees: The science of a connected age*. WW Norton & Company.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of "small-world"networks. *Nature*, 393(6684):440.

Wei, J.; Li, S.-J.; Shi, H.; Wang, H.-Y.; Rong, C.-T.; Zhu, P.; Jin, S.-H.; Liu, J. & Li, J.-Y. (2013). Characterisation of lyzls in mice and antibacterial properties of human lyzl6. *Asian journal of andrology*, 15(6):824.

Yachdav, G.; Wilzbach, S.; Rauscher, B.; Sheridan, R.; Sillitoe, I.; Procter, J.; Lewis, S. E.; Rost, B. & Goldberg, T. (2016). Msaviewer: interactive javascript visualization of multiple sequence alignments. *Bioinformatics*, 32(22):3501--3503.

Yang, T.; Snider, B. B. & Oprian, D. D. (1997). Synthesis and characterization of a novel retinylamine analog inhibitor of constitutively active rhodopsin mutants found in patients with autosomal dominant retinitis pigmentosa. *Proceedings of the National Academy of Sciences*, 94(25):13559--13564.

Zanotti, G.; Cendron, L.; Ramazzina, I.; Folli, C.; Percudani, R. & Berni, R. (2006). Structure of zebra fish hiuase: insights into evolution of an enzyme to a hormone transporter. *Journal of molecular biology*, 363(1):1--9.

Zhang, K.; Gao, R.; Zhang, H.; Cai, X.; Shen, C.; Wu, C.; Zhao, S. & Yu, L. (2005). Molecular cloning and characterization of three novel lysozyme-like genes, predominantly expressed in the male reproductive system of humans, belonging to the c-type lysozyme/alpha-lactalbumin family. *Biology of reproduction*, 73(5):1064--1071.

Zucchi, R.; Chiellini, G.; Scanlan, T. & Grandy, D. (2006). Trace amine-associated receptors and their ligands. *British journal of pharmacology*, 149(8):967--978.

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

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

Zvelebil, M. J.; Barton, G. J.; Taylor, W. R. & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of molecular biology*, 195(4):957--961.

Apêndice A

Material Suplementar

A.0.1 Lisozimas de tipo C/Alfa-lactoalbuminas

	Community 8					Community 21		
Sequence	**Aromatic116**	**Very Small164**	**D103**	**E81**	**Non-Polar218**	**Pos. Charged128**	**N71**	**G126**
LALB1_HORSE	M59	I72	E49	S33	Y103	Q65	E25	N64
LALBA_BOVIN	I78	I91	E68	T52	Y122	Q84	E44	D83
LALBA_HUMAN	L78	I91	E68	T52	Y122	Q84	E44	S83
LALBA_MOUSE	F79	I92	E69	T53	Y123	E85	E45	S84
LALBA_PIG	L78	I90	E68	I52	Y121	Q84	E44	N83
LALBA_RABIT	L78	I91	E68	T52	H122	Q84	E44	K83
LALBA_RAT	N78	I91	E68	T52	Y122	E84	E44	S83
LALBA_SHEEP	I78	I91	E68	T52	Y122	Q84	E44	D83
LYSC1_HORSE	W63	A75	D53	E35	A107	K69	N27	N68
LYSC1_PIG	Y61	A74	D51	E35	A106	K67	N27	G66
LYSC1_RAT	Y81	A94	D71	E53	A126	K87	D45	G86
LYSC1_SHEEP	W81	G94	D71	E53	A125	K87	N45	G86
LYSC2_BOVIN	W81	G94	D71	E53	A125	K87	N45	G86
LYSC2_PIG	Y79	A92	D69	E53	A124	-	N26	-
LYSC2_RAT	Y81	A94	D71	E53	A126	K85	N45	G84
LYSC3_BOVIN	W81	G94	D71	E53	A125	K87	N45	G86
LYSC3_PIG	Y81	A94	D71	E53	A126	K87	N45	G86
LYSC3_SHEEP	W81	G94	D71	E53	A125	K87	N45	G86
LYSC_BOVIN	W81	G94	D71	E53	A125	K87	N45	G86
LYSC_CHICK	W80	L93	D70	E53	A125	K87	N45	G86
LYSC_HUMAN	Y81	A94	D71	E53	A126	R86	N45	G85
LYSCK_SHEEP	W63	A76	D53	E35	A108	K87	N45	G86
LYSCN_BOVIN	W81	A94	D71	E53	A126	K69	N27	G68
LYSCT_BOVIN	W81	G94	D71	E53	A125	K87	N45	G86
SACA3_BOVIN	K97	L109	N87	A70	S141	K87	N45	G86
SACA3_CANLF	K97	V109	N87	T70	S141	N103	D62	L102
SACA3_HUMAN	R149	V161	N139	T122	Y193	-	D114	L154
SACA3_MOUSE	R155	L167	N145	T128	Y199	A161	D120	L160
SACA3_PANTR	R149	V161	N139	T122	Y193	-	D114	L154

Figura A.1: Comunidades especificas de Lisozimas C

	Community 10				Community 19		
Sequence	**E103**	**E71**	**V91**	**QE40**	**D181**	**D186**	**D187**
LALB1_HORSE	E49	E25	V42	Q2	D82	D87	D88
LALBA_BOVIN	E68	E44	V61	Q21	D101	D106	D107
LALBA_HUMAN	E68	E44	V61	Q21	D101	D106	D107
LALBA_MOUSE	E69	E45	V62	E22	D102	D107	D108
LALBA_PIG	E68	E44	V61	Q21	D100	D105	D106
LALBA_RABIT	E68	E44	V61	Q21	D101	D106	D107
LALBA_RAT	E68	E44	V61	E21	D101	D106	D107
LALBA_SHEEP	E68	E44	V61	Q21	D101	D106	D107
LYSC1_HORSE	D53	N27	N44	V2	D85	D90	D91
LYSC1_PIG	D51	N27	N44	V2	D84	Q89	D90
LYSC1_RAT	D71	D45	N62	I20	Q104	Q109	A110
LYSC1_SHEEP	D71	N45	N62	V20	E104	K109	A110
LYSC2_BOVIN	D71	N45	N62	V20	E104	K109	A110
LYSC2_HORSE	-	N26	-	V2	-	-	-
LYSC2_PIG	D69	N45	N62	V20	D102	Q107	D108
LYSC2_RAT	D71	N45	N62	V20	Q104	Q109	A110
LYSC3_BOVIN	D71	N45	N62	V20	E104	K109	A110
LYSC3_PIG	D71	N45	N62	V20	D104	Q109	D110
LYSC3_SHEEP	D71	N45	N62	V20	E104	K109	A110
LYSC_BOVIN	D71	N45	N62	V20	E104	K109	A110
LYSC_CHICK	D70	N45	N62	V20	S103	A108	S109
LYSC_HUMAN	D71	N45	N62	V20	Q104	D109	A110
LYSCK_SHEEP	D53	N27	N44	V2	Q86	Q91	A92
LYSCN_BOVIN	D71	N45	N62	V20	Q104	Q109	A110
LYSCT_BOVIN	D71	N45	N62	T20	K104	Q109	A110
SACA3_BOVIN	N87	D62	D79	V37	N119	D124	T125
SACA3_CANLF	N87	D62	D79	V37	N119	D124	T125
SACA3_HUMAN	N139	D114	D131	L89	N171	D176	T177
SACA3_MOUSE	N145	D120	D137	V95	N177	D182	S183
SACA3_PANTR	N139	D114	D131	L89	N171	D176	T177

Figura A.2: Comunidades especificas de Alfa-lactoalbuminas

	Community 18		
Sequence	**E93**	**M192**	**N104**
LALB1_HORSE	N44	K93	Y50
LALBA_BOVIN	N63	K112	Y69
LALBA_HUMAN	N63	K112	Y69
LALBA_MOUSE	D64	K113	Y70
LALBA_PIG	D63	K111	Y69
LALBA_RABIT	N63	M112	Y69
LALBA_RAT	N63	K112	Y69
LALBA_SHEEP	N63	K112	Y69
LYSC1_HORSE	K46	K96	Y54
LYSC1_PIG	N46	K95	Y52
LYSC1_RAT	N64	K115	Y72
LYSC1_SHEEP	N64	K115	Y72
LYSC2_BOVIN	N64	K115	Y72
LYSC2_HORSE	-	-	-
LYSC2_PIG	N64	K113	Y70
LYSC2_RAT	N64	K115	Y72
LYSC3_BOVIN	N64	K115	Y72
LYSC3_PIG	N64	K115	Y72
LYSC3_SHEEP	N64	K115	Y72
LYSC_BOVIN	N64	K115	Y72
LYSC_CHICK	N64	K114	Y71
LYSC_HUMAN	N64	K115	Y72
LYSCK_SHEEP	N46	K97	Y54
LYSCN_BOVIN	N64	K115	Y72
LYSCT_BOVIN	N64	K115	Y72
SACA3_BOVIN	E81	M130	S88
SACA3_CANLF	E81	M130	N88
SACA3_HUMAN	E133	M182	N140
SACA3_MOUSE	E139	M188	N146
SACA3_PANTR	E133	M182	N140

Figura A.3: Comunidades especificas das proteínas associadas a membrana do acrossomo do espermatozoide

A.0.2 HIUases e Transtirretinas

	Community 7					
Sequence	Hydrophobic32	Hydroxyl33	H35	Hydrophobic296	Polar327	Pos. Charged328
HIUH1_CAEEL	I27	S28	H30	I111	Y132	R133
HIUH1_RHIME	L11	T12	H14	I99	Y120	R121
HIUH2_CAEEL	I28	S29	H31	I112	Y133	R134
HIUH2_RHIME	L7	T8	H10	I96	Y117	R118
HIUH_BACHD	V5	T6	H8	L95	Y116	R117
HIUH_BACSU	L4	T5	H7	L90	Y111	R112
HIUH_CAUCR	L4	T5	H7	V91	Y112	R113
HIUH_DANRE	L28	S29	H31	I114	Y135	R136
HIUH_DEIRA	L7	T8	H10	V95	Y116	R117
HIUH_ECOLI	L29	S30	H32	I113	Y134	R135
HIUH_MOUSE	L8	T9	H11	I94	Y115	R116
HIUH_PSEAE	L13	T14	H16	I102	Y123	R124
HIUH_RALSO	L4	T5	H7	V93	Y114	R115
HIUH_RHILO	L12	T13	H15	M101	Y121	R122
HIUH_SALTY	L28	S29	H31	I112	Y133	R134
HIUH_SCHPO	L13	T14	H16	I101	Y121	R122
TTHL_ARATH	I206	T207	H209	V298	Y321	R322
TTHY_BOVIN	-	-	K35	A117	-	-
TTHY_CHICK	-	-	K38	A120	-	-
TTHY_HUMAN	-	-	K35	A117	-	-
TTHY_MONDO	-	-	K37	A119	-	-
TTHY_MOUSE	-	-	K35	A117	-	-
TTHY_PANTR	-	-	K35	A117	-	-
TTHY_PIG	-	-	K35	A117	-	-
TTHY_PONAB	-	-	K35	A117	-	-
TTHY_RABIT	-	-	K15	A97	-	-
TTHY_RAT	-	-	K35	A117	-	-
TTHY_SHEEP	-	-	K35	A117	-	-
TTHY_XENTR	-	-	K38	A119	-	-

Figura A.4: Comunidades especificas de HIUases

	Community 2		Community 3		Community 6											Community 8	
Sequence	V34	V223	T129	W233	L30	M31	K35	F107	E120	H123	H269	A296	G303	A315	T330	K222	E273
HIUH1_CAEEL	A29	L85	P74	Y93	-	-	H30	I58	R68	D70	Y102	I111	-	P121	-	R84	E106
HIUH1_RHIME	T13	L70	L58	L78	-	-	H14	L42	R52	D54	L90	I99	-	P109	-	E69	P94
HIUH2_CAEEL	A30	L86	P75	Y94	-	-	H31	I59	R69	D71	Y103	I112	-	P122	-	R85	E107
HIUH2_RHIME	T9	L66	M54	F74	-	-	H10	I38	R48	D50	L87	I96	-	P106	-	E65	P91
HIUH_BACHD	T7	L65	L53	F73	-	-	H8	L37	R47	D49	L86	L95	-	P105	-	E64	P90
HIUH_BACSU	T6	M60	L48	F68	-	-	H7	M32	R42	D44	L81	L90	-	P100	-	V59	T85
HIUH_CAUCR	T6	L61	-	F69	-	-	H7	L35	R45	R47	L82	V91	-	P101	-	R60	V86
HIUH_DANRE	T30	M88	T77	W96	-	-	H31	L61	R71	P73	Y105	I114	-	P124	-	K87	E109
HIUH_DEIRA	T9	L66	I54	F74	-	-	H10	V38	R48	D50	L86	V95	-	P105	-	E65	T90
HIUH_ECOLI	V31	V87	P76	F95	-	-	H32	L60	R70	K72	F104	I113	-	P123	-	R86	P108
HIUH_MOUSE	T10	L68	T57	W76	-	-	H11	L41	R51	P53	Y85	I94	-	P104	-	K67	E89
HIUH_PSEAE	T15	L72	L60	Y80	-	-	H16	I44	R54	D56	L93	I102	-	P112	-	Q71	V97
HIUH_RALSO	T6	L63	L51	Y71	-	-	H7	I35	R45	E47	L84	V93	-	P103	-	E62	T88
HIUH_RHILO	T14	L71	L59	L79	-	-	H15	L43	R53	D55	L92	M101	-	P110	-	E70	P96
HIUH_SALTY	V30	V86	P75	F94	-	-	H31	L59	R69	K71	F103	I112	-	P122	-	R85	P107
HIUH_SCHPO	A15	F75	V63	F83	-	-	H16	I47	R57	T59	Y92	I101	-	P110	-	T74	E96
TTHL_ARATH	T208	I276	D264	S284	-	-	H209	V248	R258	G260	F289	V298	Q302	P310	-	R275	S293
TTHY_BOVIN	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_CHICK	V37	V94	T83	W102	L35	M36	K38	F67	E77	H79	H111	A120	G124	A131	T142	R93	D115
TTHY_HUMAN	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_MONDO	V36	V93	N82	W101	L34	M35	K37	F66	E76	H78	H110	A119	G123	A130	T141	K92	D114
TTHY_MOUSE	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	R90	D112
TTHY_PANTR	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_PIG	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_PONAB	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_RABIT	V14	V71	T60	W79	L12	M13	K15	F44	E54	H56	H88	A97	G101	A108	T119	K70	E92
TTHY_RAT	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	R90	E112
TTHY_SHEEP	V34	V91	T80	W99	L32	M33	K35	F64	E74	H76	H108	A117	G121	A128	T139	K90	E112
TTHY_XENTR	V37	I93	T82	W101	L35	M36	K38	I66	E76	H78	H110	A119	G123	A130	S140	K92	D114

Figura A.5: Comunidades especificas de Transtirretinas

A.0.3 Amidases

	Community 1														Community 3								
Sequence	**T3044**	**R2390**	**Y2289**	**G2387**	**E2560**	**L2667**	**D3023**	**Q1564**	**Y2725**	**Y2714**	**Y2391**	**R2376**	**D2384**	**D2737**	**L2366**	**G2532**	**D2991**	**F2538**	**A1544**	**S2352**	**Y2764**	**S1546**	**T2641**
DUR1_YEAST	T410	-	E314	-	P340	Y352	G407	C219	-	D356	-	-	-	-	-	E336	-	S337	C214	Y322	-	K215	A349
FAAH1_HUMAN	P503	D395	T377	K393	L421	L433	L500	S279	S438	K437	F396	Q390	F392	R439	F388	L417	S492	L418	E274	G385	A441	A275	S430
FAAH1_MOUSE	P503	D395	A377	K393	L421	L433	L500	S279	P438	C437	F396	Q390	F392	R439	F388	L417	S492	L418	T274	G385	A441	A275	A430
FAAH1_PIG	P503	D395	T377	E393	M421	L433	L500	S279	P438	R437	Y396	Q390	F392	R439	L388	L417	S492	L418	V274	G385	A441	A275	A430
FAAH2_HUMAN	T474	D379	I361	F377	I406	L418	-	C268	S421	Y420	L380	P374	K376	N422	K372	S402	-	V403	Q263	H369	K424	E264	E415
FAAH1_CAEEL	G501	D393	F375	L391	P419	M431	L498	N276	R436	T435	I394	N388	L390	D437	V386	I415	S490	L416	P271	G383	F439	L272	G428
AMDS_ASPOR	T474	A387	N370	I385	V412	-	L471	V267	-	-	S388	Y382	H384	-	-	D408	-	L409	E262	S378	-	T263	R421
AMDS_EMENI	T473	S387	S369	S385	L411	-	L470	V266	-	-	-	R382	I384	-	V380	N407	-	M408	E261	S377	-	T262	N420
HYIN1_AGRVS	V404	H309	L291	S307	L330	E342	G401	T207	-	-	-	A304	L306	-	L302	Q326	N393	V327	S202	P299	-	P203	L339
HYIN_BRADU	I461	-	C365	-	A391	L402	W458	A272	-	-	-	-	-	-	-	D387	A450	F388	A267	M373	A406	H268	N400
HYIN1_AGRVS	V404	H309	L291	S307	L330	E342	G401	T207	-	-	-	A304	L306	-	L302	Q326	N393	V327	S202	P299	-	P203	L339
HYIN_BRADU	I461	-	C365	-	A391	L402	W458	A272	-	-	-	-	-	-	-	D387	A450	F388	A267	M373	A406	H268	N400
GATA_YEAST	T421	R304	Y286	G302	E330	L342	D418	T188	F347	A346	Y305	R299	D301	K348	L297	K326	S410	F327	S183	A294	N350	Q184	N339
GATA_ARATH	T471	R372	Y354	G370	E397	L409	D468	V249	Y414	Y413	Y373	R367	D369	D415	L365	G393	D460	F394	A244	S362	Y417	S245	T406
GATA_MAIZE	T476	R377	Y359	G375	E402	L414	D473	V254	Y419	Y418	Y378	R372	D374	D420	L370	G398	D465	F399	A249	S367	Y422	S250	T411
GATA_HUMAN	T449	Q346	H328	G344	V372	L384	D446	V229	Y389	N388	Y347	R341	D343	E390	M339	G368	N438	F369	V224	A336	Y392	N225	N381
GATA_BACSU	T428	R328	Y310	G326	E354	L366	D425	Q212	Y371	Y370	Y329	R323	D325	D372	L321	G350	D417	F351	A207	S318	Y374	S208	T363
GATA_PSEAE	T426	R326	Y308	G324	E352	L364	D423	Q210	Y369	Y368	Y327	R321	D323	D370	L319	G348	D415	F349	A205	S316	Y372	S206	T361

	Community 15			Community 17			Community 21		Community 24		Community 36	
Sequence	**D898**	**R2594**	**S2357**	**S956**	**M3000**	**F2947**	**Y2651**	**A1541**	**E1908**	**S2182**	**G2955**	**E973**
DUR1_YEAST	D132	S344	Q323	L138	N403	T391	K350	P212	N270	D301	Q394	T141
FAAH1_HUMAN	P188	L425	H386	Y194	L496	L479	A431	G272	T332	L364	N482	S197
FAAH1_MOUSE	P188	L425	C386	Y194	L496	L479	A431	G272	T332	L364	N482	S197
FAAH1_PIG	P188	L425	K386	Y194	L496	L479	G431	G272	T332	L364	N482	S197
FAAH2_HUMAN	S177	G410	D370	Y183	P468	P459	E416	G261	H312	K349	H462	S186
FAAH1_CAEEL	P185	V423	Q384	Y191	L494	P477	M429	G269	D331	Y363	D480	S194
AMDS_ASPOR	P176	Q416	T379	C182	V467	V454	-	G260	F326	T358	N457	V185
AMDS_EMENI	P175	H415	A378	C181	V466	V453	-	G259	F325	T357	D456	V184
HYIN1_AGRVS	H120	Q334	R300	I126	T397	K382	I340	P200	S246	D278	G385	D129
HYIN_BRADU	P179	M395	V374	G185	Q454	P441	V401	M265	I321	E353	R444	D188
HYIN1_AGRVS	H120	Q334	R300	I126	T397	K382	I340	P200	S246	D278	G385	D129
HYIN_BRADU	P179	M395	V374	G185	Q454	P441	V401	M265	I321	E353	R444	D188
GATA_YEAST	D98	R334	S295	S104	S414	G398	Y340	A181	E242	S274	R401	V107
GATA_ARATH	D162	R401	S363	S168	M464	Y452	Y407	A242	E310	S342	G455	E171
GATA_MAIZE	D167	R406	S368	S173	M469	Y457	Y412	A247	E315	S347	G460	E176
GATA_HUMAN	D122	R376	S337	S128	S442	V428	F382	P222	E284	S316	L431	T131
GATA_BACSU	D125	R358	A319	S131	M421	F409	F364	A205	E266	S298	G412	E134
GATA_PSEAE	D123	R356	S317	S129	Q419	W407	Y362	A203	E264	S296	G410	Q132

Figura A.6: Comunidades especificas de Glu-tRNA amidotransferases

	Community 22		Community 28	
Sequence	**E964**	**H2952**	**Q150**	**P1961**
DUR1_YEAST	V139	Q393	P47	K278
FAAH1_HUMAN	D195	L481	K109	A341
FAAH1_MOUSE	D195	L481	K109	A341
FAAH1_PIG	D195	L481	R109	A341
FAAH2_HUMAN	E184	H461	P83	D325
FAAH1_CAEEL	N192	H479	E98	A340
AMDS_ASPOR	E183	H456	Q88	P335
AMDS_EMENI	E182	H455	Q87	P334
HYIN1_AGRVS	T127	V384	T36	H255
HYIN_BRADU	F186	P443	P92	A330
HYIN1_AGRVS	T127	V384	T36	H255
HYIN_BRADU	F186	P443	P92	A330
GATA_YEAST	G105	I400	S19	G251
GATA_ARATH	T169	I454	P76	G319
GATA_MAIZE	T174	I459	P78	G324
GATA_HUMAN	G129	Y430	K35	E293
GATA_BACSU	S132	I411	D38	E275
GATA_PSEAE	S130	I409	P37	R273

Figura A.7: Comunidades especificas de Acetamidases

	Community 2										Community 20	
Sequence	**W1703**	**D3028**	**G1238**	**C186**	**G1541**	**L1174**	**D1909**	**S1254**	**N3010**	**P1940**	**D1722**	**R1987**
DUR1_YEAST	D243	T408	I173	A51	P212	V168	I271	V174	R405	G275	E246	V281
FAAH1_HUMAN	F303	D501	G229	-	G272	L224	D333	S230	N498	P338	D306	R344
FAAH1_MOUSE	F303	D501	G229	-	G272	L224	D333	S230	N498	P338	D306	R344
FAAH1_PIG	F303	D501	G229	-	G272	L224	D333	S230	N498	P338	D306	R344
FAAH2_HUMAN	K292	A472	C218	G87	G261	T213	D313	S219	N470	V322	K295	M328
FAAH1_CAEEL	F300	D499	G226	-	G269	L221	D332	S227	N496	P337	D303	R343
AMDS_ASPOR	W291	D472	G217	C92	G260	M212	D327	G218	N469	P332	D294	R338
AMDS_EMENI	W290	D471	G216	C91	G259	I211	D326	G217	N468	P331	D293	R337
HYIN1_AGRVS	T231	I402	L161	A40	P200	S156	H247	M162	T399	L252	-	A258
HYIN_BRADU	R296	G459	I220	A96	M265	A215	P322	G221	V456	V327	W299	A333
HYIN1_AGRVS	T231	I402	L161	A40	P200	S156	H247	M162	T399	L252	-	A258
HYIN_BRADU	R296	G459	I220	A96	M265	A215	P322	G221	V456	V327	W299	A333
GATA_YEAST	K212	V419	L142	-	A181	S137	F243	V143	I416	M248	T215	R254
GATA_ARATH	F273	I469	Q203	C80	A242	A198	T311	C204	A466	V316	T276	S322
GATA_MAIZE	M278	I474	Q208	S82	A247	A203	T316	C209	A471	V321	T281	S327
GATA_HUMAN	R253	I447	T183	A39	P222	A178	Y285	C184	Q444	L290	T256	S296
GATA_BACSU	M236	I426	E166	A42	A205	A161	Y267	V167	A423	V272	T239	E278
GATA_PSEAE	K234	I424	L164	S41	A203	A159	Y265	L165	L421	L270	T237	D276

Figura A.8: Comunidades especificas de amidases de ácido carboxílico

	Community 8						Community 12			Community 27	
Sequence	**C619**	**C1544**	**G1363**	**L956**	**T933**	**R980**	**N3069**	**Y683**	**F3075**	**P988**	**T3263**
DUR1_YEAST	C101	C214	G186	L138	T136	R142	N411	Y106	F412	P144	T437
FAAH1_HUMAN	L156	E274	I242	Y194	F192	N198	V507	G161	P508	L200	Q550
FAAH1_MOUSE	L156	T274	I242	Y194	L192	N198	V507	G161	P508	L200	Q550
FAAH1_PIG	L156	V274	I242	Y194	F192	N198	V507	G161	P508	L200	Q550
FAAH2_HUMAN	L145	Q263	I231	Y183	M181	N187	G475	D150	V476	I189	Q501
FAAH1_CAEEL	F153	P271	V239	Y191	L189	N195	V505	Y158	P506	L197	Q547
AMDS_ASPOR	Y144	E262	I230	C182	M180	N186	V478	G149	P479	I188	Q521
AMDS_EMENI	Y143	E261	I229	C181	M179	N185	V477	N148	P478	I187	Q520
HYIN1_AGRVS	T88	S202	V174	I126	F124	N130	D408	R93	P409	T132	Q434
HYIN_BRADU	V147	A267	L233	G185	F183	N189	S465	D152	A466	L191	Q490
HYIN1_AGRVS	T88	S202	V174	I126	F124	N130	D408	R93	P409	T132	Q434
HYIN_BRADU	V147	A267	L233	G185	F183	N189	S465	D152	A466	L191	Q490
GATA_YEAST	S66	S183	V155	S104	M102	H108	V422	N71	P423	I110	Q445
GATA_ARATH	S130	A244	V216	S168	M166	A172	V472	H135	N473	A174	Q501
GATA_MAIZE	S135	A249	V221	S173	M171	G177	V477	G140	N478	A179	Q506
GATA_HUMAN	S90	V224	T196	S128	M126	D132	Q450	G95	A451	V134	Q476
GATA_BACSU	S93	A207	I179	S131	M129	N135	I429	N98	P430	A137	Q454
GATA_PSEAE	S91	A205	I177	S129	M127	S133	I427	N96	T428	H135	Q452

Figura A.9: Comunidades especificas de amidases de ureia

A.0.4 Receptores acoplados a proteína G

	Community 4		
Sequence	Neg. Charged827	W3030	G3055
5HT1A_HUMAN	D116	W387	G389
ACM1_HUMAN	D105	W405	C407
ADA1A_HUMAN	D106	W313	G315
ADRB1_HUMAN	D138	W364	G366
DRD1_HUMAN	D103	W318	G320
HRH1_HUMAN	D107	W455	G457
TAAR1_HUMAN	D103	W291	G293
AGTR1_HUMAN	V108	C289	A291
APJ_HUMAN	I109	C296	S298
BKRB1_HUMAN	I117	F299	A301
EDNRA_HUMAN	Q165	L349	A358
GALR1_HUMAN	F115	C290	A292
GNRHR_HUMAN	K121	L310	A312
KISSR_HUMAN	Q122	C310	S312
MCHR1_HUMAN	D192	S367	G369
MSHR_HUMAN	T124	A285	I287
NMUR1_HUMAN	F141	I343	F345
NPFF1_HUMAN	Q123	W316	A318
NPSR1_HUMAN	Q128	N317	P319
NPBW1_HUMAN	D116	S294	S296
NPY1R_HUMAN	Q120	L307	A309
NTR1_HUMAN	R148	A351	F353
OPRM_HUMAN	D149	A325	G327
OX1R_HUMAN	Q126	W345	V347
PAR1_HUMAN	F182	C358	S360
SSR1_HUMAN	D137	I310	G312
CCR1_HUMAN	Y113	V288	A290
ACKR2_HUMAN	Y124	S299	A301
CX3C1_HUMAN	F109	T280	A282
CXCR1_HUMAN	K117	I292	G294
FSHR_HUMAN	T449	L613	H615
LSHR_HUMAN	T446	L610	Y612
TSHR_HUMAN	T501	L665	Y667
PKR1_HUMAN	R144	C329	A331
FFAR1_HUMAN	H86	-	-
LT4R1_HUMAN	C97	A272	A274
LPAR1_HUMAN	I128	L298	A300
CNR1_HUMAN	V196	M384	C386
PTAFR_HUMAN	F97	C280	L282
PD2R_HUMAN	M112	R310	L312
PE2R1_HUMAN	M117	R338	A340
PE2R2_HUMAN	M116	R302	L304
PE2R3_HUMAN	M137	R333	A335
PE2R4_HUMAN	L99	R316	A318
PF2R_HUMAN	M115	R291	A293
PI2R_HUMAN	M99	R279	Y281
MTR1A_HUMAN	M107	Y282	A284
AA2AR_HUMAN	V84	V275	S277
P2RY1_HUMAN	F131	G311	A313
GPBAR_HUMAN	P92	-	-
GPER1_HUMAN	L137	L311	A313
HCAR1_HUMAN	L95	S265	T267
OXGR1_HUMAN	F113	P289	A291
SUCR1_HUMAN	L102	P282	A284
OPN3_HUMAN	G121	L296	A298
OPN4_HUMAN	G150	V337	A339
OPN5_HUMAN	G113	L293	A295
OPSB_HUMAN	G114	F290	S292
OPSD_HUMAN	A117	F293	A295
OPSG3_HUMAN	V133	F309	A311
OPSG_HUMAN	V133	F309	A311
OPSR_HUMAN	V133	Y309	A311

Figura A.10: Comunidades especificas de GPCR's aminérgicas

	Community 9	
Sequence	**Hydrophilic3025**	**K3059**
5HT1A_HUMAN	I385	Y390
ACM1_HUMAN	G403	Y408
ADA1A_HUMAN	V311	Y316
ADRB1_HUMAN	F362	Y367
DRD1_HUMAN	F316	W321
HRH1_HUMAN	T453	Y458
TAAR1_HUMAN	L289	Y294
AGTR1_HUMAN	T287	Y292
APJ_HUMAN	C294	Y299
BKRB1_HUMAN	A297	F302
EDNRA_HUMAN	-	T359
GALR1_HUMAN	A288	Y293
GNRHR_HUMAN	F308	F313
KISSR_HUMAN	A308	Y313
MCHR1_HUMAN	A365	Y370
MSHR_HUMAN	F283	I288
NMUR1_HUMAN	S341	Y346
NPFF1_HUMAN	A314	F319
NPSR1_HUMAN	I315	A320
NPBW1_HUMAN	I292	Y297
NPY1R_HUMAN	C305	M310
NTR1_HUMAN	T349	Y354
OPRM_HUMAN	C323	Y328
OX1R_HUMAN	S343	Y348
PAR1_HUMAN	C356	S361
SSR1_HUMAN	S308	Y313
CCR1_HUMAN	T286	Y291
ACKR2_HUMAN	T297	F302
CX3C1_HUMAN	T278	F283
CXCR1_HUMAN	T290	F295
FSHR_HUMAN	L611	P616
LSHR_HUMAN	L608	P613
TSHR_HUMAN	L663	P668
PKR1_HUMAN	V327	M332
FFAR1_HUMAN	-	-
LT4R1_HUMAN	L270	F275
LPAR1_HUMAN	F296	E301
CNR1_HUMAN	C382	L387
PTAFR_HUMAN	T278	S283
PD2R_HUMAN	A308	S313
PE2R1_HUMAN	A336	S341
PE2R2_HUMAN	A300	S305
PE2R3_HUMAN	A331	S336
PE2R4_HUMAN	A314	S319
PF2R_HUMAN	A289	T294
PI2R_HUMAN	A277	A282
MTR1A_HUMAN	S280	Y285
AA2AR_HUMAN	A273	H278
P2RY1_HUMAN	T309	S314
GPBAR_HUMAN	-	-
GPER1_HUMAN	V309	F314
HCAR1_HUMAN	T263	Y268
OXGR1_HUMAN	S287	A292
SUCR1_HUMAN	T280	F285
OPN3_HUMAN	S294	K299
OPN4_HUMAN	P335	K340
OPN5_HUMAN	P291	K296
OPSB_HUMAN	P288	K293
OPSD_HUMAN	P291	K296
OPSG3_HUMAN	P307	K312
OPSG_HUMAN	P307	K312
OPSR_HUMAN	P307	K312

Figura A.11: Comunidades especificas de GPCR's sensoriais

	Community 7		
Sequence	**G612**	**W1324**	**R3030**
5HT1A_HUMAN	V85	A186	W387
ACM1_HUMAN	I74	Q177	W405
ADA1A_HUMAN	L75	I175	W313
ADRB1_HUMAN	M107	C215	W364
DRD1_HUMAN	V73	G174	W318
HRH1_HUMAN	V76	K179	W455
TAAR1_HUMAN	L72	G181	W291
AGTR1_HUMAN	F77	V179	C289
APJ_HUMAN	F78	Q180	C296
BKRB1_HUMAN	F86	A188	F299
EDNRA_HUMAN	Y129	T238	L349
GALR1_HUMAN	Y84	F186	C290
GNRHR_HUMAN	E90	H199	L310
KISSR_HUMAN	F91	Y190	C310
MCHR1_HUMAN	F161	G262	S367
MSHR_HUMAN	V87	-	A285
NMUR1_HUMAN	V109	V218	I343
NPFF1_HUMAN	V92	S202	W316
NPSR1_HUMAN	T97	Q196	N317
NPBW1_HUMAN	F86	Q187	S294
NPY1R_HUMAN	V89	V197	L307
NTR1_HUMAN	T115	V223	A351
OPRM_HUMAN	A119	D218	A325
OX1R_HUMAN	V95	V201	W345
PAR1_HUMAN	F151	T253	C358
SSR1_HUMAN	L107	A207	I310
CCR1_HUMAN	F83	T182	V288
ACKR2_HUMAN	F95	N194	S299
CX3C1_HUMAN	F80	E174	T280
CXCR1_HUMAN	F88	V186	I292
FSHR_HUMAN	I411	I516	L613
LSHR_HUMAN	M408	I513	L610
TSHR_HUMAN	M463	I568	L665
PKR1_HUMAN	V111	F216	C329
FFAR1_HUMAN	L55	V169	-
LT4R1_HUMAN	V67	L167	A272
LPAR1_HUMAN	A98	N194	L298
CNR1_HUMAN	G166	V263	M384
PTAFR_HUMAN	F66	R172	C280
PD2R_HUMAN	G75	W182	R310
PE2R1_HUMAN	G87	W187	R338
PE2R2_HUMAN	G81	W186	R302
PE2R3_HUMAN	G102	W207	R333
PE2R4_HUMAN	G68	W169	R316
PF2R_HUMAN	G80	W185	R291
PI2R_HUMAN	G63	W169	R279
MTR1A_HUMAN	V76	S176	Y282
AA2AR_HUMAN	V55	A165	V275
P2RY1_HUMAN	Y100	T201	G311
GPBAR_HUMAN	T64	N154	-
GPER1_HUMAN	L108	F206	L311
HCAR1_HUMAN	L64	E166	S265
OXGR1_HUMAN	Y82	A182	P289
SUCR1_HUMAN	F72	T171	P282
OPN3_HUMAN	V90	G187	L296
OPN4_HUMAN	M119	S220	V337
OPN5_HUMAN	I82	S182	L293
OPSB_HUMAN	L83	S183	F290
OPSD_HUMAN	M86	S186	F293
OPSG3_HUMAN	E102	S202	F309
OPSG_HUMAN	E102	S202	F309
OPSR_HUMAN	E102	S202	Y309

Figura A.12: Comunidades especificas de GPCR's de prostanoides

	Community 10			
Sequence	**Y633**	**E883**	**D2832**	**P3059**
5HT1A_HUMAN	L88	T121	F354	Y390
ACM1_HUMAN	F77	N110	F374	Y408
ADA1A_HUMAN	T78	T111	F281	Y316
ADRB1_HUMAN	L110	T143	F333	Y367
DRD1_HUMAN	L76	T108	F281	W321
HRH1_HUMAN	V79	T112	F424	Y458
TAAR1_HUMAN	L75	S108	F260	Y294
AGTR1_HUMAN	-	Y113	F249	Y292
APJ_HUMAN	-	Y114	F257	Y299
BKRB1_HUMAN	-	F122	F259	F302
EDNRA_HUMAN	I132	G170	F315	T359
GALR1_HUMAN	F87	L120	F256	Y293
GNRHR_HUMAN	I93	Y126	F276	F313
KISSR_HUMAN	C94	Q127	F272	Y313
MCHR1_HUMAN	-	F197	F334	Y370
MSHR_HUMAN	S90	L129	F250	I288
NMUR1_HUMAN	V112	L146	F306	Y346
NPFF1_HUMAN	F95	S128	F279	F319
NPSR1_HUMAN	V100	Y133	F283	A320
NPBW1_HUMAN	-	F121	C260	Y297
NPY1R_HUMAN	M92	T125	F272	M310
NTR1_HUMAN	L118	Y153	F312	Y354
OPRM_HUMAN	-	F154	F291	Y328
OX1R_HUMAN	I98	S131	F307	Y348
PAR1_HUMAN	-	Y187	F322	S361
SSR1_HUMAN	-	F142	F280	Y313
CCR1_HUMAN	-	Y118	F248	Y291
ACKR2_HUMAN	-	Y129	F259	F302
CX3C1_HUMAN	-	F114	F240	F283
CXCR1_HUMAN	-	Y122	F251	F295
FSHR_HUMAN	Y414	E454	D581	P616
LSHR_HUMAN	Y411	E451	D578	P613
TSHR_HUMAN	Y466	E506	D633	P668
PKR1_HUMAN	V114	Y149	Y293	M332
FFAR1_HUMAN	-	Y91	L233	-
LT4R1_HUMAN	-	Y102	F230	F275
LPAR1_HUMAN	A101	T133	F267	E301
CNR1_HUMAN	I169	T201	L352	L387
PTAFR_HUMAN	-	Y102	F241	S283
PD2R_HUMAN	L78	L117	F274	S313
PE2R1_HUMAN	-	L122	S306	S341
PE2R2_HUMAN	L84	L121	F273	S305
PE2R3_HUMAN	L105	L142	L291	S336
PE2R4_HUMAN	L71	L104	V281	S319
PF2R_HUMAN	I83	L120	S258	T294
PI2R_HUMAN	F66	L104	M248	A282
MTR1A_HUMAN	Y79	I112	F247	Y285
AA2AR_HUMAN	L58	Q89	F242	H278
P2RY1_HUMAN	-	Y136	F269	S314
GPBAR_HUMAN	-	L97	F233	-
GPER1_HUMAN	D111	Y142	F268	F314
HCAR1_HUMAN	-	A100	F229	Y268
OXGR1_HUMAN	-	Y118	F250	A292
SUCR1_HUMAN	-	Y107	F241	F285
OPN3_HUMAN	F93	I126	F264	K299
OPN4_HUMAN	T122	I155	F305	K340
OPN5_HUMAN	V85	C118	F261	K296
OPSB_HUMAN	F86	L119	F258	K293
OPSD_HUMAN	G89	E122	F261	K296
OPSG3_HUMAN	I105	I138	F277	K312
OPSG_HUMAN	I105	I138	F277	K312
OPSR_HUMAN	I105	I138	Y277	K312

Figura A.13: Comunidades especificas de GPCR's de hormônios glicoproteicos

Classe	**Precisão**	**Revocação**	***F1 Score***	**Suporte**
5HT	0,73	0,87	0,79	76
Adenosina	1,00	0,94	0,97	31
Adrenoceptores	0,76	0,97	0,85	67
Aminas Traço	0,00	0,00	0,00	5
Angiotensina	0,59	0,71	0,65	14
Apelina	0,00	0,00	0,00	7
Bombesina	0,36	1,00	0,53	12
Bradicinina	1,00	1,00	1,00	13
Canabinoide	1,00	1,00	1,00	12
Chemerina	0,00	0,00	0,00	5
Colecistocinina	1,00	1,00	1,00	12
Dopamina	0,00	0,00	0,00	32
Endotelina	1,00	1,00	1,00	14
Estrogênio	1,00	1,00	1,00	5
Fator de Ativação de Plaquetas	1,00	0,86	0,92	7
Formilpeptídeo	1,00	0,93	0,97	15
Galanina	0,00	0,00	0,00	9
GPR18, 55 e 119	0,75	0,33	0,46	9
Grelina	1,00	1,00	1,00	6
Histamina	0,69	0,52	0,59	21
Hormona de Concentração de Melanina	0,83	0,62	0,71	8
Hormona Glicoproteica	1,00	1,00	1,00	26
Hormona Liberadora de Tirotropina	1,00	1,00	1,00	6
Leucotrieno	0,75	0,35	0,48	17
Liberador de Gonadotropina	0,90	0,69	0,78	13
Lisofosfolipídeo LPA	0,00	0,00	0,00	20
Lisofosfolipídeo S1P	0,65	0,94	0,77	18
Melanocortina	0,97	1,00	0,99	36
Melatonina	0,67	0,73	0,70	11
Muscarínico	0,87	0,97	0,92	34
Neuromedina U	1,00	1,00	1,00	7
Neuropeptídeo FF	0,00	0,00	0,00	5
Neuropeptídeo W	0,00	0,00	0,00	6
Neuropeptídeo Y	1,00	0,93	0,96	28
Neurotensina	0,00	0,00	0,00	8

Classe	**Precisão**	**Revocação**	***F1 Score***	**Suporte**
Opiáceo	0,83	0,91	0,87	22
Opsinas	1,00	1,00	1,00	78
Orexina	1,00	1,00	1,00	10
P2Y	0,60	0,89	0,71	35
Peptídeo Complementar	0,57	0,85	0,69	27
Procineticina	1,00	1,00	1,00	8
Prostanoide	1,00	0,93	0,96	41
Proteinase Ativada	0,32	0,62	0,43	16
Quimiocinas	0,84	0,98	0,91	100
Relaxina	1,00	0,82	0,90	11
Somatostatina	0,81	0,65	0,72	20
Taquicinina	0,76	0,94	0,84	17
Urotensina	0,00	0,00	0,00	5
Vasopressina e Ocitocina	0,87	1,00	0,93	20
Ácido Biliar	0,83	1,00	0,91	5
Ácido Graxo Livre	0,88	0,54	0,67	13
Ácido Hidroxicarboxílico	1,00	1,00	1,00	6
Média/Total	0,76	0,81	0,77	1049

Tabela A.1: Taxa de acerto da classificação de todas as subclasses de GPCRs presentes no Swiss-Prot. Foi utilizado como fonte de informação os vetores de frequência média de cada sequência para cada comunidade da rede 2.

Sequence	**Label**	**Probability**
TAAR3_RAT	Aminérgico	0.997533387413
TAAR3_MOUSE	Aminérgico	0.997533387413
TAAR4_RAT	Aminérgico	0.996617104667
TAAR4_MOUSE	Aminérgico	0.996532516943
TAAR2_HUMAN	Aminérgico	0.993696874697
TAAR2_MOUSE	Aminérgico	0.993696874697
TAAR2_RAT	Aminérgico	0.993571639625
GP151_MOUSE	Peptídeo	0.985223688724
GP151_RAT	Peptídeo	0.983866534138
GP151_HUMAN	Peptídeo	0.982737275931
TAAR3_HUMAN	Aminérgico	0.960967278346

Sequence	Label	Probability
GP152_MOUSE	Peptídeo	0.960181230207
GP152_HUMAN	Peptídeo	0.960181230207
MRGRF_MOUSE	Peptídeo	0.947058386573
TAAR5_PANTR	Aminérgico	0.942385759924
TAAR5_HUMAN	Aminérgico	0.942385759924
TAAR6_RAT	Aminérgico	0.939949692315
TAAR6_MOUSE	Aminérgico	0.939949692315
TAAR6_PANTR	Aminérgico	0.939949692315
TAAR6_HUMAN	Aminérgico	0.939949692315
TAAR5_MOUSE	Aminérgico	0.939224347215
TAAR5_RAT	Aminérgico	0.939224347215
GPR19_MOUSE	Peptídeo	0.934331513268
GPR19_RAT	Peptídeo	0.934331513268
GPR19_HUMAN	Peptídeo	0.934331513268
TAAR8_HUMAN	Aminérgico	0.929888963099
GPR33_HUMAN	Peptídeo	0.924774171689
GPR33_PANTR	Peptídeo	0.924774171689
GP176_HUMAN	Peptídeo	0.924235048841
GP176_MOUSE	Peptídeo	0.924235048841
GP176_RAT	Peptídeo	0.917784264991
GPR84_HUMAN	Peptídeo	0.917129306496
GPR84_MOUSE	Peptídeo	0.917129306496
GPR1_HUMAN	Peptídeo	0.905191504967
GPR33_MOUSE	Peptídeo	0.902836556708
GPR1_MACMU	Peptídeo	0.877487982463
GP146_HUMAN	Peptídeo	0.875796893273
GPR83_MOUSE	Peptídeo	0.863833496672
GPR83_CANLF	Peptídeo	0.863833496672
GPR83_HUMAN	Peptídeo	0.863833496672
GPR25_HUMAN	Peptídeo	0.860537199729
GPR37_MOUSE	Peptídeo	0.858893898102
GPR37_HUMAN	Peptídeo	0.858893898102
GPR37_RAT	Peptídeo	0.858893898102
TAAR9_RAT	Aminérgico	0.846804374486
MRGRF_HUMAN	Peptídeo	0.833925217828

Sequence	Label	Probability
TAAR9_MOUSE	Aminérgico	0.833157500148
TAAR9_HUMAN	Aminérgico	0.833157500148
GPR25_MOUSE	Peptídeo	0.826117947429
GP183_MOUSE	Peptídeo	0.825588937127
GP183_BOVIN	Peptídeo	0.825588937127
GP183_RAT	Peptídeo	0.825588937127
GP183_HUMAN	Peptídeo	0.825588937127
GP146_MOUSE	Peptídeo	0.819927918564
GP174_MOUSE	Lipídico	0.808811619796
GP174_HUMAN	Lipídico	0.808811619796
GP161_HUMAN	Nucleotídeo	0.807973378692
GP101_MOUSE	Nucleotídeo	0.804562098958
GPR84_BOVIN	Peptídeo	0.804202551712
GP101_HUMAN	Nucleotídeo	0.790616586135
GP148_HUMAN	Peptídeo	0.786075487585
GPR15_MOUSE	Peptídeo	0.78559620701
MRGX3_MACMU	Peptídeo	0.780939866911
GPR78_HUMAN	Aminérgico	0.776661175248
GP161_BOVIN	Nucleotídeo	0.763832838018
GP161_XENTR	Nucleotídeo	0.763832838018
GP161_DANRE	Nucleotídeo	0.763832838018
GP161_MOUSE	Nucleotídeo	0.763832838018
GPR32_HUMAN	Peptídeo	0.760928558222
GPR61_MOUSE	Aminérgico	0.75687384377
GPR34_MOUSE	Lipídico	0.75265400862
GPR34_PANTR	Lipídico	0.75265400862
GPR34_GORGO	Lipídico	0.75265400862
GPR34_HUMAN	Lipídico	0.75265400862
GP139_MOUSE	Peptídeo	0.75145896163
GP139_RAT	Peptídeo	0.75145896163
GP139_HUMAN	Peptídeo	0.75145896163
GP153_HUMAN	Esteróide	0.748402551626
MRGX3_HUMAN	Peptídeo	0.740960691031
GP132_MOUSE	Peptídeo	0.718335439494
GPR22_MOUSE	Lipídico	0.717984294339

Sequence	Label	Probability
GPR22_HUMAN	Lipídico	0.717984294339
GPR61_HUMAN	Aminérgico	0.7101900407
GPR17_RAT	Peptídeo	0.703305882414
GPR17_MOUSE	Peptídeo	0.703305882414
GPR17_HUMAN	Peptídeo	0.699138751887
MAS_RAT	Lipídico	0.689246425888
MAS_HUMAN	Lipídico	0.689246425888
MAS_MOUSE	Lipídico	0.689246425888
GPR12_RAT	Lipídico	0.689128473739
GPR12_MOUSE	Lipídico	0.689128473739
GPR12_HUMAN	Lipídico	0.689128473739
LGR5_MOUSE	Lipídico	0.673143539765
LGR4_DANRE	Peptídeo	0.670608094968
GPR75_HUMAN	Peptídeo	0.668971251339
GPR27_HUMAN	Aminérgico	0.665148854027
GP149_MOUSE	Esteróide	0.661898835987
GPR3_RAT	Lipídico	0.659630940687
GPR3_MOUSE	Lipídico	0.659630940687
GPR3_HUMAN	Lipídico	0.659630940687
GP162_HUMAN	Lipídico	0.656293869642
GP162_MOUSE	Lipídico	0.656293869642
GPR75_MOUSE	Peptídeo	0.647655565905
GPR1_RAT	Peptídeo	0.631899550148
GPR4_PIG	Nucleotídeo	0.620303635806
LGR5_RAT	Lipídico	0.612166970387
GPR1_MOUSE	Peptídeo	0.59275000976
GP150_HUMAN	Peptídeo	0.585956462136
LGR4_RAT	Peptídeo	0.581153717053
MRGX4_HUMAN	Peptídeo	0.574890566062
GPR4_MOUSE	Nucleotídeo	0.570048080364
GPR4_RAT	Nucleotídeo	0.570048080364
GPR4_HUMAN	Nucleotídeo	0.570048080364
GPR4_BOVIN	Nucleotídeo	0.570048080364
GPR88_HUMAN	Aminérgico	0.569902288693
LGR4_BOVIN	Peptídeo	0.569615920181

Sequence	Label	Probability
LGR4_HUMAN	Peptídeo	0.569615920181
GPR88_MOUSE	Aminérgico	0.553564324833
GPR88_RAT	Aminérgico	0.553564324833
MRGX2_PANTR	Peptídeo	0.548118955262
GP182_HUMAN	Proteico	0.545154558609
GP135_HUMAN	Aminérgico	0.530410032101
GPR39_HUMAN	Peptídeo	0.528857787482
GPR15_HUMAN	Peptídeo	0.525975329497
GPR15_PANTR	Peptídeo	0.525975329497
GPR15_MACMU	Peptídeo	0.525975329497
GP153_MOUSE	Esteróide	0.525693525402
GPR21_HUMAN	Peptídeo	0.512803193188
GPR21_MOUSE	Peptídeo	0.512803193188
GPR39_BOVIN	Peptídeo	0.510140959374
GPR39_PIG	Peptídeo	0.510140959374
MRGX2_MACMU	Peptídeo	0.509281749759
MRGX2_HUMAN	Peptídeo	0.509281749759
MRGX2_GORGO	Peptídeo	0.509281749759
GP150_MOUSE	Peptídeo	0.50471518039
GPR20_MOUSE	Peptídeo	0.497113196276
GPR20_HUMAN	Peptídeo	0.497113196276
GPR87_HUMAN	Peptídeo	0.495365348452
GPR87_MOUSE	Peptídeo	0.495365348452
GPR35_HUMAN	Peptídeo	0.495104009142
GP182_RAT	Proteico	0.491034136461
GP182_MOUSE	Proteico	0.491034136461
GP135_RAT	Aminérgico	0.488044021463
GP135_MOUSE	Aminérgico	0.488044021463
MRGRD_RAT	Peptídeo	0.481448252115
GPR35_MOUSE	Peptídeo	0.455714027199
GPR85_DANRE	Aminérgico	0.452508394219
GP132_HUMAN	Peptídeo	0.451457125598
LGR4_MOUSE	Peptídeo	0.437049214976
MRGX1_HUMAN	Peptídeo	0.429089300667
GP171_MOUSE	Lipídico	0.428169681431

Sequence	Label	Probability
GPR39_MOUSE	Peptídeo	0.423264002369
GP141_HUMAN	Proteico	0.41648577227
GP141_MOUSE	Proteico	0.41648577227
GPR52_HUMAN	Nucleotídeo	0.414489485364
GPR52_MOUSE	Nucleotídeo	0.414489485364
GPR52_BOVIN	Nucleotídeo	0.414489485364
GPR85_HUMAN	Aminérgico	0.403606914764
GPR85_PONAB	Aminérgico	0.403606914764
GPR85_RAT	Aminérgico	0.403606914764
GPR85_MOUSE	Aminérgico	0.403606914764
GPR26_RAT	Peptídeo	0.395985951548
GPR26_HUMAN	Peptídeo	0.395985951548
GPR26_MOUSE	Peptídeo	0.395985951548
MRGRD_MOUSE	Peptídeo	0.39138262033
MRGX1_MOUSE	Peptídeo	0.390958949058
GPR63_MOUSE	Peptídeo	0.388734009386
GPR63_HUMAN	Peptídeo	0.388734009386
LGR4_XENTR	Peptídeo	0.386534083519
GP173_DANRE	Peptídeo	0.376036002498
GP173_RAT	Lipídico	0.368553815658
GP173_BOVIN	Lipídico	0.368553815658
GP173_MOUSE	Lipídico	0.368553815658
GP173_HUMAN	Lipídico	0.368553815658
MRGRF_RAT	Peptídeo	0.365813304071
MRGX1_RAT	Peptídeo	0.354578041629
MRGRD_HUMAN	Peptídeo	0.352865809966
GPR31_MOUSE	Peptídeo	0.352699558911
GPR31_HUMAN	Peptídeo	0.34234026613
GPR82_HUMAN	Lipídico	0.341510774355
GPR27_MOUSE	Peptídeo	0.330900814045
GPR27_RAT	Peptídeo	0.330900814045
GP171_HUMAN	Alicarboxílico	0.327996613116
GP171_BOVIN	Alicarboxílico	0.327996613116
GPR82_MOUSE	Lipídico	0.32713106437
GPR45_MOUSE	Aminérgico	0.326502940382

Sequence	Label	Probability
GPR45_HUMAN	Aminérgico	0.326502940382
GPR6_HUMAN	Esteróide	0.325850785528
GPR6_MOUSE	Esteróide	0.325850785528
GPR6_RAT	Esteróide	0.325850785528
GPR62_HUMAN	Aminérgico	0.306593910654
GPR62_MOUSE	Aminérgico	0.289288013498
GP142_MOUSE	Peptídeo	0.249366156799
GP142_HUMAN	Peptídeo	0.219596352415
LGR6_DANRE	Esteróide	0.209333483894

Tabela A.2: Tabela de classificação das GPCRs órfãs.

A.0.5 CONAN

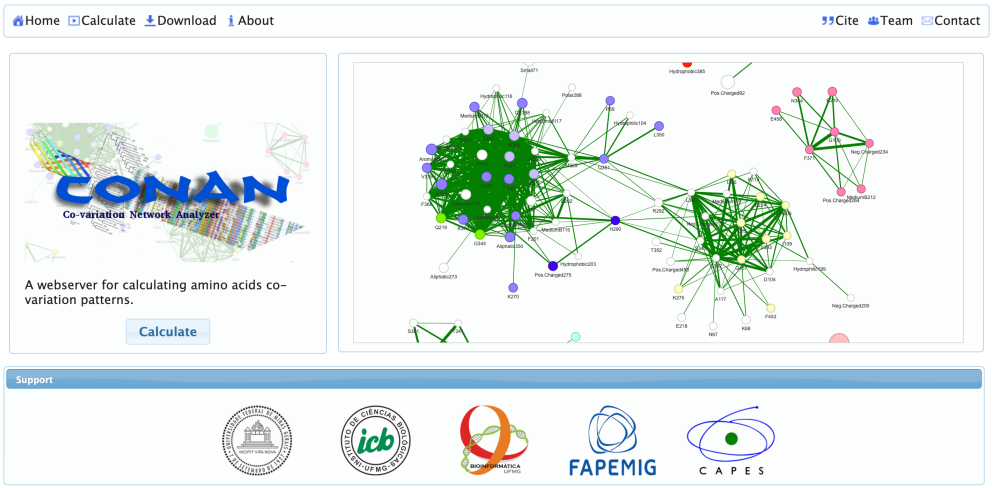


Figura A.14: Página principal do CONAN

A.0.6 CEvADA

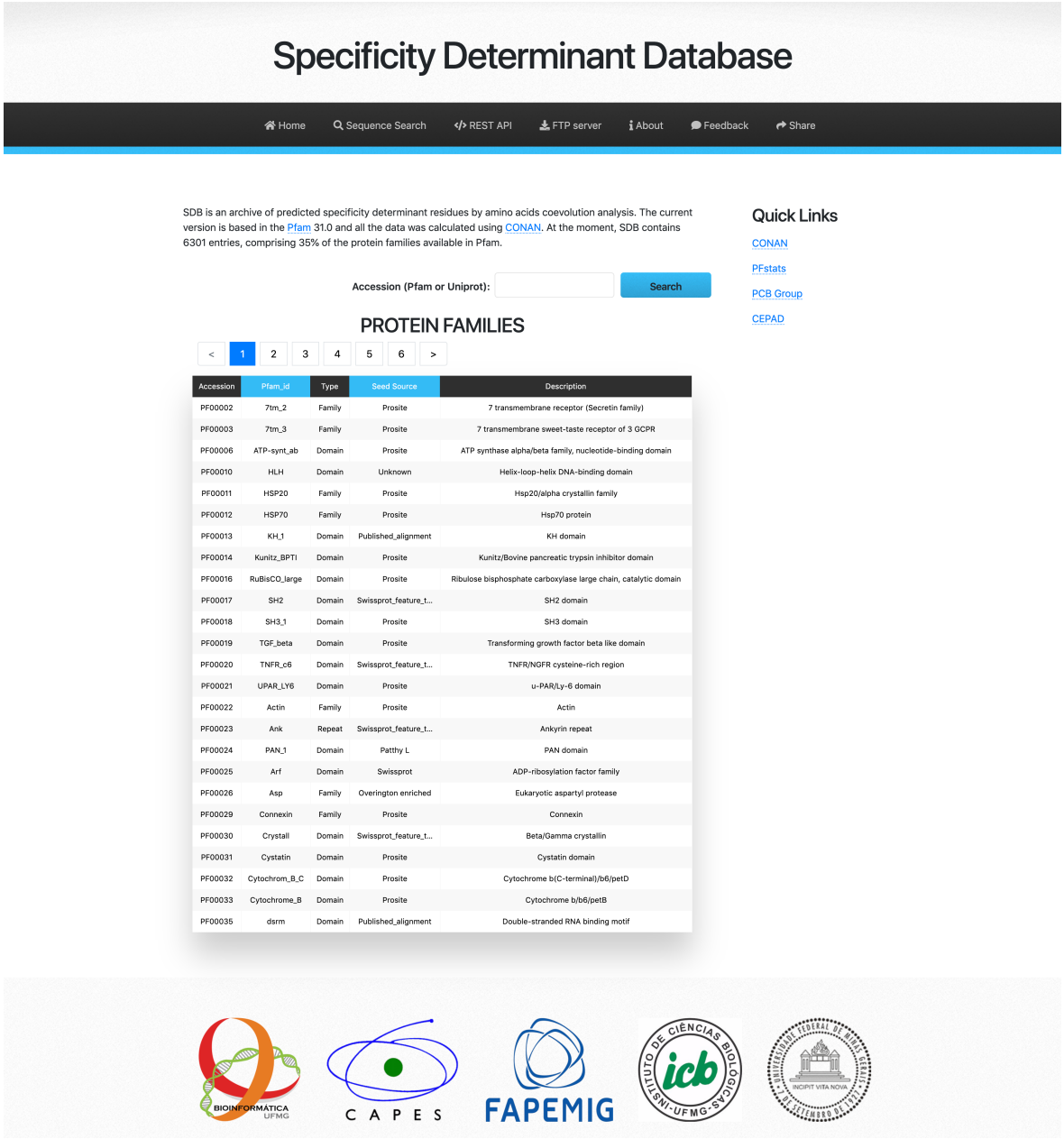


Figura A.15: Página principal do CEvADA

A.0.6.1 Rest API

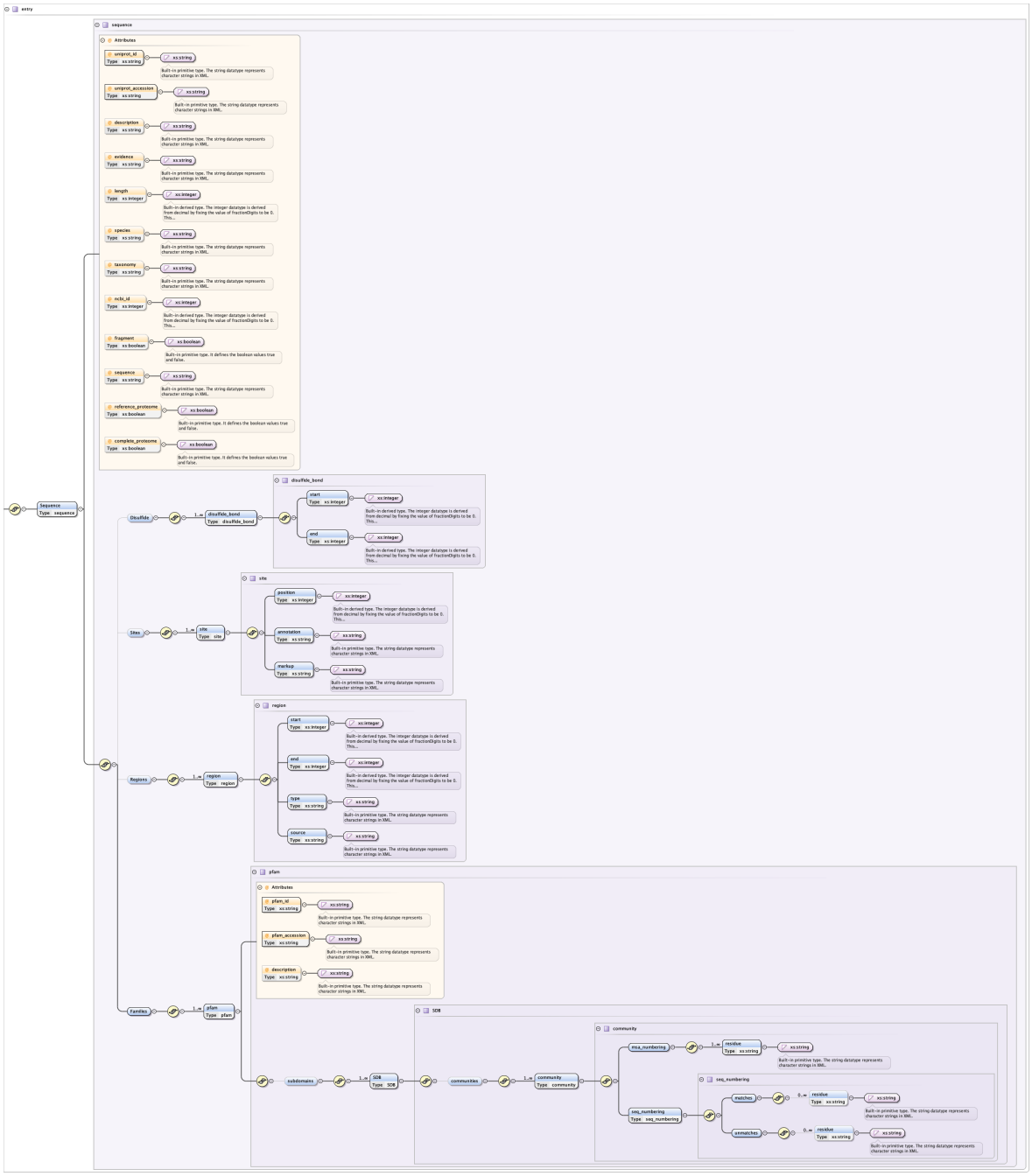


Figura A.16: Ponto final de sequência da REST API do CEvADA

Figura A.17: Ponto final de família da REST API do CEvADA