DETECÇÃO DE SUBFAMÍLIAS PROTEICAS ISOFUNCIONAIS UTILIZANDO INTEGRAÇÃO DE DADOS E AGRUPAMENTO ESPECTRAL

ELISA BOARI DE LIMA

DETECÇÃO DE SUBFAMÍLIAS PROTEICAS ISOFUNCIONAIS UTILIZANDO INTEGRAÇÃO DE DADOS E AGRUPAMENTO ESPECTRAL

Tese apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para a obtenção do grau de Doutora em Bioinformática.

ORIENTADORA: RAQUEL CARDOSO DE MELO-MINARDI

COORIENTADOR: WAGNER MEIRA JÚNIOR

Belo Horizonte

29 de outubro de 2015

Lima, Elisa Boari de.

Detecção de Subfamílias Proteicas Isofuncionais Utilizando Integração de Dados e Agrupamento Espectral / Elisa Boari de Lima. — Belo Horizonte, 2015.
xviii, 176 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais

Orientadora: Melo-Minardi, Raquel Cardoso de.
Coorientador: Meira Júnior, Wagner.

1. Subfamílias Proteicas Isofuncionais. 2. Agrupamento Funcional de Proteínas. 3. Integração de Dados. 4. Mineração de Dados. 5. Agrupamento Espectral. 6. Programação Genética. I. Título.

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"DETECÇÃO DE SUBFAMÍLIAS PROTEICAS ISOFUNCIONAIS UTILIZANDO INTEGRAÇÃO DE DADOS E AGRUPAMENTO ESPECTRAL"

Elisa Boari de Lima

Tese aprovada pela banca examinadora constituída pelos Professores:

Profa Raquel Cardoso de Melo Minardi - Orientadora
UFMG

Prof. Wagner Meira Junior - Co-Orientador
UFMG

Prof. Carlos Henrique da Silveira
Universidade Federal de Itajubá

Prof. Lucas Bleicher
UFMG

Prof. Marcos Augusto dos Santos
UFMG

Profa Cristiane Neri Nobre
PUC/MG

Belo Horizonte, 29 de outubro de 2015.

Aos meus queridos pais, José Maria e Annete, e ao amor da minha vida, Anderson, sem os quais nenhum sucesso seria possível.

Agradecimentos

Agradeço primeiramente a Deus por mais uma oportunidade de crescimento pessoal e profissional, assim como pela força, determinação e persistência para contornar os obstáculos encontrados pelo caminho. Não teria concluído mais esse desafio sem a ajuda de alguns anjos que ele me enviou:

Agradeço aos meus queridos pais, doutores José Maria de Lima e Annete de Jesus Boari Lima, por serem exemplos de conduta, pelos incontáveis ensinamentos, e pela imensa ajuda, amor, apoio e encorajamento de sempre.

Palavras não podem expressar a profunda gratidão que sinto pelo meu amado noivo Anderson Stacanelli Pedroso pelo amor e apoio incondicionais, pela infinita paciência e dedicação, pela motivação e por estar ao meu lado sempre, mesmo a grande distância.

Meu muito obrigada a todos os meus familiares, principalmente meus queridos irmãos, Drielle e Alexandre, pela torcida, incentivo e apoio, e por entenderem a minha ausência em tantas ocasiões.

Meus sinceros agradecimentos à minha orientadora, Dra. Raquel Cardoso de Melo-Minardi, pela inspiração, motivação, atenção e amizade, e ao meu coorientador Dr. Wagner Meira Júnior, pelo suporte e paciência. Obrigada por acreditarem em mim e me aceitarem como orientanda desde o mestrado. Agradeço também aos doutores François Artiguenave e Marcel Salanoubat por me receberem no Genoscope durante o período de doutorado sanduíche na França, e ao Dr. Mohammed Zaki, por acompanhar o meu trabalho e contribuir com ele desde que me recebeu e orientou no Rensselaer Polytechnic Institute durante o mestrado. Faço um agradecimento especial ao Dr. Carlos Henrique da Silveira pelas ideias e discussões valiosas, paciência, disponibilidade e incentivo. Muito obrigada!

Agradeço aos amigos e colegas da Ciência da Computação e da Bioinformática pelas discussões estimulantes, companheirismo e torcida.

Meu agradecimento aos Programas de Pós-Graduação em Ciência da Computação e em Bioinformática da UFMG pela minha formação multidisciplinar. Estendo um agradecimento especial à querida Sheila Santana, pela compreensão, paciência e ajuda na parte burocrática do processo.

Obrigada a todos que de alguma forma contribuíram para o meu sucesso nesta empreitada.

Por fim, meu agradecimento especial à Universidade Federal de Minas Gerais por ser minha segunda casa desde 2009, quando cheguei para o mestrado em Ciência da Computação. Me despeço, agora, Doutora em Bioinformática! *Au revoir!*

"A persistência é o menor caminho do êxito."

(Charles Chaplin)

Resumo

Apesar dos melhores esforços de pesquisa, uma quantidade substancial e crescente de proteínas ainda apresenta função desconhecida. À medida que novos genomas são sequenciados, a grande maioria das proteínas previstas apenas pode ser anotada computacionalmente, devido aos altos custos e dificuldade da investigação experimental. Isso enfatiza a necessidade por métodos computacionais para determinar funções proteicas rápida e confiavelmente. No entanto, não há abordagens de larga escala capazes de revelar as funções de todos os genes hipotéticos nos genomas já sequenciados. Esse objetivo só pode ser alcançado por meio de numerosos esforços de pesquisa, e o presente trabalho é um esforço computacional visando a dar um passo em direção a esse objetivo.

Acredita-se que dividir uma família de proteínas em subtipos de mesma especificidade, que compartilham funções específicas incomuns à família proteica como um todo, seja um primeiro passo para reduzir a complexidade do problema de anotação de funções proteicas. Por isso, o propósito desta tese é a detecção de subfamílias isofuncionais em uma família de proteínas de função desconhecida, além da identificação dos resíduos responsáveis pela diferenciação entre elas. Para tanto, a similaridade entre pares de proteínas em relação a vários tipos de dados é estudada e interpretada como evidência de similaridade funcional. Dados são integrados usando programação genética e, então, fornecidos a um algoritmo de agrupamento espectral, que cria grupos de proteínas similares.

A técnica proposta foi aplicada a famílias proteicas bem conhecidas, assim como a uma família de função desconhecida, e seus resultados foram comparados àqueles obtidos pelo ASMC, uma técnica similar da literatura. Resultados mostraram que a técnica proposta, totalmente automatizada, obteve grupos melhores que o ASMC para Nucleotidil Ciclases e Proteínas Cinases, além de resultados equivalentes para Serino Proteases e para a família DUF849, cujos grupos foram definidos com intervenção manual. Os grupos produzidos pela técnica proposta apresentaram grande correspondência com as subfamílias conhecidas, além de serem mais contrastantes do que aqueles produzidos pelo ASMC. Além disso, para as famílias cujas posições determinantes de especificidade são conhecidas, tais resíduos estavam entre os considerados pela técnica proposta como mais importantes para diferenciar um determinado grupo. Os melhores resultados consistentemente envolveram múltiplos tipos de dados, confirmando a hipótese inicial de que similaridades segundo diferentes domínios do conhecimento podem ser usadas como evidências de similaridade funcional. As principais contribuições desta tese são a estratégia proposta para selecionar e integrar dados, assim como a capacidade de trabalhar com dados ruidosos ou incompletos; o uso de conhecimento de domínio para detectar subfamílias em uma família proteica com diferentes especificidades, reduzindo a complexidade do problema de caracterização funcional; e a identificação de resíduos responsáveis pela especificidade.

Abstract

Despite the best research efforts, a substantial and ever-increasing amount of predicted proteins still lack functional annotation. As increasingly more genomes are sequenced, the vast majority of proteins may only be annotated computationally, given experimental investigation is difficult, expensive, and time-consuming. This highlights the need for computational methods to determine protein functions quickly and reliably. However, no large-scale approaches currently exist capable of revealing the functions of all hypothetical genes in the already sequenced genomes. This goal can only be reached through numerous research efforts, and the work presented herein is a computational effort aiming to take a step toward that goal.

We believe dividing a protein family into same-specificity subtypes, which share specific functions uncommon to the family as a whole, is a first step toward reducing the function annotation problem's complexity. Hence, this work's purpose is to detect isofunctional subfamilies inside a family of unknown function, as well as to identify residues responsible for subfamily differentiation. For this purpose, the similarity between protein pairs according to various data types is studied and interpreted as functional similarity evidence. Data are integrated using genetic programming and, then, provided to a spectral clustering algorithm, which creates clusters of similar proteins.

Four case studies were performed, applying the proposed framework to well-known protein families and to a family of unknown function, and comparing its results to those obtained by ASMC, a similar method found in the literature. Results showed our fully automated technique obtained better clusters than ASMC for the nucleotidyl cyclases and protein kinases families, besides equivalent results for serine proteases and the DUF849 family, for which clusters were defined with manual intervention. Clusters produced by our framework showed great correspondence with the known subfamilies, besides being more contrasting than those produced by ASMC. Additionally, for the families whose specificity determining positions are known, such residues were among those our technique considered most important to differentiate a given group. Best results consistently involved multiple data types, thus confirming our initial hypothesis that similarities according to different knowledge domains may be used as functional similarity evidence. Our main contributions are the proposed strategy for selecting and integrating data types, along with the ability to work with noisy and incomplete data; the use of domain knowledge for detecting isofunctional subfamilies in a protein family with different specificities, thus reducing the complexity of the experimental function characterization problem; and the identification of residues responsible for specificity.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Siglas

ASMC	*Active Sites Modelling and Clustering*.
ATP	Adenosina trifosfato, do inglês *Adenosine TriPhosphate*.
BLAST	*Basic Local Alignment Search Tool*.
BLOSUM	*BLOcks SUbstitution matrix*.
cAMP	Monofosfato cíclico de adenosina, do inglês *cyclic Adenosine MonoPhosphate*.
cGMP	Monofosfato cíclico de guanosina, do inglês *cyclic Guanosine MonoPhosphate*.
CSM	*Cutoff Scanning Matrix*.
DNA	Ácido desoxirribonucleico, do inglês *DeoxyriboNucleic Acid*.
DUF	Domínio de função desconhecida, do inglês *Domain of Unknown Function*.
EC	*Enzyme Commission*.
GA	Algoritmo genético, do inglês *Genetic Algorithm*.
GO	*Gene Ontology*.
GP	Programação genética, do inglês *Genetic Programming*.
GRAVY	Índice de hidropaticidade, do inglês *GRand AVerage of hYdropathicity*.
GTP	Guanosina trifosfato, do inglês *Guanosine TriPhosphate*.
HMM	Modelo oculto de Markov, do inglês *Hidden Markov Model*.
KAH	β-ceto-5-amino-hexanoato.
KNN	K vizinhos mais próximos, do inglês K*-Nearest Neighbors*.
MI	Informação Mútua, do inglês *Mutual Information*.
MSA	Alinhamento múltiplo de sequências, do inglês *Multiple Sequence Alignment*.
PDB	*Protein Data Bank*.
PMI	Informação mútual pontual, do inglês *Pointwise Mutual Information*.
PSI-BLAST	*Position-Specific Iterated BLAST*.
RNA	Ácido ribonucleico, do inglês *RiboNucleic Acid*.
RMN	Ressonância Magnética Nuclear.
STRING	*Search Tool for the Retrieval of Interacting Genes*.
TIM	Triose-fosfato isomerase, do inglês *Triose-phosphate IsoMerase*.

Sumário

Capítulo 1
Introdução

Sequenciamento de genomas, identificação de genes e anotação funcional de produtos gênicos são passos básicos para a compreensão do amplo espectro de processos biológicos que ocorrem em um organismo. A determinação experimental de funções proteicas é provavelmente a etapa mais desafiadora (White, 2006), tendo a caracterização funcional de proteínas tornado-se o passo limitador do uso de informação biológica na prática (Brown et al., 2006).

A função de uma proteína é um conceito que pode ter diferentes interpretações, dependendo do contexto biológico no qual ela está inserida. Geralmente, essa função compreende aspectos bioquímicos, celulares e fenotípicos dos eventos moleculares que envolvem a proteína, incluindo o modo como ela interage com outros componentes no seu ambiente (p. ex.: com pequenos compostos ou patógenos). De acordo com Radivojac et al. (2013), é difícil prever a função de uma proteína por diversas razões. Primeiro, porque a função é definida sob vários aspectos e em múltiplos níveis: descreve, por exemplo, os eventos bioquímicos envolvendo a proteína e também como cada proteína afeta vias metabólicas, células, tecidos, e o organismo inteiro. Segundo, porque função proteica e sua caracterização experimental são dependentes de contexto: é pouco provável que um experimento científico determine o repertório funcional inteiro de uma proteína sob todas as condições às quais encontra-se submetida (p. ex.: temperatura, pH ou presença de possíveis ligantes). Terceiro, porque muitas proteínas são multifuncionais ou promíscuas: considerando apenas as proteínas experimentalmente anotadas no banco de dados Swiss-Prot, 30% têm mais de uma função (termo-folha da árvore) na ontologia de função molecular do *Gene Ontology*, e 60% na ontologia de processo biológico. Quarto, porque além de serem incompletas, as anotações funcionais disponíveis estão sujeitas a erros na interpretação do experimento ou problemas na curadoria. Por último, porque os esforços atuais em grande parte mapeiam funções proteicas a nomes de genes, confundindo assim as funções de potenciais isoformas distintas de um mesmo gene.

A anotação precisa de funções proteicas é fundamental para a compreensão da vida em nível molecular e tem enormes implicações biomédicas e farmacêuticas, assim como em diversas outras áreas, devido ao papel fundamental das proteínas, principalmente enzimas, nas indústrias alimentícia, têxtil, química e de papel e celulose, por exemplo (Radivojac et al., 2013). No entanto, apesar dos melhores esforços de pesquisa, um número substancial e cada vez maior de proteínas previstas ainda apresenta função desconhecida (Babbitt, 2003). Segundo Erdin et al. (2011), os cerca de novecentos projetos genoma já completados produziram mais de treze milhões de sequências proteicas, das

quais apenas 1% possuem anotações de função experimentais, 64% têm anotações inferidas e 35% têm função desconhecida no banco de dados UniProt. Além disso, cerca de 40% das estruturas resolvidas por projetos de genômica estrutural possuem função desconhecida no banco de estruturas proteicas *Protein Data Bank* e, mesmo considerando anotações automáticas, quase três mil estruturas permanecem listadas como não anotadas no *Structural Genomics Knowledgebase*. Ademais, é fundamental considerar que as anotações existentes não são necessariamente precisas, visto que a maioria depende de homologia, supondo, muitas vezes equivocadamente, que proximidade evolutiva implica em compartilhamento de função.

O aumento sem precedentes do número de novas sequências proteicas sendo produzidas por projetos de genômica e proteômica, além das muitas estruturas proteicas de função desconhecida sendo resolvidas pela genômica estrutural, enfatiza diretamente a necessidade por métodos computacionais para determinar, rápida e confiavelmente, as funções moleculares e celulares dessas proteínas, uma vez que a investigação experimental é difícil e tem altos custos financeiro e temporal (Zhang e Kim, 2003; Lee et al., 2007). Sendo assim, à medida que o número de genomas sequenciados cresce rapidamente, a imensa maioria dos produtos gênicos somente pode ser anotada computacionalmente. Por isso, a anotação computacional de função proteica surgiu como a vanguarda da biologia computacional e molecular (Radivojac et al., 2013).

Embora muitas soluções tenham sido propostas nas últimas décadas (Bork et al., 1998; Rost et al., 2003; Watson et al., 2005; Friedberg, 2006; Sharan et al., 2007; Lee et al., 2007; Punta e Ofran, 2008; Rentzsch e Orengo, 2009; Xin e Radivojac, 2011), a tarefa de inferência computacional de função geralmente utiliza abordagens tradicionais como a identificação de domínios ou a busca por similaridades entre proteínas com função experimentalmente determinada. Mais recentemente, a disponibilidade de informação de sequência em nível genômico para milhares de espécies, combinada com uma quantidade massiva de dados experimentais de larga escala, criou novas oportunidades para a anotação automática de função (Radivojac et al., 2013). Um grande número de métodos foram propostos para explorar esses dados, incluindo anotação de função a partir de sequências de aminoácidos (Jensen et al., 2002; Wass e Sternberg, 2008; Martin et al., 2004; Hawkins et al., 2006; Clark e Radivojac, 2011), relacionamentos evolutivos inferidos e contexto genômico (Pellegrini et al., 1999; Marcotte et al., 1999; Enault et al., 2005; Engelhardt et al., 2005; Gaudet et al., 2011), redes de interação proteína-proteína (Deng et al., 2003; Letovsky e Kasif, 2003; Vazquez et al., 2003; Nabieva et al., 2005), dados de estrutura proteica (Pazos e Sternberg, 2004; Pal e Eisenberg, 2005; Laskowski et al., 2005b) e dados de microarranjo (Huttenhower et al., 2006). Segundo Galperin e Koonin (2010), atualmente não existem abordagens de larga escala capazes de revelar a função de todos os genes hipotéticos nos genomas já sequenciados. Esse objetivo só é alcançável por meio dos esforços de biólogos experimentais, computacionais e estruturais. O presente trabalho é um esforço computacional visando a dar um passo em direção a esse objetivo.

A abordagem mais comum para a anotação funcional de proteínas é a transferência de anotações por meio de homologia, que utiliza o conhecimento de que proteínas com sequências similares frequentemente desempenham funções semelhantes (Lee et al., 2007). Os algoritmos em geral aplicam métricas de similaridade sob a premissa de que proteínas suficientemente parecidas em sequência e estrutura realizarão funções semelhantes. Tais métodos apresentam várias limitações, visto que homologia não implica, necessariamente, em isofuncionalidade (Smith, 2012). Devido à combinação da imprecisão inerente às bases de dados com a plasticidade das funções proteicas, vários aspectos de função não podem ser confiavelmente transferidas entre sequências similares indiscriminadamente (Devos e Valencia, 2000). De fato, métodos baseados em transferência de anotações por homologia são considerados uma das principais fontes de erros de anotação em consequência da aplicação excessivamente liberal de herança de função (Lee et al., 2007), que falha quando não é possível identificar proteínas similares ou quando essas também não apresentam anotações confiáveis (Dobson e Doig, 2005; Boareto et al., 2012). A transferência de anotações por homologia falha também no caso de proteínas que desempenham a mesma função embora apresentem sequências e estruturas diferentes (convergência funcional) (Kumar e Choudhary, 2012), assim como para proteínas sequencial e estruturalmente semelhantes que divergiram funcionalmente ao longo da evolução (Boareto et al., 2012).

A literatura mostra que a utilização de apenas um tipo de dados como, por exemplo, similaridade de sequências, é insuficiente para anotar funções proteicas com precisão, devido à imensa quantidade de fatores envolvidos na determinação de uma função e à consequente complexidade do problema de anotação automática. A combinação de diversos tipos de informação é crucial para a anotação de funções proteicas (Furnham et al., 2012). Segundo Lee et al. (2007), frequentemente observa-se que o poder de anotação funcional de uma abordagem combinada é maior do que o dos componentes utilizados individualmente. Sendo assim, fica destacada a grande importância e demanda por métodos de análise automática de função proteica capazes de integrar vários tipos de dados. Felizmente, pesquisas em genômica e proteômica geram imensas quantidades e variedades de informações tais como sequências de DNA, RNA e aminoácidos, dados sobre vias metabólicas, e estruturas e domínios proteicos. No entanto, esse conhecimento encontra-se disperso em múltiplas bases de dados, dificultando sua integração.

Além das dificuldades mencionadas, o problema de atribuir-se uma função a uma família de proteínas (um grupo de proteínas que compartilham uma origem evolutiva, o que reflete-se em suas funções relacionadas e suas similaridades em sequência ou estrutura) é ainda mais complexo quando considera-se que muitas delas são compostas por proteínas com múltiplos enovelamentos ou funções. De acordo com Devos e Valencia (2000), nesses casos a determinação de possíveis subfamílias pode levar a informações importantes sobre a função e estrutura de uma proteína de função desconhecida associada à família, assim como sobre a diversificação funcional adquirida pela família ao longo da evolução. Uma família de proteínas homólogas pode, então, ser dividida em subtipos que compartilham funções específicas mas que não são comuns à família como um todo (Capra e Singh, 2008), o

que vem sendo feito por muitos repositórios de famílias proteicas que buscam superar o problema da diversidade funcional em uma família (Lee et al., 2007). Ainda segundo esses autores, o sucesso de um repositório em geral pode ser aumentado por meio da identificação de resíduos específicos que discriminam entre funções.

O propósito deste trabalho é a detecção de subfamílias isofuncionais em uma família de proteínas de função desconhecida, além da identificação dos resíduos responsáveis pela diferenciação entre elas. Para tanto, são utilizados algoritmos de agrupamento, uma das principais tarefas da mineração de dados, que permitem agrupar um conjunto de objetos de modo que aqueles em um mesmo grupo sejam mais similares entre si do que a objetos de outros grupos. Acredita-se que a determinação dessas subfamílias seja um primeiro passo para reduzir a complexidade do problema de anotação funcional de proteínas. Uma vez que as muitas dificuldades enfrentadas para a realização de anotação automática de função podem ser estendidas ao problema de determinar subfamílias, para contornar as principais falhas comuns a métodos de anotação baseados em homologia, é adotada uma abordagem que integra diversos tipos de dados que são possíveis indicadores de similaridade entre proteínas. Assim, há duas perguntas principais a serem respondidas ao longo do desenvolvimento deste trabalho: (i) é possível detectar subfamílias funcionais proteicas por meio da integração de vários tipos de dados? e (ii) quais desses tipos de dados têm maior capacidade de discriminar subfamílias?

Objetivos

Neste trabalho, a similaridade entre proteínas em relação a vários tipos de dados é estudada e interpretada como evidência, ainda que fraca, de similaridade funcional. O principal objetivo é analisar o modo como a integração de informação proveniente de diferentes domínios de conhecimento é capaz de direcionar um processo não-supervisionado de agrupamento a detectar, em uma família de proteínas de função desconhecida, subfamílias possivelmente isofuncionais. São objetivos específicos:

- ◇ estudar técnicas para integrar diferentes tipos de dados biológicos;
- ◇ propor e implementar uma técnica de integração de dados;
- ◇ estudar algoritmos de agrupamento descritos na literatura;
- ◇ implementar e aplicar técnicas de agrupamento para o problema de detecção de subfamílias proteicas;
- ◇ analisar os perfis dos grupos de proteínas obtidos;
- ◇ estudar a utilidade e relevância dos diferentes tipos de dados para discriminar subfamílias;
- ◇ comparar os resultados obtidos com aqueles oriundos de uma técnica similar da literatura.

Justificativa

A aplicação das técnicas aqui propostas ao cenário de detecção de subfamílias proteicas pode levar à obtenção de informações importantes sobre função e estrutura de proteínas e sobre a ocorrência de diversificação funcional durante a evolução das famílias. Esse tipo de metodologia, que integra informações provenientes de fontes diversas e possivelmente incompletas, é de grande interesse para um cenário como esse, visto que dados oriundos de experimentos biológicos são geralmente imprecisos, principalmente devido à natureza dinâmica dos fenômenos investigados, além de que alguns tipos de informação são relativamente escassos. Além disso, a função molecular de uma proteína é determinada por diversos fatores, e a complementaridade das diferentes fontes de dados permite que os algoritmos trabalhem com o máximo de informação possível. A metodologia proposta permite levantar hipóteses sobre fatores determinantes de especificidade em uma família de proteínas com função desconhecida, o que será útil para experimentos posteriores que buscam, por exemplo, caracterizar enzimas com potencial biotecnológico e estudar sua reengenharia.

A metodologia desenvolvida neste trabalho visa a detectar subfamílias de proteínas utilizando vários tipos de dados interpretados como evidências de similaridades funcionais. As principais contribuições são a verificação de que a integração de informações provenientes de diversos domínios de conhecimento é capaz de melhorar a qualidade dos grupos detectados; a estratégia adotada para integrar os diversos tipos de dados antes de fornecê-los aos algoritmos de agrupamento; e, principalmente, a possibilidade de utilizar conhecimento de domínio para detectar subfamílias proteicas em uma família de função desconhecida.

Por tratar-se de uma tese multidisciplinar, o restante do texto está organizado como segue. No Capítulo 2 são apresentados os conceitos biológicos fundamentais à compreensão da motivação deste trabalho, assim como do conhecimento de domínio empregado como evidência de similaridade funcional, enquanto no Capítulo 3 são apresentados os conceitos e ferramentas computacionais empregados. No Capítulo 4 é feita uma revisão da literatura relacionada à anotação automática de funções e à busca por diferentes especificidades em famílias de proteínas. O Capítulo 5 detalha a metodologia desenvolvida e as famílias de proteínas nas quais foi aplicada. Resultados e discussões são mostrados no Capítulo 6, enquanto o Capítulo 7 apresenta conclusões e considerações finais.

Capítulo 2
Proteínas

Neste capítulo são abordados conceitos relacionados a proteínas necessários à compreensão da motivação deste trabalho, assim como do conhecimento de domínio utilizado como evidência de similaridade funcional. Todos os conceitos e propriedades aqui descritos são estudados e empregados explícita ou implicitamente por estarem de algum modo relacionados a função proteica. Por isso, a semelhança entre pares de proteínas em relação aos domínios de conhecimento apresentados é interpretada como possível indicador de similaridade funcional. Os modos como são empregados são detalhados no Capítulo 5.

Após a água, a maior parte da célula é constituída por proteínas, polímeros de aminoácidos que são, possivelmente, as biomoléculas mais versáteis. Algumas possuem atividade catalítica e funcionam como enzimas, enquanto outras atuam como elementos estruturais, receptores e transmissores de sinais, ou transportadores de substâncias específicas para dentro ou fora das células. Cada proteína tem uma sequência de aminoácidos característica que pode chegar a milhares de resíduos e contém a informação que dá à molécula sua estrutura tridimensional e suas funções biológicas. Por exemplo, uma dada sequência de aminoácidos produz uma estrutura forte e fibrosa encontrada em cabelo e lã; outra sequência produz uma proteína que transporta oxigênio no sangue; uma terceira liga-se a outras proteínas e catalisa a clivagem de ligações entre seus aminoácidos (Nelson e Cox, 2005).

Há quatro níveis de estrutura proteica. Uma descrição de todas as ligações covalentes entre resíduos em uma cadeia polipeptídica é a **estrutura primária**, cujo elemento mais importante é a *sequência* de resíduos. A **estrutura secundária** refere-se a arranjos particularmente estáveis de resíduos que dão origem a padrões estruturais recorrentes, sendo α-hélices e folhas-β as mais comuns, enquanto a **estrutura terciária** descreve todos os aspectos do enovelamento tridimensional de um polipeptídeo. Quando uma proteína tem mais de uma subunidade polipeptídica, seu arranjo espacial é chamado **estrutura quaternária**. Uma proteína pode associar-se a outras proteínas, ácidos nucleicos ou lipídios para formar complexos como cromossomos, ribossomos e membranas. Cada molécula nesses complexos tem sítios de ligação específicos e de alta afinidade umas com as outras, de modo a formarem complexos funcionais espontaneamente na célula (Nelson e Cox, 2005).

2.1 Função Proteica

Sequenciamento de genomas, identificação de genes e anotação funcional de produtos gênicos são passos básicos para a compreensão do amplo espectro de processos biológicos que ocorrem em um organismo. A determinação experimental de funções proteicas é provavelmente a etapa mais desafiadora (White, 2006). De acordo com Brown et al. (2006), a caracterização funcional de proteínas tornou-se o passo limitador do uso prático de informação biológica. O aumento sem precedentes do número de novas sequências proteicas sendo produzidas por projetos de genômica e proteômica, além das muitas estruturas proteicas de função desconhecida sendo resolvidas pela genômica estrutural, enfatiza diretamente a necessidade por métodos para determinar, rápida e confiavelmente, as funções moleculares e celulares dessas proteínas, uma vez que a investigação experimental tem altos custos financeiro e temporal (Zhang e Kim, 2003; Lee et al., 2007). Esforços de anotação funcional em larga escala têm focado em métodos computacionais para anotação de função proteica, visto que métodos mais tradicionais somente são eficientes quando aplicados a pequenos subconjuntos dos dados disponíveis. Métodos automatizados de anotação funcional a partir de sequências e estruturas são importantes para preencher a lacuna existente entre a quantidade de estruturas conhecidas e a quantidade de estruturas com funções anotadas (Dobson e Doig, 2005). Dado que a avaliação experimental da função de cada proteína em cada genoma sequenciado está além dos recursos previsíveis, o conhecimento sobre a maioria das proteínas novas será derivado de anotações funcionais baseadas em buscas por similaridade (Devos e Valencia, 2000).

Ainda não há uma definição geral de função proteica que seja suportada por uma ontologia bem definida válida entre domínios e organismos distintos (Devos e Valencia, 2000). Segundo Lee et al. (2007), pode-se considerar função proteica em diferentes níveis interdependentes, dividindo-a em três categorias principais: **função molecular**, que descreve a atividade proteica em nível molecular, comumente prevista por métodos que identificam homólogos ou ortólogos; **processo biológico**, que descreve funções mais amplas desempenhadas por combinações de funções moleculares; e **componente celular**, que descreve compartimentos celulares nos quais a proteína realiza sua função. Para Szklarczyk et al. (2015), o conceito de função proteica é de certa forma hierárquico e, em todos os níveis dessa hierarquia, interações entre proteínas ajudam a descrever e refinar a função de uma proteína: sua estrutura tridimensional pode tornar-se significativa apenas no contexto de um dado conjunto de proteínas, suas ações moleculares podem ser reguladas por ligações cooperativas ou alosteria, e seu contexto celular pode ser controlado por uma multitude de interações de transporte, sinalização e captura.

Para um conjunto limitado de genes essenciais, a noção de função parece direta: a função é o que o produto gênico precisa fazer para permitir o crescimento celular. Operacionalmente, se um gene é nocauteado, a célula morre e pode-se supor que a causa da morte é a função do gene em questão. O caso de genes não-essenciais é mais complexo, uma vez que as funções de muitas, senão todas,

as proteínas são inerentemente multifacetadas e complexas. Uma única Oxirredutase, por exemplo, pode usar um leque de substratos e uma variedade de aceptores de elétrons, dificultando, ou mesmo impossibilitando, a atribuição precisa de função (Galperin e Koonin, 2010).

Segundo Boareto et al. (2012), um grande número de estudos sugere que funções proteicas evoluem fundamentalmente por recrutamento, de modo que um mesmo motivo estrutural ou enovelamento pode ser utilizado para realizar diferentes funções enzimáticas. Alguns aminoácidos específicos são responsáveis pela atividade de uma enzima, sendo, portanto, essenciais. Tais aminoácidos devem pertencer a sítios ativos, por isso um método efetivo de anotação funcional precisa ser capaz de reconhecê-los.

O entendimento de funções moleculares geralmente pode ser imensamente facilitado por conhecimento de domínio como, por exemplo, de similaridade estrutural. Segundo Sadowski e Jones (2009), a íntima associação, para a maioria das proteínas, entre função e estrutura nativa é hoje muito bem conhecida por meio de vários estudos. Uma abordagem para atribuir a função molecular de uma proteína consiste em primeiro determinar sua estrutura tridimensional por meio de cristalografia de raios-x ou ressonância magnética nuclear e, então, comparar a estrutura resolvida contra bancos de dados de estruturas conhecidas. Caso hajam homólogos estruturais significativos, prevê-se que a proteína hipotética tenha funções moleculares similares às dos homólogos, apesar da ausência de similaridades de sequência. As previsões podem, então, ser experimentalmente testadas (Zhang e Kim, 2003). No entanto, infelizmente ainda há uma pequena fração de estruturas resolvidas experimentalmente em comparação com o grande número de sequências proteicas disponíveis. Mesmo assim, a capacidade de resolução de estruturas ainda é maior que a de atribuição experimental de funções, motivo pelo qual muitas estruturas novas carecem de anotação funcional.

2.2 Proteínas de Função Desconhecida

Apesar dos melhores esforços de biólogos experimentais e computacionais, um número substancial e cada vez maior de proteínas deduzidas apresentam função desconhecida. Segundo Zhang e Kim (2003), em todos os genomas já sequenciados, uma grande porção das regiões codificadoras previstas codificam polipeptídeos de funções biológicas desconhecidas, ou seja, proteínas hipotéticas. Um grande desafio é o de encontrar modos de determinar as funções molecular (bioquímica e biofísica) e celular de tais proteínas de modo confiável. De acordo com Galperin e Koonin (2010), alguns genes muito comuns entre bactérias, arqueias e eucariotos ainda não possuem anotação funcional. Refere-se variadamente aos produtos proteicos desses genes como hipotéticos (*hypothetical*), hipotéticos conservados (*conserved hypothetical*), não-caracterizados (*uncharacterized*) ou mesmo putativos não-caracterizados (*putative uncharacterized*). No entanto, um número significativo de atividades enzimáticas órfãs ainda não foram associadas a nenhuma sequência proteica, o que sugere que alguns genes hipotéticos podem desempenhar funções bem conhecidas. Atualmente, não existem

abordagens de alto rendimento que revelem a função de todos os genes hipotéticos codificados nos genomas já sequenciados. Esse objetivo só é alcançável por meio de esforços de inúmeros biólogos experimentais, computacionais e estruturais. O presente trabalho é um esforço computacional visando a dar um passo em direção a esse objetivo.

Segundo Punta et al. (2012), a versão 26 do banco de dados de famílias proteicas Pfam continha 3.526 famílias anotadas como DUFs (*Domain of Unknown Function*), que são famílias para as quais não há anotação funcional. Tais famílias constituíam mais de um quarto de todas as famílias dessa versão do Pfam, e seu número aumentou sistematicamente com o passar dos anos. Famílias DUFs são, em média, menos amplamente distribuídas na árvore evolutiva que famílias funcionalmente anotadas. Então, apesar de representarem 26,5% das famílias dessa versão, elas respondem por apenas 6,7% da cobertura do banco de dados de sequências proteicas UniProt. Normalmente, quando a função de pelo menos uma proteína em uma família DUF é experimentalmente determinada, a família é renomeada. Embora o número de famílias DUFs funcionalmente caracterizadas que foram renomeadas venha crescendo, esse crescimento é inferior ao do número de novas famílias DUFs geradas. O Pfam também contém vários domínios que receberam o nome de proteínas representativas de bactérias mas cuja função ainda não foi caracterizada, de modo que permanecem na categoria de domínio de função desconhecida. As famílias DUFs incluem numerosas famílias potencialmente de grande interesse para caracterização experimental de função. Entre essas, estão cerca de trezentas famílias DUFs encontradas em mais de cem genomas representativos. A ampla distribuição taxonômica dessas famílias sugere que elas provavelmente estejam associadas a funções celulares importantes. Além disso, a lista de famílias DUFs da versão 26 do Pfam incluía mais de 400 famílias com pelo menos uma proteína humana.

Segundo Erdin et al. (2011), os cerca de novecentos projetos genoma já completados até então produziram mais de treze milhões de sequências proteicas, das quais 1% possuem anotações experimentais, 64% têm anotações inferidas e 35% permanecem rotuladas como putativas, não-caracterizadas, hipotéticas ou de função desconhecida no UniProt. Além disso, 40% das quase dez mil estruturas resolvidas por projetos de genômica estrutural possuem função desconhecida no *Protein Data Bank* (PDB) e, mesmo considerando anotações automáticas, quase três mil estruturas permanecem listadas como não anotadas no *Structural Genomics Knowledgebase*. Ainda é preciso considerar que as anotações existentes não são necessariamente precisas, visto que a maioria depende de homologia, supondo que a proximidade evolutiva implica compartilhamento de função. Algoritmos de anotação funcional aplicam métricas de similaridade entre proteínas sob a premissa de que aquelas suficientemente parecidas em sequência e estrutura realizarão funções idênticas. Segundo os autores, as análises de rede que integram tais métricas são a promessa de ganhos rápidos na especificidade de anotação funcional.

De acordo com Devos e Valencia (2000), resultados obtidos analisando um número significativo de similaridades sequenciais verdadeiras, derivadas diretamente de alinhamentos estruturais, apontam a complexidade do problema de anotação automática de função. Diferentes aspectos de função proteica, incluindo classificação de função enzimática, anotações funcionais na forma de palavras-chave, classes de função celular e conservação de sítios de ligação, não podem ser confiavelmente transferidos entre sequências similares indiscriminadamente. O motivo para tal dificuldade é uma combinação da inevitável imprecisão inerente às bases de dados e da plasticidade das funções proteicas. Em casos em que uma proteína de função desconhecida pode ser ligada a uma superfamília de múltiplos enovelamentos ou funções, a determinação de possíveis subfamílias pode levar a informação importante sobre a função e estrutura da proteína, assim como sobre a diversificação funcional adquirida pela família ao longo da evolução. Neste trabalho são pesquisadas formas de integrar vários tipos de dados, visando à detecção de subfamílias possivelmente isofuncionais em famílias proteicas. Acredita-se que a determinação dessas subfamílias seja um primeiro passo para reduzir a complexidade do problema de anotação funcional de proteínas.

2.3 Enzimas

Uma grande porção do proteoma é constituída por enzimas, proteínas altamente especializadas que catalisam as reações químicas envolvidas no metabolismo de todos os seres vivos (Friedberg, 2006), transformando substratos em produtos enquanto elas mesmas permanecem inalteradas (Alberts et al., 2010). Reações chave do metabolismo celular são catalisadas por enzimas, que permitem que algumas reações aconteçam rapidamente, enquanto as taxas de outras não são alteradas (Neitzel, 2010). Praticamente todo processo biológico requer uma enzima em algum momento, pois elas são ferramentas naturais que constroem ou quebram moléculas que as células precisam para seu crescimento, reparação e manutenção (Bartlett et al., 2002).

Enzimas controlam a atividade de um sistema vivo, e a capacidade de catalisarem reações efetiva e seletivamente é vital, pois são necessárias para que quase todos os processos celulares ocorram a taxas significativas e necessárias à manutenção da vida, acelerando reações de 10^5 a 10^{17} vezes. Reações necessárias para a digestão de alimentos e envio de sinais nervosos, por exemplo, simplesmente não ocorreriam a taxas necessárias sem catálise. O notável poder catalítico das enzimas é muito maior que o de catalisadores sintéticos ou inorgânicos, por isso seu estudo é de grande importância prática. Algumas doenças, por exemplo, podem ser causadas por um aumento na taxa de atividade enzimática, enquanto outras, especialmente desordens genéticas herdáveis, podem ser relacionadas à deficiência ou ausência de determinadas enzimas (Nelson e Cox, 2005). Essas catalisadoras naturais são importantes ferramentas práticas em diversas áreas, tais como medicina, agricultura e indústrias química, farmacêutica e alimentícia.

A extrema especificidade de substrato e alta eficiência catalítica das enzimas reflete a existência de um ambiente que é primorosamente adaptado para que uma dada reação ocorra: o **sítio ativo**, que geralmente tem a forma de uma cavidade (Murray et al., 2003). Enquanto alguns sítios ativos estão posicionados em cavidades na superfície da enzima, outras estão bastante enterrados, sendo conectados ao exterior por um ou mais canais (Pravda et al., 2014). O sítio ativo consiste de resíduos que formam ligações temporárias com o substrato (o sítio de ligação) e de resíduos que catalisam a reação no mesmo (o sítio catalítico). Segundo Nelson e Cox (2005), a superfície do sítio ativo é revestida de resíduos com grupos substituintes que ligam-se ao substrato e catalisam sua transformação química. Muitas vezes, o sítio ativo envolve um substrato, sequestrando-o completamente da solução. De acordo com Murray et al. (2003), substratos ligam-se ao sítio ativo em uma região complementar que não passa por mudanças químicas permanentes durante o curso da reação, fazendo com que porções do substrato que irão sofrer alterações sejam alinhadas, simultaneamente, aos grupos quimicamente funcionais dos resíduos do sítio ativo. Além do substrato, o sítio ativo também liga e orienta espacialmente cofatores ou grupos prostéticos (componentes não-proteicos de proteínas conjugadas que são essenciais para a atividade biológica das mesmas). Muitos resíduos de várias partes da cadeia polipeptídica contribuem para o tamanho relativamente grande e para o caráter tridimensional do sítio ativo. Segundo Neitzel (2010), embora seja possível identificar aminoácidos específicos envolvidos nas reações químicas da catálise, a capacidade da enzima em atuar como um catalisador efetivo também pode depender de interações entre aminoácidos distantes do sítio ativo.

Segundo Bartlett et al. (2002), resíduos do sítio ativo são considerados catalíticos se eles têm envolvimento direto no mecanismo catalítico; se exercem um efeito sobre outro resíduo ou molécula de água que está diretamente envolvido no mecanismo catalítico que auxilia a catálise; se estabilizam um estágio intermediário de transição; ou se exercem um efeito sobre um substrato ou cofator que auxilia a catálise. Resíduos que apenas ligam-se ao substrato, cofator ou metal não são considerados catalíticos a não ser que realizem uma dessas funções. Os autores mostraram que resíduos catalíticos, assim como seu ambiente tridimensional, são altamente conservados, claramente mais do que a média, e que a pressão seletiva sobre eles é muito maior do que sobre resíduos na vizinhança do sítio ativo, o que é importante para o reconhecimento do substrato. Adicionalmente, mostraram que resíduos estruturalmente próximos a resíduos catalíticos são mais conservados que aqueles próximos a eles na sequência.

2.4 Evolução Proteica

O principal objetivo do processo de inferência de relações evolutivas é modelar os mecanismos moleculares pelos quais as sequências proteicas evoluíram (Zaha et al., 2012). Essa evolução não ocorreu por um caminho simples e linear; o processo evolutivo envolve substituições, inserções e deleções na sequência de aminoácidos, e esses resíduos variáveis podem ser utilizados para traçar a evolução.

Algumas proteínas têm mais resíduos variáveis que outras, de forma que proteínas evoluem a taxas diferentes umas das outras (Nelson e Cox, 2005). Substituições de resíduos são pseudoaleatórias. No entanto, em algumas posições na sequência de uma proteína, a necessidade de manter a função pode significar que apenas determinadas substituições possam ser toleradas. Segundo Sigrist et al. (2010), é evidente, no estudo de sequências de famílias proteicas, que algumas regiões foram mais conservadas que outras durante a evolução. Tais regiões geralmente são importantes para a função de uma proteína e/ou para a manutenção de sua estrutura tridimensional e, por serem essenciais à atividade proteica, são conservados ao longo do tempo evolutivo (Sigrist et al., 2010; Nelson e Cox, 2005). Já resíduos funcionalmente menos importantes podem variar, ou seja, um aminoácido pode vir a substituir outro. Sequências curtas conservadas são chamadas **motivos** e podem indicar função proteica, sendo muito úteis para o problema de anotação funcional (Dobson e Doig, 2005) e, por essa razão, são considerados neste trabalho como fonte de evidência de similaridade funcional.

Podem haver modificações extensas entre as sequências de proteínas distantemente relacionadas, de modo a produzir enovelamentos em que números e orientações de estruturas secundárias variam consideravelmente (Zaha et al., 2012). Segundo Shah e Hunter (1997), um relacionamento evolutivo entre proteínas não implica que elas sejam funcionalmente relacionadas, visto que muitas proteínas evolutivamente relacionadas podem ter divergido funcionalmente, assim como proteínas não relacionadas evolutivamente podem ter convergido para desempenharem a mesma função.

2.5 Transferência de Anotações por Homologia

Definir a função de uma proteína não é uma tarefa simples, por isso é uma área de pesquisa muito ativa. A abordagem mais comum e acessível para anotar proteínas é a herança por meio de homologia, que utiliza o conhecimento de que proteínas com sequências similares frequentemente desempenham funções semelhantes (Lee et al., 2007). Convencionalmente, a homologia de sequência tem sido utilizada como fonte-chave de informação para a transferência de anotações por meio da comparação das novas sequências com uma base de dados de genes anotados (Chitale et al., 2009).

Proteínas homólogas são aquelas que descenderam, geralmente com divergência, de uma sequência ancestral comum (Lee et al., 2007). Segundo Smith (2012), dois genes ou produtos gênicos de diferentes organismos são homólogos quando há evidência suficiente para afirmar que eles evoluíram a partir de um mesmo gene ancestral. Para Zaha et al. (2012), assume-se que quaisquer genes ou sequências nucleotídicas com identidade parcial ou total compartilhem uma origem evolutiva. Em alguns casos, similaridades entre estruturas também refletem origens evolutivas comuns. Diferentes tipos de homologia existem conforme os processos evolutivos que originaram os genes homólogos, de modo que inferir a conservação de função a partir de genes homólogos requer que a identificação do tipo de homologia envolvido (Smith, 2012).

Um evento evolutivo de surgimento de novas espécies (especiação) leva à criação de duas espécies a partir de uma única ancestral. Genes com um ancestral comum separados apenas por eventos de especiação são chamados **ortólogos** e conservam função, ocupando nichos funcionais iguais ou similares nas diferentes espécies. Já um evento de duplicação gênica resulta em genes **parálogos** que, para serem mantidos em um genoma ao longo da evolução, estão suscetíveis de evoluir para desempenharem diferentes funções, variando desde diferenças sutis de substrato (p. ex.: Malato e Lactato Dehidrogenases), a similaridades fracas em função molecular (p. ex.: Hidrolases), até uma total diferença de localização e função celular (p. ex.: homologia entre o domínio de sinalização intracelular e o fator de crescimento secretado). Além disso, a função molecular de parálogos pode ser mantida, alterando apenas a função celular (p. ex.: enzimas com diferentes especificidades de tecido). Parálogos podem existir em espécies diferentes devido a um evento de especiação posterior ao de duplicação. Um evento de transferência gênica horizontal, quando um trecho de DNA é transferido de um organismo de uma espécie a outro de espécie diferente, gera genes **xenólogos** (Smith, 2012; Nelson e Cox, 2005; Zaha et al., 2012; Lee et al., 2007; Betts e Russell, 2003). A distância evolutiva entre organismos pode impossibilitar a distinção entre ortólogos e parálogos por métricas simples de similaridade de sequências (Betts e Russell, 2003). Adicionalmente, existem genes **análogos**, que apresentam funções moleculares similares embora tenham evoluído separadamente (Smith, 2012).

Segundo Smith (2012), uma vez estabelecida a homologia entre um par de genes, uma informação sobre um gene pode fornecer suporte para a anotação funcional do outro. A similaridade de sequências é frequentemente analisada para detectar homólogos, mas isso está propenso a erros e leva à detecção de falsas homologias. De acordo com Lee et al. (2007), muitos estudos de anotação automática de função buscaram estabelecer medidas de similaridade de sequências para transferir função com segurança entre proteínas relacionadas. No entanto, genes evoluem a taxas diferentes devido a pressões seletivas desiguais sobre suas funções e a taxas de mutação inerentes às espécies, o que significa que é muito difícil estabelecer uma medida de similaridade que seja confiável em todos os casos. Para Smith (2012), domínios proteicos conservados são sinais de homologia mais orientados a função do que similaridade de sequências, visto que proteínas contêm partes estruturalmente distintas que são unidades de enovelamento, função e evolução proteica, e as subsequências desses domínios são, geralmente, muito mais conservadas que o restante da sequência. Considerando sua relação com funções dos produtos gênicos, domínios conservados podem ser um marcador forte do relacionamento evolutivo que conserva a função proteica, por isso sua utilidade na determinação de famílias de genes e proteínas homólogos. No entanto, segundo Redfern et al. (2009), a presença de um domínio em uma determinada enzima, por exemplo, não significa necessariamente que ele contribua com sua atividade catalítica. Ele pode ser responsável por interações proteína-proteína ou por outros aspectos da função, como o posicionamento da proteína em uma determinada parte da célula.

De acordo com Smith (2012), métodos de transferência de anotação com base em homologia apresentam várias limitações. Além da simples divergência sugerida por paralogia, a homologia

não necessariamente implica isofuncionalidade, dado que funções proteicas podem não ser únicas e podem divergir muito rapidamente com poucas mutações. Por exemplo, muitas proteínas podem desempenhar diferentes papéis celulares, enquanto algumas enzimas podem catalisar conjuntos de reações metabólicas similares ou mesmo reações completamente diferentes. Algumas funções proteicas foram selecionadas durante a evolução como a função "principal" de uma proteína, enquanto muitas outras funções "secundárias" podem ter surgido ao acaso, seja ocorrendo no mesmo sítio ativo da função principal ou em outras partes da proteína.

Muitas anotações incorretas encontradas em bases de dados são consequência da aplicação excessivamente liberal de herança de função por homologia (Lee et al., 2007), que falha quando não é possível identificar proteínas similares ou quando essas também não apresentam anotações confiáveis (Dobson e Doig, 2005; Boareto et al., 2012). Para Fitch (2000), os conceitos de similaridade filogenética, similaridade estrutural e similaridade funcional não são necessariamente intercambiáveis com homologia de sequência, e esses termos mapeiam-se imperfeitamente uns com os outros. Segundo Radivojac et al. (2013), estudos recentes mostraram que a correlação entre similaridade de sequências e similaridade funcional é fraca quando aplicada a pares de proteínas, e que anotações de domínio, sozinhas, não são suficientes para resolver a função de uma proteína. Já foi demonstrado que comparações simples entre pares de sequências são inadequadas para a transferência de anotações funcionais entre proteínas com menos de 30 a 40% de identidade (Brown et al., 2006). Além disso, abordagens de anotação funcional baseadas em homologia estão entre as principais fontes de erros de anotação por falharem para proteínas que desempenham a mesma função embora apresentem sequências e estruturas diferentes (convergência funcional) (Kumar e Choudhary, 2012), assim como para proteínas sequencial e estruturalmente semelhantes que divergiram funcionalmente (Boareto et al., 2012).

No caso de enzimas, a transferência de anotações por homologia é complicada por várias razões. Segundo Bray et al. (2009), menos de 30% dos pares de enzimas com pelo menos 50% de identidade de sequências possuem a mesma anotação funcional. Além disso, sabe-se que similaridade estrutural nem sempre corresponde a similaridade catalítica. A evolução convergente é outro complicador, tendo sido detectados casos de mesma função para pares de enzimas sem nenhuma similaridade sequencial detectável, além de algumas com estruturas completamente distintas. Nesses casos, a similaridade funcional não pode ser reconhecida por métodos de comparação de sequências e/ou estruturas.

Segundo Galperin e Koonin (2010), devido à escassez de dados experimentais, informação estrutural raramente está inteiramente disponível, de modo que a atribuição de função para a maioria dos genes é baseada apenas na similaridade de sequências de seus produtos a proteínas experimentalmente caracterizadas em um pequeno número de organismos-modelo. Portanto, a transferência automática de anotações funcionais frequentemente leva a confusão. Para os autores, uma rota mais produtiva para uma anotação funcional razoável é substituir a anotação de proteínas individuais pela anotação de famílias proteicas. À medida que ensaios experimentais ficam cada vez mais atrasados

em relação à avalanche de dados genômicos, a validação experimental de uma função proteica prevista torna-se gradativamente menos provável. Em contraste, a classificação de famílias proteicas está tornando-se cada vez mais robusta. O sucesso notável de bases de dados de famílias proteicas como Pfam, InterPro, *Clusters of Orthologous Groups* (COGs) e *Conserved Domain Database* (CDD) deve-se às suas coleções abrangentes contendo muitas características úteis, além de abandonarem o objetivo esquivo de anotar toda e qualquer sequência e, ao invés disso, concentrarem-se nas características comuns a famílias proteicas.

De acordo com Lee et al. (2007), no caso de proteínas que realizam funções similares ou relacionadas mas que não apresentam similaridade global de sequências significativa, pode-se esperar que elas compartilhem algumas características, visto que devem compartilhar o mesmo maquinário celular, além de atuarem em ambientes similares. Segundo Laskowski et al. (2005b), há casos em que sequências divergiram significativamente ao longo da evolução, mas as regiões funcionalmente ativas das estruturas proteicas foram relativamente bem preservadas. Nesses casos, a inferência de função com base em similaridade de sequências torna-se arriscada, mas pode ser fortalecida uma vez que as estruturas tridimensionais sejam conhecidas e suas similaridades estruturais locais, detectadas.

Segundo Erdin et al. (2011), muitos aspectos da evolução proteica confundem a sensitividade e especificidade de esforços automáticos para determinar função proteica como, por exemplo, a existência de proteínas multifuncionais; a presença ou ausência de relacionamento evolutivo podem ser enganosas, visto que parálogos podem desenvolver funções inteiramente não-relacionadas, assim como há inúmeros exemplos de proteínas não-relacionadas convergindo para desempenharem funções similares e tal convergência é difícil de discernir mesmo em nível molecular; a resposta funcional a mutações pode variar entre alterações bruscas de enovelamento e mudanças na especificidade funcional, ou mesmo podem não haver alterações funcionais apesar de variações nas características das cadeias laterais ou em posições de resíduos catalíticos. Para Devos e Valencia (2000), dificuldades na transferência de funções estão relacionadas, em parte, à definição teórica de função e também a problemas práticos. Alguns desses são a identificação de sequências similares em grandes bases de dados; a persistência de erros sistemáticos na detecção de homologia devido a regiões composicionalmente enviesadas de naturezas diferentes; as sequências incorretamente anotadas em diferentes bases de dados; e a propagação de erros devido à cópia repetida de anotações entre sequências similares.

Em uma família de proteínas homólogas, essas podem estar agrupadas em subtipos que compartilham funções específicas que não são comuns à família como um todo (Capra e Singh, 2008). Este trabalho visa à detecção de possíveis subfamílias isofuncionais em uma família de proteínas de função desconhecida, além da determinação de resíduos responsáveis pela diferenciação entre elas. Falhas comuns aos métodos de anotação de função por homologia são contornadas estudando o modo como a integração de diversas fontes de conhecimento de domínio afeta a qualidade dos grupos obtidos por algoritmos de agrupamento.

2.6 Alinhamento de Sequências

O alinhamento de sequências é o método mais utilizado para procurar similaridades e diferenças entre sequências de nucleotídeos ou aminoácidos, visando a inferir analogias estruturais ou funcionais e relações evolutivas (Zaha et al., 2012). O alinhamento de sequências associa explicitamente resíduos de duas ou mais sequências, sendo um dos objetivos o de determinar quando duas sequências são suficientemente similares para justificar uma inferência de homologia. Segundo Duan et al. (2006), em muitos casos, a similaridade de sequências é um forte indicativo de similaridade funcional.

Alinhar duas sequências consiste em estabelecer uma correspondência entre seus resíduos obedecendo a ordem em que aparecem. Um alinhamento global entre pares de sequências pode ser visto como o processo de deslizar uma sobre a outra até que um bom pareamento seja encontrado (Nelson e Cox, 2005). Atribui-se uma pontuação positiva a cada posição em que os resíduos nas duas sequências são idênticos, de modo a fornecer uma métrica de qualidade do alinhamento. Em alguns casos, dois segmentos alinham bem entre as proteínas, mas são conectados, em cada uma, por sequências menos relacionadas e de tamanhos diferentes. Então, lacunas (*gaps*) podem ser introduzidas em uma das sequências para permitir alinhar tais segmentos, como ilustra a Figura 2.1. No entanto, inserindo um número suficiente de lacunas, quase todo par de sequências poderia ser alinhado de alguma forma. Por isso, aplica-se penalidades quando lacunas são formadas visando a reduzir a pontuação total do alinhamento, evitando assim a obtenção de alinhamentos não-informativos (Nelson e Cox, 2005).

Figura 2.1: Emprego de lacunas no alinhamento de sequências. Fonte: Nelson e Cox (2005).

Diferentes esquemas de pontuação para igualdades, desigualdades e lacunas dão origem a alinhamentos diferentes. A Figura 2.2 apresenta um exemplo de alinhamento cujo algoritmo atribui o valor +1 a uma identidade (*match*) de caracteres, -1 a uma desigualdade (*mismatch*) e -2 quando é inserida uma lacuna.

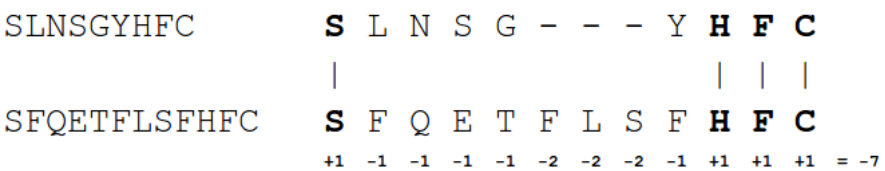


Figura 2.2: Exemplo de pontuação de um alinhamento de sequências. Fonte: Zaha et al. (2012).

Alinhamentos com pontuação alta, porém apenas matematicamente significativa, precisam ser diferenciados daqueles alinhamentos com pontuação moderada ou baixa, porém biologicamente importantes (Zaha et al., 2012). A identidade de aminoácidos em geral é inadequada para identificar

proteínas relacionadas ou determinar quão próximas elas estão em escala evolutiva. Uma análise mais útil considera as propriedades físico-químicas dos aminoácidos substituídos. Quando substituições ocorrem em uma família proteica, muitas delas podem ser conservativas, ou seja, um resíduo é substituído por outro com propriedades similares (Nelson e Cox, 2005). Assim, adota-se um esquema de pontuação em que substituições de aminoácidos mais conservativas recebam maiores pontuações do que as não-conservativas, compatibilizando assim a pontuação e a significância biológica do alinhamento. Tal ponderação é descrita por matrizes de substituição de aminoácidos, que determinam a quantidade de pontos atribuída às várias possíveis substituições, levando em consideração as taxas observadas ao longo de grandes distâncias evolutivas (Zaha et al., 2012).

Segundo Betts e Russell (2003), matrizes de substituição ou de mutação são conjuntos de números que descrevem as propensões de troca de um aminoácido por outro. Os valores são geralmente calculados utilizando um modelo de tempo evolutivo. Tais matrizes são bastante úteis como guias aproximados de quão impactante é uma determinada substituição de aminoácidos. Apesar da grande utilidade para alinhamentos de sequência, essas matrizes não descrevem precisamente a probabilidade e os efeitos de substituições em posições específicas. Matrizes de substituição de posições específicas podem ser geradas para uma família proteica de interesse como é feito, por exemplo, com os perfis de modelos ocultos de Markov (HMM, do inglês *Hidden Markov Model*) gerados pelo programa HMMER e fornecidos pela base de dados de famílias proteicas Pfam.

Entre as matrizes de substituição de aminoácidos mais conhecidas estão as matrizes BLOSUM (*BLOcks of amino acid SUbstitution Matrix*), utilizadas para alinhamento de sequências de proteínas evolutivamente divergentes, baseando-se em alinhamentos locais e levando em conta blocos correspondentes a regiões muito conservadas de famílias proteicas (Zaha et al., 2012). Segundo Nelson e Cox (2005), testes mostraram que a matriz BLOSUM62, gerada a partir de sequências com pelo menos 62% de identidade de resíduos, fornece os alinhamentos mais confiáveis para uma grande variedade de famílias proteicas e é a matriz padrão em muitos programas de alinhamento de sequência. Por isso, a BLOSUM62 é a matriz de substituição de aminoácidos empregada neste trabalho.

Existem dois modelos de alinhamento de sequências: o **global**, em que a similaridade é considerada ao longo de toda a extensão das sequências, e o **local**, no qual as regiões de similaridade são trechos das sequências. Alinhamentos globais são úteis quando as sequências estudadas são similares e com tamanho aproximadamente igual. Já alinhamentos locais são mais interessantes para sequências dissimilares que possam conter regiões de similaridade como motivos sequenciais. Segundo Zaha et al. (2012), sítios funcionais estão localizados em regiões consideradas curtas, de modo que buscar por similaridades locais pode levar a resultados mais biologicamente significativos do que aqueles obtidos por alinhamentos globais. Uma vez que neste trabalho são consideradas famílias proteicas definidas com base na presença de domínios funcionais nas sequências, ambos os tipos de alinhamento são utilizados como possíveis indicadores de similaridade funcional.

2.7 Contexto Genômico

Embora a definição de um gene seja alterada conforme os avanços das pesquisas, genes geralmente podem ser entendidos como segmentos de cromossomos que correspondem à informação necessária para produzir proteínas (Nelson e Cox, 2005). Segundo von Mering et al. (2003), proteínas funcionalmente associadas são codificadas por genes com pressões seletivas semelhantes, pois eles precisam ser mantidos e regulados juntos para que as proteínas que codificam possam interagir no mesmo tempo e mesma localização celular. A necessidade de manter juntos genes funcionalmente associados pode tornar-se observável como uma concordância em padrões de ocorrência em vários genomas: os genes tendem a estar presentes ou ausentes ao mesmo tempo, ou seja, têm o mesmo perfil filogenético. De acordo com Deng et al. (2004), acredita-se que genes com padrões filogenéticos similares tenham maior probabilidade de desempenharem funções semelhantes, de modo que a ligação funcional entre genes pode ser prevista com base nos padrões filogenéticos.

Do mesmo modo, a necessidade de regulação similar frequentemente é refletida como uma tendência de genes funcionalmente associados serem vizinhos próximos em genomas procarióticos, onde geralmente têm a mesma orientação transcricional e poucas ou nenhuma sequência entre eles, sugerindo que são óperons (grupos de genes transcritos e regulados como unidade), recorrendo em composição similar mas não idêntica em vários genomas (von Mering et al., 2003). Segundo Dobson e Doig (2005), alguns métodos de anotação funcional baseados em vizinhança gênica buscam proteínas colocalizadas em um cromossomo, visto que foi observado que proteínas funcionalmente similares frequentemente agrupam-se no genoma. Genes vizinhos em bactérias costumam ser funcionalmente relacionados devido ao seu envolvimento em um mesmo óperon (Kolesov et al., 2002). De acordo com Zhao et al. (2013), vias metabólicas bacterianas são frequentemente codificados por "vizinhanças gênicas", o que pode fornecer pistas importantes para a atribuição de função.

Finalmente, genes cujos produtos precisam interagir proximamente na célula têm tendência a serem fundidos em um único gene, codificando assim um polipeptídeo combinado em que as proteínas têm maior chance de interagir produtivamente (von Mering et al., 2003). Segundo Huynen (2000), proteínas codificadas por genes cujos contextos apresentam maior sintenia possuem maiores chances de estarem envolvidas em funções semelhantes, enquanto aquelas pertencentes à mesma família, mas cujos genes codificadores estão em contextos muito distintos, provavelmente apresentam diferentes especificidades de substrato. Neste trabalho, a similaridade entre contextos genômicos é considerada uma fonte de evidência de similaridade funcional.

2.8 Estrutura Tridimensional de Proteínas

Os resíduos em uma sequência proteica interagem entre si criando padrões de enovelamento complexos que determinam a estrutura terciária da proteína (Figura 2.3), diretamente relacionada à sua

função. Segundo Smith (2012), além da grande porção de aminoácidos da proteína que são responsáveis pelas suas estruturas secundária e terciária, vários resíduos podem afetar diretamente seu mecanismo de ação oferecendo, por exemplo, um grupo doador de prótons em uma posição específica ou deformando a nuvem eletrônica de um substrato de modo a facilitar substituições de grupo. Outros resíduos são responsáveis por aprisionar moléculas envolvidas na reação em conformações espaciais mais favoráveis ao mecanismo de ação.

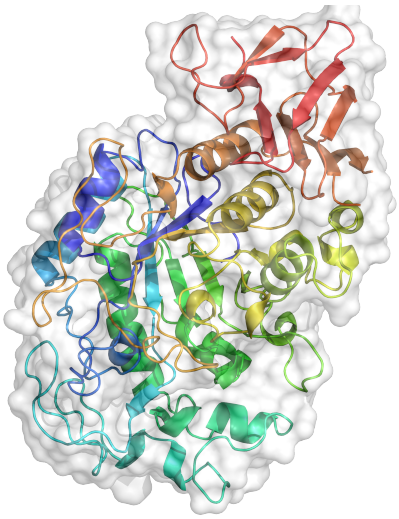


Figura 2.3: Exemplo de estrutura tridimensional: Amilase Salivar Humana.

De acordo com Thornton et al. (2000), a estrutura revela a organização geral da sequência proteica em três dimensões e, a partir disso, pode-se identificar os resíduos que estão enterrados no núcleo proteico ou expostos a solventes na superfície da proteína, a forma e composição molecular da superfície, e a justaposição relativa de grupos individuais. Ela também revela a estrutura quaternária da proteína no ambiente de cristal ou em solução de alta concentração. Complexos proteína-ligante são possivelmente a informação funcional mais útil, visto que mostram a natureza do ligante, o local de sua ligação na proteína e, se tratar-se de uma enzima, a disposição de resíduos no sítio ativo, a partir do qual um mecanismo catalítico pode ser postulado.

Segundo Gherardini e Helmer-Citterich (2008), acredita-se que a disponibilidade de informação estrutural seja de grande ajuda na anotação automática de função por duas razões: i) métodos de comparação estrutural são potencialmente capazes de identificar relacionamentos evolutivos muito distantes, e dados estruturais possibilitam a identificação de sítios funcionais que evoluíram independentemente; e ii) a função depende da estrutura, de modo que a estrutura de uma proteína indica os determinantes mecanísticos de sua função.

Salvo exceções, proteínas desnaturadas geralmente não são funcionais, e mutações que interrompem estrutura e dinâmica gerais muitas vezes têm consequências drásticas tanto para a função quanto para o fenótipo associado à proteína (Sadowski e Jones, 2009; Betts e Russell, 2003). Pe-

quenas alterações na sequência podem ter um impacto profundo na estrutura e, consequentemente, na função. Por exemplo, Melamina Deaminase e Atrazina Clorohidrolase têm 98% de identidade de sequência, mas apresentam funções diferentes (Erdin et al., 2011). Frequentemente, não se sabe se mutações causam doenças, tampouco qual a intensidade do seu efeito na função proteica. A previsão de resíduos em sítios de ligação e sítios ativos é importante para caracterizar funções proteicas e para prever os efeitos de mutações pontuais (Kawabata, 2010). Segundo Betts e Russell (2003), aminoácidos conservados em todos os homólogos provavelmente desempenham papéis estruturais chave ou papéis com um tema funcional comum. Outros aminoácidos podem desempenhar papéis em um determinado grupo de ortólogos (p. ex.: conferindo especificidade de substrato), por isso podem variar quando considera-se todos os homólogos.

Uma grande porção do espaço de sequências é coberta por um número muito menor de enovelamentos proteicos. No entanto, o espaço funcional é mais complexo, incluindo regiões em que muitas sequências relacionadas correspondem a uma única função, regiões em que pequenas mudanças na sequência correspondem a diferenças funcionais importantes, e regiões em que mesmo sequências não-relacionadas convergiram para a mesma função (Devos e Valencia, 2000). Proteínas estão sujeitas a restrições fortes para alcançar estabilidade, o que reflete-se no fato de que os mesmos elementos de esqueleto (hélices, folhas, laços) são reutilizados em todo o espaço de estruturas (Gherardini e Helmer-Citterich, 2008).

Compreender o relacionamento entre função e estrutura é útil para a anotação funcional em larga escala, visto que padrões estruturais são mais conservados que padrões sequenciais (Boareto et al., 2012). Similaridades estruturais fornecem pistas poderosas sobre função bioquímica que podem não ser evidentes a partir da sequência. Além disso, mesmo quando não há ligantes ou homólogos estruturais próximos de uma proteína hipotética, a estrutura tridimensional pode, algumas vezes, sugerir uma ou mais funções celulares e moleculares testáveis (Zhang e Kim, 2003). De fato, há muitos métodos para anotar funções proteicas a partir de estruturas tridimensionais. Segundo Laskowski et al. (2005b), a maioria depende da detecção de similaridade estrutural, seja local ou global, entre a proteína alvo e uma estrutura de função conhecida. Esses métodos baseados em estruturas precisam realizar uma busca por enovelamentos similares, identificar motivos estruturais associados a funções específicas e localizar sítios ativos, além de possíveis interfaces de ligação proteína-proteína ou resíduos mais prováveis de terem papel funcional. Alguns motivos estruturais locais, que podem resultar de evolução tanto divergente quanto convergente, capturam a essência da função bioquímica e podem ser usados para atribuir função (Thornton et al., 2000).

Para Redfern et al. (2009), embora métodos de comparação global de estrutura possam ser utilizados para transferir anotações funcionais, o relacionamento entre enovelamento e função é complexo, particularmente em superfamílias funcionalmente diversas que evoluíram por meio de diferentes alterações a um núcleo estrutural comum. Além disso, segundo Thornton et al. (2000), o mesmo

enovelamento muitas vezes ocorre em famílias diferentes com funções distintas. Foi encontrada baixa correlação entre a função enzimática específica e o enovelamento geral da proteína, o que está de acordo com a observação de que diferentes enovelamentos podem desempenhar a mesma função, algumas vezes com os mesmos grupo catalítico e mecanismo. Vários artigos destacaram a variedade de funções bioquímicas que podem ser realizadas por proteínas com um mesmo enovelamento ou, ainda, por membros de uma única família de proteínas homólogas. Sendo assim, segundo Roy et al. (2012), muitas abordagens de anotação com base em estrutura foram projetadas para identificar similaridade local de resíduos funcionalmente importantes. No entanto, a anotação com base apenas na estrutura local pode resultar em altas taxas de falsos-positivos, especialmente nos casos de baixa identidade de sequência ou de estruturas de baixa resolução.

A comparação do enovelamento proteico ou de motivos estruturais em uma proteína a bases de dados estruturais pode revelar similaridades a partir das quais informações de função bioquímica e biológica podem ser inferidas (Thornton et al., 2000), visto que famílias proteicas com padrões de enovelamento distintos tendem a desempenhar funções diferentes. No entanto, segundo Sadowski e Jones (2009), grande parte da dificuldade para atribuir função a partir de estruturas vem da convergência funcional, que implica na possibilidade de proteínas com estruturas distintas desempenharem a mesma função. Além disso, sequências proteicas diferentes podem adquirir estruturas tridimensionais similares e manter a função, ou seja, estruturas apresentam maior conservação que as sequências correspondentes (Smith, 2012). O processo de divergência funcional, em que proteínas com estruturas semelhantes desempenham funções distintas, é outro complicador para a anotação funcional com base em estruturas. Determinados enovelamentos como, por exemplo, o barril TIM e o enovelamento de Rossman são encontrados em muitas famílias de funções distintas (Kinoshita e Nakamura, 2003). A extrema fluidez do relacionamento estrutura-função proteicos é a principal dificuldade para a utilização de informação estrutural para anotação de função (Sadowski e Jones, 2009). Neste trabalho busca-se contornar tais problemas combinando, além de informações estruturais, diversas outras fontes de informação sobre as proteínas estudadas.

Segundo Lee et al. (2007), embora a estrutura de uma proteína conserve-se mais que a sequência, conhecer o enovelamento específico adotado por uma dada proteína não implica diretamente uma determinada função. O uso de similaridade estrutural para prever função é também problemático devido a artefatos do processo de cristalização. Em alguns casos, quaisquer alterações conformacionais que ocorram durante a ligação do substrato podem causar alterações significativas na estrutura geral. Mesmo estruturas da mesma proteína podem apresentar diferenças significativas quando superpostas. No entanto, dados estruturais podem ser utilizados para detectar proteínas de funções similares mas cujas sequências divergiram ao longo da evolução além de um nível de similaridade que possa ser detectado confiavelmente utilizando métodos de comparação de sequências. Conhecer a estrutura tridimensional de uma proteína pode fornecer uma visão crucial sobre seu modo de ação. No entanto, apenas uma porcentagem muito pequena das sequências tiveram suas estruturas resolvidas

experimentalmente. O número limitado de estruturas proteicas disponíveis é fator restritivo nas tentativas de realizar anotação funcional em larga escala com base em informação estrutural (Norin e Sundström, 2002).

Modelagem comparativa

Dado que a estrutura tridimensional de uma proteína tende a ser mais conservada do que sua sequência e que o número de enovelamentos distintos é limitado (Chothia, 1992), quando uma proteína não possui estrutura determinada experimentalmente, a modelagem comparativa ou por homologia frequentemente fornece um modelo tridimensional útil para uma proteína que seja relacionada a pelo menos uma estrutura conhecida. A modelagem comparativa prevê a estrutura tridimensional de uma dada sequência proteica (alvo) com base principalmente em seu alinhamento com sequências de uma ou mais proteínas de estrutura conhecida (moldes) (Eswar et al., 2006). Tais métodos estão tornando-se cada vez mais confiáveis (Tramontano et al., 2001; Tramontano e Morea, 2003; Moult, 2005).

Segundo Eswar et al. (2006), a modelagem comparativa consiste de quatro passos principais: i) atribuição de enovelamento de proteína, que identifica similaridade entre o alvo e pelo menos uma estrutura molde conhecida; ii) alinhamento da sequência alvo com o(s) respectivos molde(s); iii) construção de um modelo baseado no alinhamento com o(s) molde(s) escolhido(s); e iv) previsão de erros no modelo. Existem vários programas e servidores Web que automatizam a modelagem comparativa. Neste trabalho, utilizamos o programa Modeller (Eswar et al., 2006) para realizar a modelagem de proteínas de interesse cujas estruturas não foram determinadas.

De acordo com Gherardini e Helmer-Citterich (2008), os principais problemas e limitações do uso de estruturas para anotar função de proteínas são a avaliação da significância estatística da similaridade estrutural e o quanto os métodos dependem da precisão e da disponibilidade de estruturas. À medida que mais tipos de enovelamento são caracterizados, a modelagem por homologia de um número crescente de proteínas deve tornar-se não somente possível, mas mais confiável. Sendo assim, proteínas alvos para determinação estrutural são selecionados entre aquelas que apresentam identidade muito baixa com proteínas de estrutura conhecida. Consequentemente, um grande número de estruturas conhecidas pertence a proteínas de função desconhecida.

Sobreposição de estruturas

Muitos casos reportados na literatura evidenciam que a estrutura de uma proteína fornece pistas essenciais para a descoberta de sua função. Na maioria dos exemplos, a chave para a anotação funcional foi o uso de métodos de comparação de enovelamentos (Gherardini e Helmer-Citterich, 2008). Comparações de estruturas proteicas são utilizadas em quase todos os ramos da biologia estrutural, desde classificação de enovelamento e modelagem de estruturas até anotação de função baseada em estrutura. O alinhamento ou sobreposição de estruturas proteicas é frequentemente utilizado para detectar

similaridade funcional, visto que é uma ferramenta valiosa para comparação de proteínas com baixa similaridade sequencial, caso em que relacionamentos evolutivos entre proteínas não podem ser facilmente identificados pelas técnicas de alinhamento de sequências (Zhang e Skolnick, 2005). Análogo ao alinhamento de sequências, o objetivo da sobreposição estrutural é encontrar subestruturas que podem ser sobrepostas de forma a maximizar uma pontuação objetiva (Salem e Zaki, 2009). Devido à alta complexidade e imensa quantidade de combinações possíveis, encontrar a sobreposição estrutural ótima é impraticável. Por isso, algoritmos de sobreposição de estruturas são heurísticos, ou seja, não garantem encontrar a solução ótima, mas buscam uma solução satisfatória.

Duas proteínas que apresentam alta similaridade estrutural ao longo de toda a extensão de suas sequências provavelmente apresentam funções iguais ou semelhantes. Quando analisa-se a significância da similaridade entre duas estruturas, é importante considerar tanto a qualidade da sobreposição quanto o número de resíduos sobrepostos. Uma vantagem de métodos estruturais é que eles geralmente produzem alinhamentos melhores do que os métodos sequenciais quando a identidade cai abaixo de 40% (Lee et al., 2007). No entanto, utilizar a comparação global de estruturas para atribuir função é limitado pelo fato de que pequenas alterações no sítio ativo podem causar divergências funcionais. Resíduos catalíticos podem frequentemente mover-se em relação uns aos outros quando da ligação do substrato, de modo que sua geometria varia entre estruturas com e sem ligantes. O ambiente ao redor do sítio ativo frequentemente apresenta maior similaridade de sequências do que é evidenciado com um alinhamento global.

Sobreposições estruturais globais são úteis para explorar relacionamentos de estruturas recém-resolvidas com enovelamentos e famílias (Roy et al., 2012). Por isso, segundo Gherardini e Helmer-Citterich (2008), são utilizados principalmente para classificar a estrutura proteica e para identificar ligações evolutivas entre homólogos distantes. Ainda segundo esses autores, comparações globais de estruturas podem ser utilizadas para anotação funcional, mas deve-se estar ciente que o relacionamento entre enovelamento e função é extremamente complexo e que são conhecidos vários exemplos de enovelamentos que hospedam uma ampla variedade de funções. A função de uma proteína geralmente depende mais da identidade e localização de alguns poucos resíduos que compõem o sítio ativo do que do enovelamento geral da proteína, por isso a utilidade dos métodos de comparação global de estruturas está em sua capacidade de identificar homólogos remotos.

De acordo com Roy et al. (2012), como alguns enovelamentos são funcionalmente diversos, sua função só pode ser precisamente anotada avaliando a similaridade entre resíduos do sítio ativo envolvido na função. Para tanto, são utilizadas comparações locais de estrutura, que buscam detectar um arranjo tridimensional similar de um pequeno conjunto de resíduos, possivelmente no contexto de estruturas completamente diferentes (Gherardini e Helmer-Citterich, 2008). Em alguns casos, motivos funcionais são conservados ao longo da evolução das proteínas para manter a função mesmo quando a similaridade global de estrutura diminui. Por isso, uma comparação de estrutura de sítios

funcionais pode fornecer um modo mais confiável de anotação funcional de proteínas (Roy et al., 2012). Métodos de sobreposição local de estruturas são úteis para comparar duas estruturas inteiras na busca por similaridades locais sem pressuposições, assim como para buscar, em uma estrutura, por um molde predefinido que represente um arranjo espacial dos resíduos envolvidos em alguma função bioquímica (Gherardini e Helmer-Citterich, 2008).

Segundo Lee et al. (2007), em superfamílias proteicas altamente variáveis, ou seja, aquelas que apresentam divergência estrutural significativa, funções diferentes podem evoluir pela inserção de elementos de estrutura secundária. Esses elementos podem originar-se de diferentes regiões na sequência, mas tendem a estar colocalizados na estrutura de modo a produzir um motivo estrutural maior ou uma característica superficial que modifique a geometria do sítio ativo ou promova interações proteína-proteína diferentes. Como princípio básico, a maioria das superfamílias com alta similaridade estrutural também apresentam alta similaridade funcional.

Neste trabalho, sobreposições estruturais são usados como fonte de evidência de similaridade funcional. Elas são realizadas utilizando o algoritmo TM-Align, que baseia-se na métrica de TM-score. Uma vez que o TM-score atribui maior peso a pares de resíduos com distâncias pequenas do que a pares com distâncias grandes, ele é mais sensível à topologia global de proteínas do que a similaridade estrutural tradicional. O TM-score contabiliza tanto a precisão quanto a cobertura da sobreposição em um único parâmetro. Geralmente, um par de proteínas com TM-score acima de 0,5 possui o mesmo enovelamento, enquanto um par com TM-score menor que 0,3 tem similaridade estrutural aleatória (Roy et al., 2012).

Cavidades

Em nível molecular, muitas funções proteicas podem ser atribuídas a, ou reguladas por, suas interações com ligantes pequenos como metabólitos ou drogas (Gao e Skolnick, 2013). Pode ser observado, em muitas estruturas cristalizadas de proteínas em complexo com seus ligantes, que interações proteína-ligante geralmente ocorrem em locais preferenciais na superfície da proteína conhecidas como cavidades (*pockets*), em contraste com a forma geométrica relativamente plana dos sítios de interação proteína-proteína. Segundo Schmitt et al. (2002), a função de uma proteína, especialmente de enzimas, com frequência está intimamente ligada ao reconhecimento e à modificação química de ligantes endógenos como agonistas, antagonistas, efetores ou substratos. Tal reconhecimento geralmente ocorre em cavidades bem-caracterizadas na superfície proteica. A identificação de sítios de ligação de ligantes em estruturas proteicas é o primeiro passo em muitas análises estruturais, incluindo a previsão de função ou de sítio catalítico, comparações de configurações atômicas de proteínas, e cálculos de docagem (Kawabata, 2010). Uma das metodologias mais simples para isso é buscar cavidades na superfície proteica.

Segundo Lee et al. (2007), uma das principais razões pelas quais enzimas catalisam reações tão eficientemente é que elas são capazes de isolar seus substratos em cavidades ou sítios de ligação, criando assim um ambiente químico único. Passos essenciais de uma reação química requerem um arranjo espacial estritamente definido de determinantes de reconhecimento molecular no sítio ativo enzimático para acomodar e aprisionar espacialmente os substratos (Schmitt et al., 2002). Sítios de ligação com propriedades físico-químicas similares, em conformações tridimensionais comparáveis, podem ser utilizados para identificar funções enzimáticas similares (Lee et al., 2007). Segundo Thornton et al. (2000), sítios ativos enzimáticos estão geralmente em cavidades nítidas na superfície da proteína, e vários autores mostraram que mais de 70% de tais sítios podem ser facilmente reconhecidos apenas encontrando a maior cavidade na superfície da proteína (Schmitt et al., 2002; Laskowski, 1995). Em contraste, sítios de interação proteína-proteína são altamente expostos e apresentam características variáveis.

De acordo com Gao e Skolnick (2013), o espaço estrutural de cavidades proteicas é degenerado e apresenta variabilidade surpreendentemente baixa. Além disso, a promiscuidade tanto de cavidade quanto de ligante é comum. Cavidades com formatos similares podem atrair um conjunto diverso de ligantes com diferentes propriedades químicas por duas razões: i) cavidades com formas similares podem ter diferentes composições de aminoácidos, gerando assim ambientes físico-químicos diversos favorecidos por ligantes quimicamente distintos (p. ex.: homólogos com diferentes especificidades de substrato); e ii) para cavidades grandes, alguns ligantes pequenos podem ligar-se a regiões parcialmente distintas das cavidades, e eles não têm, necessariamente, propriedades químicas similares.

Em geral, mesmo que ligantes tenham esqueletos muito diferentes, eles estão sujeitos às mesmas interações físicas com os mesmos resíduos da cavidade. No entanto, a plasticidade das cavidades proteicas pode permitir diferentes tipos de interações, enquanto a flexibilidade da cadeia lateral permite que sejam formados diferentes tipos de contatos com ligantes distintos. Um determinado ligante pode interagir com várias proteínas cujas estruturas não são globalmente relacionadas, mas que contêm cavidades similares, ou pode interagir com cavidades estruturalmente diferentes devido a alterações conformacionais, o que sugere que existem múltiplas possibilidades de interação entre o ligante e suas cavidades. Adicionalmente, o mesmo confórmero de um ligante pode interagir com cavidades diferentes em poses diferentes (Gao e Skolnick, 2013).

Segundo Betts e Russell (2003), enzimas tendem a ter sítios ativos altamente conservados envolvendo alguns resíduos polares, enquanto proteínas que interagem primariamente com outras proteínas o fazem sobre uma superfície grande, com praticamente qualquer aminoácido sendo importante para mediar a interação. Múltiplas funções tornam a situação ainda mais confusa como, por exemplo, Proteínas Cinases, que podem funcionar como enzimas ou ligar-se especificamente a outras proteínas. Segundo Gherardini e Helmer-Citterich (2008), além de localizadas em cavidades, vários autores relataram que resíduos de sítio ativo estão próximos do centroide da estrutura, têm efeito desestabili-

zante na mesma, interagem com um grande número de resíduos da própria proteína, têm valores de pKa perturbados, e induzem picos de potencial eletrostático ao redor da proteína.

Para detectar cavidades, são necessários métodos para identificar e delimitar depressões na superfície proteica às quais compostos pequenos provavelmente ligam-se. Em geral, várias cavidades são detectadas, de modo que é necessário realizar uma caracterização das mesmas para selecionar as que são potenciais sítios de ligação. Embora a maior cavidade tenda a corresponder ao sítio de ligação observado, essa regra não pode ser generalizada (Guilloux et al., 2009). Algumas abordagens para identificar cavidades são baseadas na análise geométrica da superfície proteica, enquanto outras envolvem cálculos de energia.

Neste trabalho, é utilizado o Fpocket (Guilloux et al., 2009), um pacote de detecção de cavidades de código aberto que utiliza o conceito de α-esferas: esferas cujas superfícies fazem contato com quatro átomos, sem átomos em seu interior. Segundo os desenvolvedores do método, os raios das α-esferas refletem a curvatura local definida pelos quatro átomos. Esferas muito pequenas estão localizadas no interior da proteína, esferas grandes, em seu exterior, e esferas com raios intermediários correspondem a cavidades e fendas. Regiões de interesse como cavidades na superfície proteica têm maior ocorrência de α-esferas. Portanto, a procura por cavidades de ligação pode ser vista como a busca por grupos de α-esferas de raios apropriados.

Segundo Guilloux et al. (2009), o Fpocket tem três passos principais: i) determinação de todo o conjunto de α-esferas da estrutura proteica e definição de uma coleção pré-filtrada de esferas; ii) detecção de grupos de esferas vizinhas, identificação de cavidades e remoção de grupos desinteressantes; e iii) pontuação de cavidades com base no cálculo de propriedades atômicas. Para tanto, é realizada uma tesselação de Voronoi do conjunto de átomos pesados, obtendo-se um conjunto de vértices de Voronoi que é podado de acordo com o tamanho máximo e mínimo das α-esferas, o que elimina esferas inacessíveis ao solvente ou superexpostas. Apenas esferas definidas por zonas de empacotamento de átomos são mantidas e rotuladas de acordo com os tipos de átomos com os quais têm contato. Um primeiro conjunto de grupos de esferas é identificado utilizando um critério de distância simples. Todos os grupos contendo uma única esfera são removidos e os centros de massa dos grupos remanescentes são calculados. Então, grupos contendo centros de massa próximos são agregados e, finalmente, uma abordagem de ligação múltipla é aplicada para realizar o agrupamento final de α-esferas. Os grupos finais são caracterizados de modo a classificar cavidades de acordo com sua capacidade de ligar pequenas moléculas. Para isso é utilizada uma função de ponderação simples que envolve cobertura de ligantes, número de α-esferas, média de densidade hidrofóbica local, proporção de esferas apolares, pontuação de polaridade e densidade de esferas.

Uma vez que o sítio ativo é essencial para a função de uma enzima, neste trabalho, um dos principais indicadores de similaridade funcional entre proteínas é a semelhança em sua composição. Cada grupo encontrado pelo algoritmo de agrupamento é descrito por um padrão de composição

de resíduos, como será detalhado no Capítulo 5. Por tratar-se de famílias proteicas, há resíduos conservados por toda a família. Assim, se buscará neste trabalho grupos que apresentem resíduos específicos, que são possíveis determinantes de especificidade das subfamílias.

2.9 Propriedades de Proteínas

As seguintes propriedades são específicas de cada proteína, por isso, neste trabalho, são utilizadas as diferenças de valores das mesmas a fim de comparar pares de proteínas. Considera-se que a ocorrência de valores próximos para duas proteínas é uma possível evidência de similaridade entre elas.

Ponto Isoelétrico (pI): pH no qual a carga elétrica líquida da proteína é nula, apresentando carga positiva (negativa) em valores de pH abaixo (acima) do pI. Pontos isoelétricos refletem a natureza das cadeias laterais ionizantes e podem ser utilizados para separação de proteínas: a aplicação de uma voltagem a uma mistura proteica em um gradiente de pH provoca a migração das proteínas até os valores de pH em que apresentam carga neutra e a migração cessa. O pI pode ser estimado a partir da sequência de resíduos da proteína e fornece informação complementar a métodos de espectrometria de massa para identificação proteica.

Peso Molecular: razão da massa da proteína para $1/12$ da massa do carbono-12. Utilizado na cromatografia e eletroforese, por exemplo, para isolar proteínas. No entanto, não é possível fazer generalizações a seu respeito em relação a funções proteicas (Nelson e Cox, 2005).

Índice de Hidropaticidade (GRAVY, do inglês *grand average of hydropathicity*): soma dos valores de hidropaticidade dos aminoácidos, dividida pelo número de resíduos. Indica solubilidade, com valores positivos (negativos) para proteínas hidrofóbicas (hidrofílicas).

Índice de Instabilidade: fornece uma estimativa da estabilidade da proteína em um tubo de ensaio. Uma proteína cujo índice de instabilidade é inferior a quarenta é considerada estável.

Estatística Dayhoff: calculada para cada tipo de aminoácido presente na proteína. Corresponde à porcentagem molar do aminoácido dividida pela estatística Dayhoff, que, por sua vez, concerne sua ocorrência relativa a cada mil aminoácidos. Segundo Dobson e Doig (2005), a composição de aminoácidos é uma informação surpreendentemente relevante para a anotação de função, tendo sido usada em técnicas de aprendizado de máquina. Neste trabalho, a estatística Dayhoff é utilizada como métrica da composição de aminoácidos, sendo considerada importante porque, quando completamente hidrolisada, cada tipo de proteína tem uma proporção ou mistura característica dos diferentes aminoácidos. Segundo Nelson e Cox (2005), os vinte aminoácidos mais comuns quase nunca ocorrem em iguais quantidades em uma proteína: alguns podem ocorrer no máximo uma vez em um dado tipo de proteína, enquanto outros podem ocorrer várias vezes. Os aminoácidos menores predominam na maioria das proteínas.

2.10 Bases de Dados Utilizadas

Esta seção descreve brevemente as bases de dados públicas utilizadas neste trabalho.

Universal Protein Resource Knowledgebase

O UniProtKB[1] é um repositório central de sequências proteicas formado pela junção dos bancos de dados Swiss-Prot, TrEMBL e *Protein Information Resource* (PIR). Trata-se de um banco de dados não-redundante que concentra informações incorporadas à sequência de proteínas por meio de programas de anotação automática e de previsão de estrutura e de domínios. Para algumas proteínas, é realizada uma anotação manual detalhada, curada por especialistas (Zaha et al., 2012). O UniProt utiliza um vocabulário controlado baseado em palavras-chave para anotação e recuperação rápidas dos dados funcionais associados às suas sequências proteicas.

PROSITE

O PROSITE[2] é um banco de dados de famílias e domínios proteicos baseado na observação de que, apesar de haver um número imenso de diferentes proteínas, a maioria delas pode ser agrupada em um número limitado de famílias com base em similaridades de sequência. Proteínas ou domínios proteicos pertencentes a uma mesma família geralmente compartilham atributos funcionais e derivam de um ancestral comum (Sigrist et al., 2010). A ideia é que famílias proteicas podem ser simplesmente caracterizadas pelo seu motivo mais conservado, obtido pelo alinhamento múltiplo de sequências. Tais motivos geralmente estão associados às respectivas funções biológicas (Zaha et al., 2012). Analisando as propriedades constantes e variáveis em sequências de famílias proteicas é possível derivar uma assinatura para uma família ou domínio que distingue seus membros de outras proteínas (Sigrist et al., 2010). Sendo assim, o PROSITE consiste de sítios, padrões e perfis biologicamente significativos que ajudam a identificar confiavelmente a quais famílias proteicas conhecidas uma nova sequência pertence.

Pfam

O Pfam[3] é um banco de dados de famílias de sequências de proteínas em que famílias são conjuntos de regiões proteicas com alto grau de similaridade de sequência, sugerindo, então, homologia. Cada família é representada por um modelo estatístico conhecido como perfil HMM, um modelo probabilístico utilizado para a inferência estatística de homologia construído a partir de um alinhamento de sequências representativas definido por um curador. Busca-se o perfil HMM em uma grande coleção

[1] `http://www.uniprot.org/`
[2] `http://prosite.expasy.org/prosite.html`
[3] `http://pfam.sanger.ac.uk/`

de sequências baseada no UniProt, de modo a encontrar todas as instâncias de uma família. Para eliminar falsos positivos, é definido um limiar de pontuação mínima para cada família. Regiões de sequência com pontuação acima desse limiar são alinhadas ao perfil HMM para produzir o alinhamento completo da família. Haviam 14.831 entradas manualmente curadas na versão 27 do Pfam (Finn et al., 2014), com 1.182 novas famílias desde a versão 26 (Punta et al., 2012), mantendo a cobertura de 80% das sequências do UniProt apesar do aumento de 50% no seu tamanho.

Repositórios de famílias proteicas facilitam medir a confiabilidade da herança funcional por homologia, uma vez que organizam os homólogos putativos de uma maneira mais informativa do que seria obtido pela busca de uma sequência contra um banco de dados de sequências não-estruturado. Sequências multi-domínio ou domínios proteicos individuais são agrupados em famílias evolutivas putativas (Lee et al., 2007). Uma vez que proteínas homólogas têm maior probabilidade de compartilhar características estruturais e funcionais, famílias Pfam podem auxiliar na anotação de sequências não-caracterizadas e guiar trabalho experimental. Segundo Punta et al. (2012), espera-se que membros da mesma família Pfam compartilhem uma história evolutiva e tenham pelo menos algum aspecto funcional em comum. Idealmente, as famílias deveriam representar unidades funcionais que, quando combinadas de diferentes modos, podem gerar proteínas com funções únicas. No entanto, homologia não é garantia de similaridade funcional, por isso deve-se ter cuidado ao transferir anotações funcionais com base somente na pertinência a uma mesma família. Em várias das famílias mais populosas, a função pode divergir consideravelmente entre parálogos. Dados adicionais disponíveis no Pfam, como a conservação dos resíduos de assinatura da família ou uma arquitetura de domínios comum, podem aumentar a confiança em uma determinada hipótese funcional.

Segundo Lee et al. (2007), muitos dos repositórios baseados em famílias proteicas estão buscando superar o problema de diversidade funcional por meio da identificação de subgrupos ou subfamílias com funções mais específicas, o que é o propósito do presente trabalho. Ainda segundo os autores, o sucesso de um repositório geralmente pode ser aumentado por meio da identificação de resíduos específicos que discriminam entre funções. Resíduos funcionalmente ativos são aqueles que têm maior probabilidade de terem sido conservados ao longo da evolução.

InterPro

O InterPro[4] é um recurso integrado de documentação para famílias, domínios, regiões e sítios de proteínas. Compreende vários bancos de dados biológicos, incluindo PROSITE e Pfam, anteriormente descritos. Fornece análise funcional de sequências proteicas classificando-as em famílias e prevendo a presença de domínios e sítios importantes (Hunter et al., 2012). Seu objetivo é a obtenção de assinaturas de proteínas, por isso combina uma série de bancos de dados que utilizam diferentes metodologias e graus de informações biológicas sobre proteínas bem caracterizadas (Zaha et al., 2012).

[4] `http://www.ebi.ac.uk/interpro/`

Segundo Finn et al. (2010), proteínas são geralmente compostas por uma ou mais regiões funcionais (domínios), com diferentes combinações de domínios originando a ampla gama de proteínas encontradas na natureza. Dessa forma, a identificação dos domínios que ocorrem em uma proteína pode fornecer pistas sobre sua função. De fato, segundo Deng et al. (2004), várias pesquisas mostraram que domínios proteicos são uma característica informativa para a anotação funcional. De acordo com Smith (2012), famílias de genes ou proteínas podem ser definidas com base na identificação dos domínios presentes em sequências ou estruturas, uma vez que pode-se esperar que proteínas com domínios ou arquiteturas de domínio similares sejam funcionalmente relacionadas e formem famílias de proteínas ou genes baseadas em domínios. Neste trabalho, quanto mais anotações duas proteínas têm em comum no InterPro, maior a evidência de similaridade entre elas.

Protein Data Bank

O PDB[5], um compêndio contendo estruturas de proteínas experimentalmente resolvidas, é um repositório de coordenadas atômicas e informações que descrevem proteínas e outras macromoléculas biológicas importantes. Cada proteína possui, além das posições de seus átomos no espaço tridimensional, um resumo contendo, por exemplo, a descrição da resolução da estrutura, do número de cadeias da proteína, de ligantes e íons metálicos, e de estruturas secundárias (Berman et al., 2000). Neste trabalho, as estruturas tridimensionais das proteínas são coletadas do PDB.

Gene Ontology

Diferentes tipos de anotação, origens profissionais e preferências pessoais moldam a forma como anotações são construídas, levando a amplas variações de terminologia que afetam negativamente o armazenamento de informação e sua recuperação por sistemas automatizados (Smith, 2012). Ontologias buscam tratar esse problema implementando um vocabulário controlado. Elas restringem descrições a termos específicos com ortografia uniforme, de modo que similaridades entre conjuntos de termos de anotação atribuídos a proteínas diferentes podem ser automaticamente quantificadas utilizando uma métrica de similaridade semântica que explica o conteúdo de informação de diferentes anotações (Lee et al., 2007).

O *Gene Ontology* (GO) [6] é um projeto que visa tornar consistentes as descrições de produtos gênicos em bases de dados diferentes, sendo uma fonte abrangente de anotações gênicas. Há três categorias ontológicas controladas que descrevem os produtos gênicos em termos de processos biológicos, componentes celulares e funções moleculares, associados de modo independente de espécie (Radivojac et al., 2013; Lee et al., 2007). Cada uma das três categorias do GO é um conjunto hierárquico de termos e relacionamentos entre eles que capturam informação funcional (Radivojac et al.,

[5] `http://www.pdb.org`
[6] `http://geneontology.org/`

2013). A ontologia *processo biológico* refere-se ao objetivo biológico para o qual o gene ou produto gênico contribui. Um processo é realizado via uma ou mais sequências ordenadas de funções moleculares. A ontologia *função molecular* define a atividade bioquímica (incluindo a ligação específica a ligantes ou estruturas) de um produto gênico, descrevendo apenas o que é feito sem especificar onde ou quando o evento ocorre. Já a ontologia *componente celular* refere-se à localização celular onde o produto gênico está ativo. Cada uma das três ontologias é atribuída independentemente das outras, visto que uma dada proteína pode funcionar em vários processos, conter domínios que desempenham diversas funções moleculares e participar em múltiplas interações alternativas com outras proteínas, organelas ou localizações celulares (Ashburner et al., 2000).

Neste trabalho, o GO é utilizado de modo semelhante aos dados do InterPro: quanto mais termos duas proteínas têm em comum, maior é a evidência de similaridade funcional entre elas.

Search Tool for the Retrieval of Interacting Genes/Proteins

O STRING[7] é uma base de dados que fornece uma coleção abrangente de associações entre proteínas para um grande número de organismos. As associações são derivadas de dados experimentais de alto rendimento, da mineração de bases de dados e literatura, e de previsões baseadas na análise de contexto genômico (von Mering et al., 2005). Segundo Smith (2012), o STRING tornou-se referência em informação de dependência funcional pré-calculada.

Três tipos de associações de contexto genômico são buscados, em genomas completos, pelo STRING: vizinhança genômica conservada, eventos de fusão de genes e coocorrência de genes em genomas diferentes. Segundo von Mering et al. (2005), essas buscas visam a identificar pares de genes que aparentam estar sob pressões seletivas comuns durante a evolução e que, portanto, são consideradas funcionalmente associadas. Por isso, os três tipos de associações são interpretadas neste trabalho como evidências de similaridade funcional.

[7] `http://string-db.org/`

Capítulo 3
Ferramentas Computacionais

Neste capítulo são descritos os algoritmos e técnicas empregados na realização deste trabalho.

3.1 Agrupamento

Agrupamento é uma técnica de mineração de dados que divide objetos em grupos por similaridade, sendo o método mais comum de análise de dados automática e não-supervisionada (Ares et al., 2012). Em praticamente todos os campos científicos que lidam com dados empíricos, procura-se ter uma primeira impressão dos dados buscando identificar grupos de "comportamento similar" (von Luxburg, 2007). Diferentemente de métodos de classificação, o agrupamento não requer qualquer informação de supervisão (p. ex.: rótulos de classe). A técnica consiste no processo de particionar um conjunto de objetos em grupos naturais, cada qual representando uma subpopulação significativa, de forma que aqueles em um mesmo grupo tenham alta similaridade entre si, enquanto objetos pertencentes a grupos distintos sejam o mais dissimilares possível (Zaki e Meira Jr., 2014). Objetos podem ser registros de um banco de dados, nós de um grafo, palavras, imagens ou qualquer coleção de indivíduos descritos por um conjunto de atributos ou relacionamentos (Basu et al., 2008). No caso deste trabalho, objetos são proteínas pertencentes a uma mesma família. Segundo Han e Kamber (2006), o agrupamento pode ser considerado uma forma de compressão de dados, visto que objetos em um grupo podem ser tratados coletivamente. Isso é justamente o que aqui busca-se: dividir uma família de proteínas em grupos que representem possíveis subfamílias isofuncionais para tratamento posterior.

Por ser adaptável a mudanças e ajudar a isolar atributos que diferenciam grupos, algoritmos de agrupamento têm sido utilizados em diversas aplicações tais como pesquisa de mercado, reconhecimento de padrões, análise de dados e processamento de imagens. Segundo Han e Kamber (2006), métodos de agrupamento podem ser utilizados como ferramentas autônomas para compreensão da distribuição dos dados, para observação das características de cada grupo ou para isolamento de um determinado conjunto de grupos para análise posterior. Além disso, agrupamento pode ser empregado como passo de pré-processamento para algoritmos de caracterização, seleção de atributos e classificação, os quais operariam nos grupos detectados (Ares et al., 2012).

Neste trabalho são empregados métodos de agrupamento baseados em particionamento, cuja ideia fundamental é: dado um conjunto de dados com N objetos, construir K partições dos dados

($K \leq N$), cada qual representando um grupo não-vazio e com cada objeto pertencendo a um único grupo. Dado o número K de partições a serem construídas, o método cria um particionamento inicial e usa uma técnica iterativa de deslocamento entre grupos que busca melhorar o particionamento, movendo objetos de um grupo para outro. Em geral, um agrupamento é considerado bom se objetos em um mesmo grupo são relacionados entre si e são, ao mesmo tempo, distintos daqueles pertencentes a outros grupos (Han e Kamber, 2006).

Para encontrar o particionamento ótimo global seria necessária uma enumeração exaustiva de todas as possíveis partições, o que tem custo computacional proibitivo. Por isso, segundo Han e Kamber (2006), a maioria das aplicações adotam um dos dois métodos heurísticos mais populares:

- ◇ K-Médias, em que cada grupo é representado pelo valor médio dos objetos nele contidos; ou
- ◇ K-Medoides, em que um dos objetos próximos ao centro do grupo é o representante do mesmo.

K-Médias

O K-Médias é um algoritmo do tipo "guloso" que minimiza o erro quadrado entre objetos e a média de seus respectivos grupos (Zaki e Meira Jr., 2014). O Algoritmo 3.1 descreve seu funcionamento. Inicialmente, K dos N objetos são sorteados para representar centros de grupos e cada objeto restante é atribuído ao grupo ao qual mais assemelha-se, ou seja, aquele cujo centro está à menor distância. O algoritmo então atualiza os centros calculando a média dos objetos pertencentes a cada grupo, e o processo itera até que os grupos estabilizem (convergência) ou que um critério de parada seja atingido. Devido ao uso da média, o K-Médias é sensível a valores discrepantes, visto que um objeto com valor extremamente alto, por exemplo, pode distorcer substancialmente a distribuição dos dados.

Algoritmo 3.1: Algoritmo de agrupamento K-Médias. Fonte: Han e Kamber (2006)

Entrada: número de grupos K e conjunto de dados D contendo N objetos
Saída: conjunto de K grupos

```
1 sortear K objetos de D como centros iniciais dos grupos
2 repetir
3 |   atribuir cada objeto ao grupo a cuja média ele está mais próximo
4 |   atualizar as médias dos objetos pertencentes a cada grupo
5 até convergir ou atingir um critério de parada
```

K-Medoides

Ao invés de utilizar o valor médio dos objetos de um grupo como ponto de referência, cada grupo pode ser representado por um dos objetos nele contidos. Assim, cada objeto não-representante é atribuído ao grupo cujo representante lhe é mais similar. O método de particionamento é então realizado

com base no princípio de minimização da soma das dissimilaridades entre cada objeto e seu ponto de referência correspondente. Em geral, o algoritmo itera até que cada objeto representante seja o medoide do seu grupo, isto é, o objeto mais centralmente localizado (Han e Kamber, 2006). Essa é a base do método de agrupamento K-Medoides, apresentado no Algoritmo 3.2.

Algoritmo 3.2: Algoritmo de agrupamento K-Medoides. Fonte: Han e Kamber (2006).

Entrada: número de grupos K e conjunto de dados D contendo N objetos
Saída: conjunto de K grupos

```
1 sortear K objetos como representantes iniciais dos grupos
2 repetir
3     atribuir cada objeto ao grupo cujo representante está mais próximo
4     sortear um objeto não representante o_rand
5     calcular o custo total S de trocar o representante o_j por o_rand
6     se S < 0 então
7         trocar o_j por o_rand formando o novo conjunto de K representantes
8 até convergir ou atingir um critério de parada
```

Assim como no algoritmo K-Médias, os objetos representantes iniciais são escolhidos aleatoriamente. O processo iterativo de substituição de representantes por objetos não-representantes continua enquanto a qualidade do agrupamento resultante aumentar, ou seja, enquanto a soma de dissimilaridades entre objetos e seus respectivos representantes diminuir.

Esses métodos heurísticos funcionam bem para encontrar grupos com formatos esféricos em bases de dados pequenas ou médias. No entanto, métodos de agrupamento baseados em particionamento precisam ser estendidos para tornarem-se aplicáveis a bases de dados maiores e para que possam encontrar grupos com formatos complexos. Uma forma de fazer isso é empregar agrupamento espectral, como utilizado neste trabalho.

Agrupamento espectral

Agrupamento espectral refere-se a uma classe de técnicas que dependem da auto-estrutura de uma matriz de similaridades para particionar um conjunto de objetos em grupos disjuntos (Bach e Jordan, 2003). O agrupamento espectral tornou-se um dos algoritmos de agrupamento modernos mais populares por apresentar muitas vantagens fundamentais em relação aos algoritmos tradicionais como o K-Médias. Ele usa o espectro da matriz de similaridades entre objetos, ou seja, seus autovalores, para reduzir a dimensionalidade antes de efetuar o agrupamento em um número menor de dimensões. Sua implementação é simples e pode ser eficientemente solucionado por meio de programas padrões de álgebra linear. Frequentemente, o agrupamento espectral tem desempenho melhor do que algoritmos tradicionais. No entanto, segundo von Luxburg (2007), o motivo pelo qual funciona e o que realmente o agrupamento espectral faz ainda não são óbvios.

Dois objetos matemáticos são usados no agrupamento espectral: grafos de similaridade e suas matrizes Laplacianas. Grafos de similaridade são uma forma simples de representar os dados quando não se tem informações além das similaridades entre objetos. Cada vértice v_i no grafo representa um objeto x_i, e dois vértices são conectados se a similaridade s_{ij} entre os objetos correspondentes for positiva ou maior que um determinado limiar de interesse, sendo s_{ij} o peso da aresta entre eles. O problema de agrupamento é reformulado para a ideia de encontrar partições do grafo de similaridades de forma que arestas entre grupos diferentes tenham pesos baixos, enquanto arestas dentro de um grupo tenham pesos altos (von Luxburg, 2007).

A partir do grafo de similaridades, calcula-se a matriz Laplaciana e seus autovalores e autovetores, tomando-se os primeiros K autovetores como dimensões na nova representação do conjunto de dados. Segundo von Luxburg (2007), a principal vantagem do agrupamento espectral é a alteração na representação dos objetos x_i no espaço original para pontos y_i no espaço K-dimensional. Essa mudança de representação acentua as propriedades de grupo presentes nos dados, de forma que grupos possam ser detectados trivialmente na nova representação, podendo ser utilizado, após a redução de dimensionalidade, um algoritmo de agrupamento simples como o K-Médias.

Construção do grafo de similaridades

Diferentes métodos existem para transformar um conjunto de objetos com distâncias ou similaridades par-a-par em um grafo. Todos têm o mesmo objetivo de modelar as relações de vizinhança locais que existem entre os objetos. Segundo von Luxburg (2007), não existem resultados teóricos sobre a forma como a escolha do grafo de similaridades usado afeta o resultado do agrupamento espectral.

O grafo de similaridades considerado neste trabalho é o totalmente conectado, que consiste simplesmente em conectar todos os pares de vértices cujas similaridades sejam positivas e ponderar cada aresta pela respectiva similaridade s_{ij}. Uma vez que supõe-se que o grafo represente vizinhanças locais, o grafo totalmente conectado somente é útil se a função de similaridade propriamente dita modelar vizinhanças locais (von Luxburg, 2007).

Outro grafo de similaridades bastante utilizado é o dos K vizinhos mais próximos (KNN, do inglês *K-Nearest Neighbors*), em que um vértice v_j é conectado ao vértice v_i se ele está entre os K vizinhos mais próximos de v_i, ou seja, se a aresta entre eles está entre as K arestas com os maiores pesos envolvendo v_i. Isso resulta em um grafo direcionado, que pode ser transformado em não-direcionado seguindo duas estratégias: ignorar as direções das arestas, o que leva ao grafo KNN; ou somente conectar v_i e v_j se eles estiverem entre os K vizinhos mais próximos um do outro, o que gera o grafo KNN mútuo. Em ambos os casos, arestas são ponderadas pela similaridade de seus vértices, mas o grafo KNN tem muito mais arestas que o KNN mútuo para o mesmo valor de K.

Segundo von Luxburg (2007), o grafo KNN pode conectar objetos em escalas diferentes, ou seja, objetos em regiões de baixa densidade podem ser conectados a outros em regiões de alta densi-

dade, o que pode ser bastante útil para algumas aplicações. O grafo KNN pode quebrar-se em vários componentes desconectados caso hajam regiões de alta densidade razoavelmente distantes umas das outras. Já o grafo KNN mútuo tende a conectar objetos dentro da mesma região de densidade, mas não conecta regiões com densidades distintas. Ele pode agir em diferentes escalas, mas não as mistura, por isso é particularmente adequado caso se queira detectar grupos de diferentes densidades. Isso é bom caso existam grupos claros induzidos por áreas separadas de alta densidade, mas pode ser ruim em situações menos óbvias, visto que partes desconexas do grafo sempre serão escolhidas como grupos. Ainda segundo o autor, a literatura é escassa em relação a resultados teóricos para guiar a escolha do valor de K.

O agrupamento espectral pode ser interpretado como uma busca por um particionamento do grafo de forma que um caminhamento aleatório permaneça bastante tempo dentro de um mesmo grupo e quase não desloque-se entre grupos. Em geral, caso o grafo de similaridades contenha mais componentes conexos do que o número de grupos que o algoritmo deve encontrar, o agrupamento espectral irá, trivialmente, retornar os componentes conexos como grupos. Sendo assim, a menos que se tenha certeza que os componentes conexos são os grupos corretos, é preciso certificar-se de que o grafo de similaridades seja conexo ou consista de apenas alguns componentes conexos e bem poucos vértices isolados (von Luxburg, 2007). Essa é a razão pela qual utiliza-se, neste trabalho, o grafo totalmente conectado. Segundo von Luxburg (2007), para o grafo totalmente conectado, é preciso assegurar que, para a maioria dos objetos, o conjunto de vizinhos com similaridade significativamente maior que zero não é nem muito pequeno e nem muito grande.

Matrizes Laplacianas

Matrizes Laplacianas de grafos são ferramentas principais para o agrupamento espectral mas, segundo von Luxburg (2007), não há uma convenção única na literatura sobre qual matriz é chamada de "Laplaciana do grafo". Considerando G um grafo ponderado não-direcionado com uma matriz de adjacências A simétrica e não-negativa, o grau d_i do vértice v_i corresponde à soma dos pesos das arestas que envolvem v_i, de modo que a matriz diagonal de graus D é formada por $d_i, \ldots, d_N$. Há dois tipos de matrizes Laplacianas: a não-normalizada, definida por $L = D - A$, e a normalizada, que pode ser simétrica, calculada pela Equação 3.1, ou assimétrica, que é aproximadamente equivalente a um caminhamento aleatório (*random walk*) no grafo, calculada pela Equação 3.2.

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2} \tag{3.1}$$

$$L_{rw} = D^{-1} L = I - D^{-1} A \tag{3.2}$$

Detalhes sobre as propriedades de cada matriz Laplaciana podem ser encontrados nos textos de von Luxburg (2007) e de Zaki e Meira Jr. (2014). Segundo von Luxburg (2007), há vários argu-

mentos a favor do uso das matrizes Laplacianas normalizadas ao invés da não-normalizada, visto que pode-se provar que o agrupamento espectral não-normalizado pode simplesmente não convergir ou pode convergir para soluções triviais, com grupos consistindo de um único objeto. No caso das normalizadas, há argumentos a favor do uso dos autovetores da matriz assimétrica L_{rw} ao invés dos de L_{sym}, visto que os primeiros são vetores indicadores de grupo, enquanto os segundos são adicionalmente multiplicados por $D^{1/2}$, o que pode levar a efeitos indesejados, além de não possuir vantagens computacionais. Por isso, neste trabalho é utilizada a matriz Laplaciana normalizada assimétrica L_{rw}.

Existem diferentes algoritmos de agrupamento espectral, cada um utilizando uma das matrizes Laplacianas do grafo de similaridade. O Algoritmo 3.3 é aplicado quando utiliza-se a matriz Laplaciana normalizada assimétrica L_{rw}, como feito neste trabalho. Os autovalores da matriz Laplaciana são sempre ordenados de forma crescente, respeitando multiplicidades. Sendo assim, no algoritmo, por "primeiros K autovetores" entende-se os autovetores correspondentes aos K menores autovalores. A multiplicidade do autovalor zero da matriz Laplaciana normalizada está relacionada ao número de componentes conexos do grafo (von Luxburg, 2007; Zaki e Meira Jr., 2014).

Algoritmo 3.3: Algoritmo de agrupamento espectral normalizado assimétrico. Fonte: von Luxburg (2007)

Entrada: número de grupos K e matriz de similaridades $S \in \Re^{N \times N}$
Saída: grupos $A_1, \ldots, A_K$ com $A_i = \{j | y_j \in C_j\}$

1 construir um grafo de similaridades usando um dos métodos de construção e definir A como sua matriz de adjacências
2 calcular a matriz Laplaciana não-normalizada L
3 calcular a matriz Laplaciana normalizada assimétrica L_{rw}
4 calcular os primeiros K autovetores $u_i, \ldots, u_K$ de L_{rw}
5 seja $U \in \Re^{N \times K}$ a matriz contendo os autovetores $u_i, \ldots, u_K$ como colunas
6 **para cada** $i = 1 \ldots N$ **faça**
7 seja $y_i \in \Re^K$ o vetor correspondente à i-ésima linha de U
8 agrupar os pontos $(y_i)_{i=1\ldots N}$ em $\Re^K$ com o K-Médias, gerando os grupos $C_1, \ldots, C_K$

Segundo von Luxburg (2007), o sucesso do agrupamento espectral é baseado principalmente no fato de que ele não pressupõe o formato dos grupos. Diferente do K-Médias, cujos grupos resultantes têm forma convexa, o agrupamento espectral pode resolver problemas muito gerais como o caso de grupos em espirais entrelaçadas. Além disso, o agrupamento espectral pode ser implementado eficientemente, mesmo para grandes conjuntos de dados, desde que assegure-se que o grafo de similaridades seja esparso.

3.2 Programação Genética

A programação genética (GP, do inglês *Genetic Programming*) é uma técnica de computação natural que soluciona problemas automaticamente, sem que o usuário precise conhecer ou especificar previamente a forma ou estrutura da solução. Consiste de um sistema de aprendizagem adaptativa baseado em muitos dos princípios de algoritmos genéticos (GA, do inglês *Genetic Algorithms*). Basicamente, evolui-se uma população de indivíduos, cada qual representando um programa, transformando estocasticamente uma população em outra presumivelmente melhor a cada geração. Segundo Poli et al. (2008), como toda heurística, a programação genética não garante bons resultados, mas sua aleatoriedade essencial pode tirá-la de armadilhas nas quais métodos determinísticos podem cair, e vem sendo muito bem sucedida na evolução de modos novos e inesperados de solucionar problemas. O Algoritmo 3.4 mostra os passos básicos de um sistema de programação genética.

Algoritmo 3.4: Algoritmo básico de programação genética. Fonte: Poli et al. (2008)

Entrada: conjunto de primitivas
Saída: melhor indivíduo

1 criar uma população inicial de programas a partir das primitivas disponíveis
2 **repetir**
3 executar cada programa (indivíduo) e obter seu valor de aptidão
4 selecionar um ou dois programas da população com probabilidade baseada em aptidão para participar das operações genéticas
5 criar novos programas aplicando operações genéticas com as probabilidades especificadas
6 **até** *encontrar uma solução aceitável ou outro critério de parada for atingido*

Tanto algoritmos genéticos quanto programação genética mantêm uma multitude de soluções independentes, representadas como indivíduos em uma população. Ambos executam ciclos chamados de gerações, nos quais membros da população atual são copiados para uma nova população, removidos, ou modificados antes de serem inseridos na nova população. Ambos utilizam um conjunto de operadores genéticos para modificar indivíduos da população, além de uma operação de seleção para determinar quais indivíduos serão movidos à próxima geração com base na aptidão dos mesmos, tipicamente medida por meio de uma função de avaliação externa.

Modelagem de indivíduos

Conforme discutido por Zongker e Punch (1996), há diferenças básicas importantes entre indivíduos de programação genética e de algoritmos genéticos:

- um indivíduo da população é representado na GP por uma árvore, enquanto a maioria das aplicações de GA utiliza uma cadeia de caracteres;

- ◇ os nós na árvore da GP são tipicamente funções ou terminais, o que permite que cada árvore seja interpretada como um programa;
- ◇ enquanto o tamanho da cadeia de caracteres do GA geralmente é fixo, o tamanho do indivíduo da GP é intrinsecamente variável.

Para definir uma aplicação de GP, deve-se fornecer os conjuntos de funções e terminais a partir dos quais as árvores são construídas. Nós terminais são os nós-folhas da árvore e representam funções que não recebem argumentos, ou seja, constantes ou variáveis. Nós funcionais são os nós internos da árvore e representam uma função cujos argumentos são obtidos pela avaliação de suas subárvores. O conjunto de *primitivas* é formado pelos terminais e funções permitidos no sistema (Poli et al., 2008).

No exemplo da Figura 3.1, as variáveis e constantes do programa (x, y, 3) são folhas da árvore, ou seja, seu conjunto de terminais. Já os operadores aritméticos ($+$, $\times$, $\max$) são nós internos (funções). Como a raiz da árvore contém a função $\max$, essa árvore representa a busca do valor máximo entre seus dois ramos: $x + x$ e $x + 3 * y$.

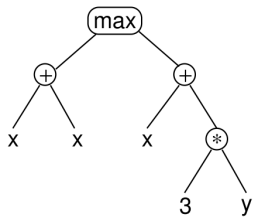


Figura 3.1: Árvore de sintaxe de um indivíduo de GP representando a função $\max(2x, x + 3y)$.

Funções de aptidão

Assim como em GAs, o sistema de GP descobre quão bom um indivíduo (programa) é comparando-o a algum comportamento ideal. Tal comparação é quantificada por um valor numérico chamado aptidão (*fitness*). Segundo Poli et al. (2008), a métrica de aptidão é o principal, senão o único, mecanismo para declarar em alto nível os requisitos do problema ao sistema de GP. Há diferenças entre as funções de aptidão utilizadas em GP e as utilizadas em outros algoritmos evolutivos, uma vez que as estruturas sendo evoluídas em GP são programas, portanto a avaliação da aptidão normalmente requer que todos os programas da população sejam executados, tipicamente múltiplas vezes. Esse é o caso do presente trabalho. Outra característica comum de métricas de aptidão de GP é que, para muitos problemas práticos, elas são multiobjetivas, ou seja, combinam dois ou mais diferentes elementos que frequentemente competem entre si.

Operações genéticas

Segundo Zongker e Punch (1996), um operador genético é um método para criar indivíduos a cada nova geração, geralmente pela recombinação de partes de indivíduos da população corrente. Cruzamento é o principal operador para recombinar soluções antigas em soluções novas potencialmente melhores tanto em algoritmos genéticos quanto em programação genética. Em GP, o cruzamento ocorre em árvores geralmente da seguinte forma: são selecionados aleatoriamente dois indivíduos e um nó em cada um, ocorrendo, então, o cruzamento por meio da troca das subárvores cujas raízes estão nos nós selecionados em cada indivíduo. O operador genético mais simples é o de reprodução, que consiste em escolher um indivíduo na população atual e copiá-lo literalmente para a nova população. Já a mutação em GP é tipicamente de ponto: são selecionados um indivíduo e um ponto de mutação, e, então, a subárvore com raiz no ponto de mutação é removida e substituída por uma subárvore gerada aleatoriamente. Há muitos outros operadores de mutação e cruzamento que, juntamente com outros detalhes sobre programação genética, são descritos em detalhes por Poli et al. (2008).

Operadores genéticos normalmente são mutuamente exclusivos e sua probabilidade de aplicação é chamada **taxa de operador**. Segundo Poli et al. (2008), em geral, o cruzamento é aplicado com maior probabilidade, com taxa de 90% ou mais, sendo a taxa de mutação bem menor, tipicamente em torno de 1%. Se a soma das taxas de cruzamento e mutação for menor que 100%, é utilizado também o operador de reprodução com taxa complementar. Operadores genéticos são aplicados aos indivíduos probabilisticamente selecionados com base em aptidão. Assim, indivíduos mais aptos apresentam maior probabilidade de serem selecionados para "procriar" e produzir novos indivíduos para a geração seguinte. De acordo com Luke e Spector (1998), a popularidade do operador de cruzamento em programação genética deve-se à crença latente de que o cruzamento em GPs, assim como em GAs, deve de alguma forma transferir "coisas de valor" de um indivíduo a outro, contribuindo para a melhoria das gerações subsequentes. Sendo assim, a literatura de GP usa livremente, com pouco suporte teórico, os conceitos e blocos de construção utilizados em algoritmos genéticos. Os autores observaram em alguns experimentos que, mesmo quando estatisticamente significativa, a diferença entre mutação e cruzamento em GPs é, em muitos casos, surpreendentemente pequena. Muitas vezes, alterar valores de outros parâmetros propiciou maior efeito do que escolher entre cruzamento e mutação. Os autores reportaram uma tendência geral de que mutação é mais bem sucedida em populações menores, e cruzamento, em populações maiores.

Segundo Munroe (2004), diferentemente do que ocorre com algoritmos genéticos, na programação genética a distinção entre cruzamento e mutação não é clara, visto que pontos de cruzamento são aleatórios e pode-se argumentar que, em ambos os casos, um dado código aleatório está sendo substituído por outro código aleatório. No entanto, uma diferença fundamental é que o cruzamento somente pode utilizar material que já existe na população, enquanto a mutação pode gerar código novo usando qualquer função, incluindo aquelas que podem ter sido previamente e, talvez, erronea-

mente, eliminadas pelo processo de seleção. Testando taxas de cruzamento e mutação variáveis com o tempo, o autor observou que há três estágios no sistema de GP: mutação tem melhor desempenho no estágio inicial; os operadores têm essencialmente o mesmo desempenho no segundo estágio; e cruzamento tem melhor desempenho no estágio final. Segundo o autor, esses estágios podem ser vistos como períodos de aprendizagem quando o cruzamento domina, ou períodos de busca cega quando mutação domina ou quando ambos são equivalentes. Para ele, a razão ótima entre taxas de cruzamento e mutação varia ao longo do tempo nos domínios de problema testados, mas a tendência geral é do limite superior de 99% de cruzamento e 1% de mutação.

Parâmetros do sistema

No sistema de GP, o parâmetro de controle mais importante é o tamanho da população. Outros operadores incluem as probabilidades de realizar operações genéticas, o tamanho máximo dos programas e outros detalhes da execução. O critério de parada pode incluir um número máximo de gerações a executar, assim como um critério de sucesso específico do problema em questão. Tipicamente, o melhor indivíduo obtido até o critério de parada ser atingido é, então, retornado como resultado da execução, embora possam ser retornados indivíduos e dados adicionais conforme necessário ou apropriado para o domínio do problema (Poli et al., 2008).

Ainda segundo Poli et al. (2008), é impossível fazer recomendações para configurar valores ótimos de parâmetros, visto que eles dependem muito dos detalhes da aplicação. No entanto, a GP é robusta na prática, de modo que é provável que vários valores diferentes de parâmetros funcionem bem. Consequentemente, em geral não é necessário gastar muito tempo ajustando os parâmetros do sistema de GP para que ele funcione adequadamente. Ainda de acordo com o autor, em muitos casos, incluindo o presente trabalho, o principal limitador do tamanho da população é o tempo necessário para avaliar as aptidões. Como regra, é preferível trabalhar com a maior população que o sistema possa suportar graciosamente. No entanto, há muitos sistemas de GP que consideram populações menores, geralmente empregando mutação mais do que cruzamento como mecanismo primário de busca. Por fim, o número de gerações é tipicamente limitado de dez a cinquenta, uma vez que a busca mais produtiva é geralmente realizada nessas primeiras gerações. Entende-se que se uma solução não for encontrada até então, é pouco provável que seja encontrada em um tempo de execução aceitável.

Implementação

Neste trabalho, o sistema de GP é implementado utilizando a biblioteca *lil-gp* (Zongker e Punch, 1996) da linguagem C. Apenas o operador de adição é utilizado como função e há um terminal representando cada uma das fontes de dados estudadas. A população inicial é criada utilizando o método *ramped half-and-half*, em que metade da população tem todos os nós folhas no mesmo nível e metade tem nós folhas em níveis variados. Em ambos os casos, a profundidade máxima inicial

do indivíduo varia para assegurar que sejam geradas árvores com tamanhos e formatos variados. A seleção de indivíduos para passar por operações genéticas é feita pelo método de seleção por torneio, o mais comumente empregado em programação genética. Nesse tipo de seleção, um número de indivíduos é selecionado aleatoriamente a partir da população e o melhor deles é escolhido para ser "pai". Para o operador de cruzamento são necessários dois "pais", portanto são realizadas duas seleções por torneio.

Capítulo 4
Trabalhos Relacionados

Vários trabalhos constituintes do referencial teórico desta tese foram apresentados nos capítulos anteriores, que abordaram os conceitos biológicos e computacionais essenciais à sua compreensão. Neste quarto capítulo, são destacados trabalhos relacionados ao objetivo da metodologia aqui desenvolvida. Sendo assim, são descritos especificamente trabalhos que envolvem a anotação automática de funções proteicas e a busca por diferentes especificidades em famílias de proteínas.

Anotação Automática de Funções Proteicas

Segundo Erdin et al. (2011), métodos automáticos de anotação de função proteica dependem de uma correlação entre métricas de similaridade funcional e de similaridade estrutural ou sequencial. Os autores resumem bem os tipos de métricas de similaridade entre proteínas encontradas na literatura:

- ◇ As métricas mais simples exploram similaridade de sequências: os alinhadores BLAST ou PSI-BLAST são rotineiramente utilizados para encontrar sequências similares a uma sequência de consulta e, então, a melhor correspondência que apresente função conhecida tem sua anotação transferida para a sequência consultada. Para os autores, uma estratégia melhor seria coletar termos do *Gene Ontology* (GO) entre todas as sequências correspondentes e transferir à sequência de consulta os termos recorrentes com frequência significativa. Neste trabalho, são utilizadas similaridades globais e locais entre pares de sequências, além do número de termos GO que elas têm em comum.

- ◇ Uma segunda métrica de similaridade foca em motivos locais de sequência, que consistem de resíduos que mediam diretamente a função e, portanto, para os autores, devem ser os mais específicos para a anotação funcional. O InterPro, uma das fontes de dados deste trabalho, reúne assinaturas funcionais de proteínas de onze bases de dados.

- ◇ Um terceiro tipo de métrica explora estruturas tridimensionais, podendo-se sobrepor estruturas diretamente umas com as outras ou, mais genericamente, encontrar o modo como elas encaixam-se em classificações estruturais mais amplas. Neste trabalho, são estudadas a similaridade estrutural por meio de sobreposições e da similaridade entre assinaturas estruturais.

- ◇ Uma quarta métrica foca em características estruturais locais como, por exemplo, a geometria local de cavidades ou a eletrostática da superfície, que são informativas para a comparação de sítios ativos e sítios de ligação em nível molecular. A composição do sítio ativo putativo é um dos principais indicadores de similaridade empregados neste trabalho.
- ◇ Finalmente, um quinto tipo de métrica de similaridade baseia-se em padrões tridimensionais, que estreitam ainda mais as buscas por similaridades estruturais locais, visto que tais padrões são compostos de poucos resíduos diretamente associados com a função e posicionados em uma geometria espacial definida. Apesar de não serem utilizados explicitamente neste trabalho, o método utilizado para sobreposição estrutural, as assinaturas estruturais e os domínios do InterPro conseguem capturar essas informações de similaridades estruturais locais.

De acordo com Erdin et al. (2011), uma vez que essas métricas de similaridade focam em diferentes características proteicas, espera-se que elas levariam a resultados melhores quando combinadas. Isto é exatamente o que este trabalho faz, integrando diversas fontes de dados distintas visando à detecção de subfamílias em famílias proteicas. A seguir, é feita uma revisão da literatura relacionada seguindo a ordem cronológica de publicação.

Shah e Hunter (1997) utilizaram todas as enzimas com atribuição de número *Enzyme Commission* (EC) completo na versão 33 do Swiss-Prot e realizaram um teste sistemático da hipótese de que atividade enzimática pode ser prevista diretamente a partir da similaridade de sequências. O sistema EC é uma classificação hierárquica de funções enzimáticas que descreve as reações catalíticas em quatro níveis de detalhe (Erdin et al., 2011). O objetivo dos autores foi estabelecer um limiar de modo que qualquer proteína que alinhasse com outra em uma determinada classe EC com similaridade maior que tal limiar garantidamente estaria na mesma classe. Os autores concluíram que aproximadamente 60% das classes EC contendo duas ou mais proteínas não podiam ser discriminadas utilizando apenas similaridade de sequências com nenhum valor de limiar, e sugeriram que deve-se tentar delimitar domínios funcionais significativos antes de atribuir classes EC. Isso mostra a ineficácia de métodos que se baseiam apenas em similaridade de sequências para transferir anotações funcionais entre proteínas.

Devos e Valencia (2000) apresentaram uma revisão sistemática do relacionamento entre similaridade de sequências e de função, analisando uma coleção de proteínas estruturalmente alinhadas. Os autores analisaram função proteica em quatro níveis: i) classificação de função enzimática; ii) anotação funcional na forma de palavras-chave; iii) classe de função celular; e iv) conservação do tipo de aminoácido no sítio de ligação. Seus resultados mostraram que o tipo de enovelamento podia ser transferido confiavelmente de uma proteína cuja estrutura era conhecida para uma sequência não-caracterizada quando a identidade entre elas era de pelo menos 20% com, no mínimo, alinhamento de cinquenta resíduos e 75% dos tamanhos das sequências. Comparando os valores de identidade sequencial com os quatro níveis que compõe os números EC, os autores observaram que os códigos

EC eram completamente iguais apenas para pares de enzimas com identidade maior que 80%. Entre 50 e 80% de identidade, apenas os três primeiros níveis eram iguais, enquanto abaixo de 50% de identidade o código EC era menos conservado, de modo que foi difícil selecionar as anotações completamente corretas. Os autores encontraram proteínas com 41% de identidade com números EC totalmente diferentes, além de proteínas com 16% de identidade com números EC idênticos. Sítios de ligação foram a característica menos conservada entre proteínas relacionadas, visto que foram observados casos extremos desses sítios que eram muito diferentes em níveis razoavelmente altos de identidade, e casos de proteínas distantemente relacionadas que retiveram sítios de ligação notavelmente conservados. Os autores concluíram que a transferência de função entre sequências similares envolve mais dificuldades do que comumente acredita-se, observando que mesmo relacionamentos verdadeiros entre pares de sequências, identificados por sua similaridade estrutural, correspondem, em muitos casos, a diferentes funções. Isto mostra a importância de métodos que integrem diversos tipos de dados, como feito neste trabalho.

Tian e Skolnick (2003) buscaram classificar famílias enzimáticas com base em similaridades de sequência e de função de modo que sequências em cada família deviam ter o mesmo número EC (completo ou até o terceiro nível). Seus resultados sugeriram que 40% de identidade de sequência poderia ser usado como limiar para transferência de função entre pares de enzimas até o terceiro nível do número EC, mas seria necessária identidade acima de 60% para transferir o número EC completo com 90% de precisão. Posteriormente, o grupo apresentou o EFICAz (*Enzyme Function Inference by Combined Approach*) (Tian et al., 2004; Arakaki et al., 2006, 2009), uma aplicação para inferência de função enzimática que combina diferentes métodos baseados em limiares de similaridade sequencial dependentes de família, na presença de padrões de domínios funcionalmente relevantes e na identificação de resíduos discriminantes.

Laskowski et al. (2005a) desenvolveram o ProFunc, um servidor Web para prever a provável função de proteínas cuja estrutura era conhecida. Vários métodos eram utilizados para analisar sequência e estrutura de uma proteína visando a identificar motivos funcionais ou relacionamentos próximos com proteínas funcionalmente caracterizadas, reportando o que tivesse sido encontrado separadamente por meio de cada método empregado. A ideia da abordagem era utilizar o máximo de métodos possível para não somente aumentar as chances de encontrar um alinhamento útil, mas também para potencializar casos em que vários métodos chegavam a respostas similares ou iguais.

Dobson e Doig (2005) descreveram um método baseado em decomposição em valores singulares (SVD, do inglês *Singular Value Decomposition*) que atribuía função a enzimas utilizando apenas atributos calculados a partir de estruturas cristalográficas como, por exemplo, conteúdo de estrutura secundária, propensões de aminoácidos e propriedades de superfície e de ligantes. As proteínas foram primeiro classificadas em enzimas ou não-enzimas e, segundo os autores, o conteúdo de estrutura secundária, a presença de cofatores e as frações de resíduos, especialmente na superfície, foram úteis

para realizar essa separação. Então, as enzimas foram divididas conforme seu número EC. Para tanto, foram criados quinze classificadores binários, cada um discriminando entre duas classes EC, e, devido aos erros resultantes do desbalanceamento de classes, os autores ajustaram a penalidade de erros conforme o tamanho de cada classe a fim de melhorar o desempenho. No entanto, em um contexto de agrupamento como o do presente trabalho, não se pode supor que as classes sejam conhecidas *a priori*, portanto uma manobra semelhante para contornar o desbalanceamento de classes torna-se inviável. Os autores observaram, por exemplo, que a presença de ferro era um forte indicador de Oxirredutases, que um grande número de atributos eram necessários para os modelos envolvendo Ligases, e que havia alta utilidade na fração de tirosina superficial em modelos envolvendo Hidrolases. Assim, demonstraram a utilidade de atributos estruturais simples na anotação de função enzimática e concluíram que a combinação desses com vários atributos baseados em sequência deveria aumentar o desempenho dos classificadores resultantes, atestando a relevância da integração de vários tipos de dados realizada neste trabalho.

Tetko et al. (2008) compararam anotações de quatro genomas bacterianos geradas automaticamente por métodos de regressão linear múltipla, K vizinhos mais próximos e redes neurais associativas com anotações geradas manualmente por curadores do *Munich Information Center for Protein Sequences* (MIPS) de acordo com o *Functional Catalog* (FunCat). Todos os dados considerados descrevem pares de sequências, com descritores derivados de pontuações de alinhamento de sequência, domínios do InterPro, informação de sintenia, comprimento de sequência e propriedades proteicas calculadas. Os autores utilizaram uma pontuação para estimar a precisão da anotação e treinaram métodos de aprendizado de máquina para prever pontuações de anotação funcional para cada par de sequências. No fim de um processo de validação cruzada de cinco iterações, as pontuações dos conjuntos de teste foram usadas para anotar proteínas, sendo que um par de proteínas protótipo e alvo com pontuação máxima era detectado e, então, a anotação da proteína protótipo era transferida para a proteína alvo. Os autores observaram que a abordagem de redes neurais teve melhor desempenho e concluíram que descritores derivados de domínios do InterPro e similaridade de sequências forneceram a mais alta contribuição para esse desempenho. Além disso, eles destacaram a importância da vizinhança gênica como indicador de conservação de função proteica, o que reforça a validade dos tipos de dados estudados no presente trabalho.

A fim de compreender a relação entre estrutura e função, Bray et al. (2009) analisaram um conjunto de 294 enzimas buscando por diferenças de características estruturais e sequenciais entre as seis principais classes enzimáticas existentes. Tais características incluíam composição de aminoácidos, conteúdo de estrutura secundária, frações de carga, hidrofobicidade, fatores B, ponto isoelétrico médio e área de superfície, tanto para a enzima inteira como para a região do sítio ativo. Os autores observaram, por exemplo, que Oxirredutases apresentavam sítios ativos mais apolares, enquanto Hidrolases formavam a maior proporção de monômeros. As características extraídas foram utilizadas para prever a classe principal das enzimas, com precisão de 26%. A adição de características

específicas de sítio ativo elevou a precisão para 33%. Segundo os autores, os resultados mostraram que há informações relevantes nas características utilizadas, particularmente do sítio ativo, que permitem a classificação nas principais classes enzimáticas. As cinco características mais significativas encontradas por eles foram as proporções totais de leucina e prolina, o número de resíduos na unidade biológica e as proporções, no sítio ativo, de aspartato e de resíduos apolares.

Chitale et al. (2009) desenvolveram o método ESG (do inglês *Extended Similarity Group*), que realiza buscas iterativas em bases de dados de sequências e anota uma sequência de consulta com termos do GO. Cada anotação é atribuída com probabilidade baseada em sua similaridade relativa com vizinhos no grafo de similaridades de proteínas. Segundo os autores, a busca iterativa foi efetiva na captura de múltiplos domínios em uma proteína de consulta, possibilitando a previsão de várias funções originárias de domínios diferentes. O ESG utiliza o PSI-BLAST como ferramenta de busca no UniProt. Ele atribui uma probabilidade a termos do GO com base na soma da significância relativa dos valores esperados de sequências anotadas com tais termos. O PSI-BLAST é executado com uma sequência de consulta, obtendo-se o primeiro nível do ESG. Então, o processo é repetido com cada uma das sequências do primeiro nível, produzindo o segundo nível, e assim por diante. Desse modo, segundo os autores, o método cobre múltiplos níveis de vizinhanças ao redor da proteína de consulta. Os vizinhos do segundo nível são usados para calcular a probabilidade de um termo do GO. O ESG extrai anotações funcionais do espaço de similaridades de sequências, que é ampliado pela busca iterativa em bases de dados.

Kumar e Choudhary (2012) apresentaram um modelo supervisionado de aprendizagem de máquina para prever classe e subclasse funcional de enzimas com base em 73 características derivadas das sequências. Os autores utilizaram o algoritmo de classificação *random forest* para construir um modelo de três camadas em que a camada superior classificava uma proteína em enzima ou não-enzima, a segunda previa a classe funcional principal e a terceira, a subclasse. No entanto, eles utilizaram o mesmo número de enzimas para cada classe e subclasse para evitar o problema de desbalanceamento, o que não pode ser feito no contexto de algoritmos não-supervisionados como os empregados nesta tese. Eles concluíram que atributos derivados de sequências capturam informação rica sobre o mecanismo funcional das enzimas até o nível de subclasse e, utilizando métodos de seleção de atributos, observaram grande relevância biológica de dados como composição de cisteína, peso molecular e número de resíduos. No entanto, concluíram também que atributos sequenciais são insuficientes para prever mecanismos enzimáticos, sugerindo sua combinação com características estruturais. Novamente, a complexidade do problema de anotação funcional mostra a grande importância de trabalhos que integram vários tipos de informação biológica, como é o caso desta tese.

Para um conjunto de enzimas, Boareto et al. (2012) analisaram a relação entre as funções definidas pelos números EC e parâmetros físico-químicos como conteúdo de estrutura secundária, frequência de aminoácidos, área e carga de superfície, volume e hidrofobicidade. Eles definiram uma

distância entre enzimas utilizando um método de aprendizado supervisionado e estudaram a geometria de diferentes grupos enzimáticos gerados por um algoritmo de agrupamento hierárquico. Não foram encontrados agrupamentos compatíveis com a classificação EC. Segundo eles, os resultados sugeriram que parâmetros estruturais globais são insuficientes para separar enzimas de acordo com a hierarquia EC, o que, para eles, indica que características essenciais para a função são locais.

Radivojac et al. (2013) reportaram os resultados do primeiro experimento em larga escala de avaliação crítica de métodos de anotação proteica baseada em comunidade (CAFA, do inglês *Critical Assessment of protein Function Annotation*), em que 54 métodos foram avaliados com um conjunto alvo de 866 proteínas provenientes de onze espécies, focando em esquemas de classificação proteica fornecidos pelo consórcio GO. O tipo de dados utilizado pela maioria dos métodos foi a sequência de aminoácidos, pressupondo que a similaridade de sequências está correlacionada com a similaridade funcional. Mais da metade dos métodos usaram dados além da similaridade de sequência, como tipos de relacionamentos evolutivos, estrutura proteica, interações proteína-proteína ou dados de expressão gênica. O desafio para tais métodos foi encontrar formas de integrar fontes de dados discrepantes e tratar adequadamente dados incompletos e ruidosos.

Ainda segundo Radivojac et al. (2013), do lado computacional, a maioria dos métodos utilizou princípios de aprendizagem de máquina. Tipicamente, eles buscaram combinações de características baseadas ou não em sequência que correlacionassem com uma função específica em um conjunto de treino de proteínas experimentalmente anotadas. Segundo os autores, apesar da camada adicional de complexidade, o aprendizado de máquina desempenhou em geral um papel positivo no aumento da precisão de anotação e, portanto, pode-se esperar que métodos de alto desempenho no futuro serão baseados em princípios bem fundados de aprendizagem e inferência estatística. Os autores concluíram que os métodos atuais de previsão funcional apresentam desempenho significativamente melhor que métodos de primeira geração amplamente utilizados, com grandes ganhos em todos os tipos de alvos. No entanto, embora os melhores métodos tenham desempenho bom o suficiente para guiar experimentos, há uma necessidade considerável de melhoria nas ferramentas atualmente disponíveis. Para os autores, uma vez que a similaridade de sequências é menos preditiva dos papéis biológicos de proteínas, uma chave para melhorar a precisão em anotar a função biológica de uma proteína será a capacidade de gerar-se dados sistêmicos de melhor qualidade e de desenvolver ferramentas computacionais que os explorem. Além disso, com base no que observaram, afirmaram que os métodos mais poderosos serão aqueles que encontrarem formas de integrar uma variedade de evidências experimentais e ponderar cada tipo de dados apropriada e separadamente para cada termo funcional.

Os resultados desses trabalhos mostram que apenas um tipo de informação é insuficiente para anotar funções proteicas com precisão, devido à imensa quantidade de fatores envolvidos e à consequente complexidade do problema de anotação automática. Eles indicam ainda que métodos para prever sítios ativos ou funções proteicas com base em estrutura e outros tipos de informação são mais

apropriados do que aqueles baseados somente em dados de sequência. Segundo Lee et al. (2007), frequentemente observa-se que o poder de anotação funcional de uma abordagem combinada é maior do que o poder dos componentes utilizados individualmente. Sendo assim, fica destacada a grande importância e demanda por métodos de análise automática de função proteica capazes de integrar vários tipos de dados conforme proposto neste trabalho.

Repositórios de famílias proteicas buscam identificar subgrupos com funções mais específicas e, segundo Lee et al. (2007), o sucesso de um repositório geralmente pode ser aumentado por meio da identificação de resíduos específicos que discriminam entre funções. Esses são justamente os propósitos do presente trabalho: agrupar famílias proteicas visando à detecção de subfamílias possivelmente isofuncionais, ao mesmo tempo em que são identificados os resíduos específicos de determinados grupos que os diferenciam dos demais. Uma vez que as muitas dificuldades enfrentadas para a realização de anotação automática de função podem ser estendidas ao problema aqui estudado, é adotada uma abordagem que integra diversos tipos de dados visando a contornar os principais problemas reportados na literatura relacionada a anotação de função. A seguir, apresentamos trabalhos relacionados à determinação de resíduos que discriminam entre grupos de proteínas.

Detecção de Resíduos Discriminantes de Grupos

Livingstone e Barton (1993) usaram alinhamentos múltiplos de sequência e propriedades físico-químicas dos resíduos para analisar conservações de resíduos por grupo. Casari et al. (1995) propuseram uma representação de sequência na forma de vetores e utilizaram técnicas de redução de dimensionalidade para projetar tais vetores em menos dimensões e detectar subgrupos e resíduos funcionalmente importantes. Hannenhalli e Russell (2000) exploraram diferentes modelos ocultos de Markov para alinhamentos múltiplos de sequência visando a detectar posições responsáveis por criar subdivisões em famílias de proteínas. Mesa et al. (2003) apresentaram um conjunto de métodos baseados em árvores filogenéticas visando a descobrir resíduos importantes para ramificações nas árvores. Posteriormente, o mesmo grupo de pesquisa publicou um método independente de filogenia (Pazos et al., 2006): o Xdet, um método supervisionado para detecção de sítios funcionais em posições conservadas em alinhamentos múltiplos de sequência. Já Yu et al. (2005) propuseram um método que, dado um alinhamento múltiplo de sequência, uma classificação de subfamílias baseada no Swiss-Prot e uma estrutura representativa, determinava resíduos de superfície que discerniam grupos. Capra e Singh (2008) descreveram uma metodologia para identificar posições determinantes de especificidade que consistiu na construção de alinhamentos múltiplos de sequência para famílias proteicas, seleção dos resíduos estruturalmente próximos a ligantes e filtragem dos resíduos conservados para detectar diferentes subfamílias.

Uma grande desvantagem dos métodos anteriores para determinar padrões de conservação de resíduos que diferenciem subgrupos de uma família proteica é a necessidade de conhecer as subfa-

mílias *a priori*. Além da escassez de informação experimental sobre subfamílias, essa necessidade é um fator extremamente limitante quando trabalha-se com famílias de função desconhecida, que é o propósito desta tese.

Bleicher et al. (2011) apresentaram um método baseado em mutações correlacionadas e análise de redes para calcular e analisar grupos de aminoácidos que poderiam estar relacionados a subclasses funcionais em famílias proteicas. Esses autores propuseram uma métrica de correlação específica para cada tipo de aminoácido que poderia ser usada para construir redes que sumarizavam padrões de correlação e anti-correlação em uma família de proteínas. Nesse caso, a conexão entre dois nós da rede implicava que sequências que apresentavam o primeiro aminoácido em determinada posição tendiam a ter, também, o segundo em outra posição específica. Eles analisaram tais redes com algoritmos de detecção de comunidades, resultando em subconjuntos de aminoácidos correlacionados que tendiam a estar presentes simultaneamente. Segundo os autores, seus resultados mostraram como os parâmetros e procedimentos propostos estavam relacionados a características biológicas observadas nas famílias proteicas estudadas, o que realçava o uso potencial da metodologia na caracterização de proteínas e anotação gênica. Além disso, o método poderia ser explorado para identificar resíduos-chave para propriedades funcionais específicas, além de ser utilizado para anotação gênica.

Active Sites Modelling and Clustering (ASMC)

De modo semelhante ao empregado no presente trabalho, o ASMC (Melo-Minardi et al., 2010) é um método que busca primeiro determinar subfamílias de uma família de enzimas utilizando informação estrutural sobre cavidades e, então, detectar resíduos determinantes de função e especificidade que caracterizem cada um dos subgrupos obtidos. Segundo Melo-Minardi et al. (2010), o ASMC combina modelagem estrutural por homologia dos membros da família com uma estrutura de referência, alinhamento estrutural dos sítios ativos modelados e o que os autores chamaram de "classificação conceitual hierárquica" das enzimas, embora a técnica empregada seja de agrupamento, não de classificação. Resumidamente, o ASMC emprega um algoritmo de agrupamento hierárquico em um conjunto de dados sobre composição de resíduos de cavidades estruturais buscando determinar padrões de composição de resíduos responsáveis por diferenciações funcionais em uma família de enzimas. A comparação dos perfis dos grupos gerados permite identificar resíduos correlacionados com a divergência de subfamílias.

De acordo com Melo-Minardi et al. (2010), dada uma família do Pfam, são excluídas da base de dados as sequências com tamanhos que diferem em mais de um desvio-padrão da média de tamanho da família, além daquelas com menos de 30% de identidade com as estruturas de referência ligadas à família, obtidas no PDB. Então, o programa Modeller (Eswar et al., 2006) é utilizado para a modelagem comparativa das sequências da família com as estruturas de referência, cujas cavidades são calculadas utilizando o Fpocket (Guilloux et al., 2009), buscando a cavidade onde é localizado

o sítio ativo. O MultiProt (Shatsky et al., 2004) é, então, usado para alinhar os modelos estruturais às estruturas de referência, recuperando os resíduos dos modelos espacialmente alinhados com a cavidade do sítio ativo putativo das estruturas de referência e construindo um alinhamento múltiplo de sequências (MSA, do inglês *Multiple Sequence Alignment*) que representa a composição do sítio ativo putativo em cada enzima da família. Esse MSA é fornecido como entrada para o algoritmo de agrupamento hierárquico Cobweb (Fisher, 1987) da biblioteca Weka, que então gera uma árvore cujos nós são grupos de proteínas e cujos níveis representam subdivisões sucessivas do conjunto de enzimas. No ASMC, Melo-Minardi et al. (2010) alteram manualmente a árvore gerada, juntando grupos com composições semelhantes a fim de obter grupos de maior interesse. Posteriormente, é feita uma análise de significância estatística de cada posição do MSA para determinar as posições mais importantes para diferenciar entre os grupos. Os autores argumentaram que a especificidade de cada grupo poderia ser explicada por uma ou mais posições determinantes de especificidade que não eram conservadas em outros grupos na mesma posição no padrão de sítio ativo, ou seja, aminoácidos chave responsáveis pela segregação entre grupos poderiam ser detectados pela comparação dos padrões de conservação dos grupos. De acordo com eles, o ASMC é aplicável a famílias de enzimas com funções bem-caracterizadas e com pelo menos uma estrutura conhecida que contém ligantes para servir de molde para a modelagem das sequências da família.

Bastard et al. (2014) descreveram uma estratégia integrada para descoberta de atividades enzimáticas catalisadas por proteínas de função desconhecida ou pouco conhecida que combina métodos computacionais e procedimentos experimentais e cujo principal método utilizado é o ASMC. Eles investigaram a família de proteínas DUF849 do Pfam, analisando 725 sequências, a maioria das quais coberta quase completamente pelo domínio que dá nome à família. Primeiramente, os autores estabeleceram uma reação química genérica que poderia corresponder à família estudada, juntamente com uma lista de potenciais substratos, e selecionaram um conjunto de proteínas representantes para triagem enzimática de alto desempenho. Então, realizaram análises estruturais e computacionais dos sítios ativos das sequências modeladas por homologia com uma estrutura de referência (um passo do ASMC), além de docagem de substratos a fim de interpretar a diversidade de atividades encontradas na família e de identificar os elementos estruturais responsáveis pela especificidade e promiscuidade. Finalmente, foram explorados os contextos genômico e metabólico, além da caracterização bioquímica, para revelar o papel *in vivo* de algumas das atividades descobertas. Foi utilizada uma abordagem de *ensemble clustering* para integrar resultados das diferentes estratégias de agrupamento utilizadas, baseadas em similaridade de sequências proteica, análise filogenética, contexto genômico e composição de sítios ativos (ASMC), uma vez que cada método levou a resultados diferentes. Manualmente, foram atribuídos diferentes pesos a cada um dos agrupamentos primários para dar-lhes maior ou menor influência no agrupamento consenso final. O agrupamento final escolhido pelos autores contém 32 subfamílias consenso com tamanhos variando de 3 a 130 proteínas.

Descrevendo de forma simples o uso do ASMC por Bastard et al. (2014), o método alinhou os modelos estruturais das 725 sequências com a estrutura de referência da família Kce e projetou linearmente os resíduos que alinharam com o sítio ativo da referência para gerar um alinhamento de sequências baseado em estrutura. Então, tais sequências de sítio ativo foram agrupadas com o algoritmo de agrupamento hierárquico Cobweb e a árvore resultante foi chamada de "árvore hierárquica de sítio ativo". Essa abordagem gerou 84 grupos-folha que foram utilizados como agrupamento primário da abordagem de *ensemble clustering*. Dos 84 grupos, 31 continham mais de uma proteína, cobrindo 92,7% da família. No entanto, os autores também editaram manualmente a árvore resultante até obterem sete grupos principais, cada qual contendo entre 50 e 156 sequências. Eles reportaram uma alta correlação entre a natureza dos compostos transformados e sua distribuição entre os sete grupos manualmente definidos com base na árvore gerada pelo ASMC. As sequências de sítio ativo pertencentes a um mesmo grupo foram utilizadas para formar padrões de conservação do grupo. Segundo os autores, com algumas exceções, esses grupos manualmente definidos concordaram com a árvore filogenética e sugeriram relacionamentos estruturais e evolutivos que levaram à diversificação da família. Ainda segundo eles, para que essa abordagem possa ser aplicada, é necessário que pelo menos uma atividade enzimática na família seja conhecida.

A fim de possibilitar a comparação da metodologia desenvolvida neste trabalho com o ASMC, adotamos parte de seu passo-a-passo e utilizamos as mesmas famílias proteicas estudadas por Melo-Minardi et al. (2010) e Bastard et al. (2014), como será descrito no Capítulo 5.

Capítulo 5
Metodologia

Este capítulo detalha a metodologia desenvolvida neste trabalho, além de descrever as famílias de proteínas às quais foi aplicada nos estudos de caso. Os principais passos são descritos a seguir.

1. *Definição da família proteica Pfam a ser estudada.* Coleta das sequências das proteínas das famílias no Pfam e posterior filtragem da base, além da recuperação das estruturas de referência associadas à família no PDB e geração dos modelos estruturais.
2. *Coleta de evidências de similaridade.* Obtenção de dados junto às fontes descritas na Seção 2.10, que abrangem similaridades de sequência, estrutura, contexto genômico, composição de aminoácidos, domínios e parâmetros físicos e químicos. Cálculo da similaridade entre pares de proteínas segundo cada fonte de dados estudada.
3. *Integração de dados.* Combinação, via programação genética, dos vários tipos de dados para compor as matrizes de similaridades fornecidas como entrada ao algoritmo de agrupamento.
4. *Agrupamento.* Definição, implementação e execução dos algoritmos de agrupamento.
5. *Avaliação de experimentos.* Análise dos resultados da aplicação dos algoritmos de agrupamento à família de proteínas em estudo e comparação com os resultados obtidos por outro método da literatura: o ASMC (Melo-Minardi et al., 2010). Verificação dos tipos de dados que apresentaram maior utilidade para o problema de detecção de subfamílias proteicas.

5.1 Famílias de Proteínas Estudadas

Como discutido na Seção 3.1, algoritmos de agrupamento independem de informação de supervisão como rótulos de classe dos objetos sendo agrupados. No entanto, a fim de avaliar desempenho e facilitar a comparação com resultados obtidos pelo ASMC (Melo-Minardi et al., 2010) para famílias proteicas bem conhecidas, a técnica proposta neste trabalho foi aplicada às mesmas famílias estudadas pelos autores: Nucleotidil Ciclases (família Pfam PF00211), Serino Proteases (PF00089) e Proteínas Cinases (PF00069 e PF07714). Essas famílias também foram consideradas no trabalho de Hannenhalli e Russell (2000), ao qual os autores compararam o ASMC. Como dito, são estudadas as mesmas proteínas utilizadas por Melo-Minardi et al. (2010), apenas eliminando da base aquelas que, desde então, foram removidas do UniProt. A filtragem das sequências obsoletas resultou em conjuntos de 461 Nucleotidil Ciclases, 1.533 Serino Proteases e 3.087 Proteínas Cinases, após a eliminação

de 75, 140 e 314 proteínas, respectivamente. Dado o propósito de detectar subfamílias possivelmente isofuncionais em famílias de função desconhecida, consideramos também as 725 proteínas da família Pfam DUF849 estudada por Bastard et al. (2014).

Estudo de Caso I: Nucleotidil Ciclases

Segundo Hannenhalli e Russell (2000), as Nucleotidil Ciclases são uma família de domínios citosólicos ou anexados a membrana que catalisam a reação que transforma um nucleotídeo trifosfato em um nucleotídeo monofosfato cíclico. Segundo Tucker et al. (1998), esta família possui duas subfamílias funcionais: Adenilato Ciclases, que utilizam ATP como substrato para formar cAMP, e Guanilato Ciclases, que catalisam a conversão análoga de GTP em cGMP. A mutação de apenas dois resíduos (glutamato para lisina e cisteína para aspartato) são suficientes para alterar completamente a especificidade do nucleotídeo de GTP para ATP. Guanilato e Adenilato Ciclases têm papéis fundamentais em uma ampla gama de processos celulares.

Estudo de Caso II: Serino Proteases

Proteases são uma grande família de enzimas envolvidas na hidrólise de ligações entre aminoácidos em uma proteína. Quase um terço de todas as Proteases podem ser classificadas como Serino Proteases, cujo nome deriva do resíduo nucleófilo de serina no sítio ativo (Hedstrom, 2002). Serino Proteases estão envolvidas em um número enorme de processos biológicos como, por exemplo, digestão, homeostase, apoptose, transdução de sinais, reprodução, resposta imunológica e coagulação sanguínea (Hedstrom, 2002; Neitzel, 2010). Esses nichos fisiológicos diversificados demandam especificidades altamente variáveis, indo desde Proteases digestivas que clivam depois de resíduos hidrofóbicos ou positivamente carregados, até as que reconhecem um sítio de clivagem de cinco resíduos ou mesmo uma única proteína (Hedstrom, 2002).

Serino Proteases possuem uma tríade catalítica composta por uma serina, um aspartato e uma histidina. Segundo Neitzel (2010), a disposição tridimensional da tríade catalítica permite o movimento de prótons para dentro e para fora do sítio ativo da enzima. A histidina está posicionada para atuar como base, um receptor de prótons, e remover o próton do grupo OH da serina que, com isso, torna-se muito mais reativa e pode facilmente formar uma nova ligação com o átomo de carbono da ligação peptídica do substrato. A carga negativa do grupo carboxila do aspartato está na posição correta para estabilizar a cadeia lateral positivamente carregada da histidina.

Todas as Serino Proteases atuam por meio de um mecanismo catalítico similar, mas apresentam diferentes preferências quanto à ligação que clivam, o que deve-se a alterações no sítio ativo (Hannenhalli e Russell, 2000; Melo-Minardi et al., 2010). De acordo com Neitzel (2010), a cavidade de ligação das Quimotripsinas é revestida por aminoácidos hidrofóbicos, por isso proteínas com resíduos hidrofóbicos como leucina ou isoleucina ligam-se fortemente e na orientação correta para que a

tríade catalítica atue. A cavidade das Tripsinas tem um aspartato negativamente carregado, então seus substratos devem ter um aminoácido positivamente carregado como a lisina ou arginina na posição adequada. Já a cavidade da Elastase é muito pequena, por isso apenas proteínas com aminoácidos de cadeias laterais curtas como glicina ou alanina podem ser quebradas. Segundo Hannenhalli e Russell (2000), o aspartato encontrado no sítio ativo de Tripsinas, que justifica a preferência por resíduos positivos, geralmente é substituído por um resíduo pequeno em Quimotripsinas (serina, o que é responsável pela sua propriedade hidrofóbica) e Elastases (glicina). Além disso, duas glicinas do sítio ativo de Tripsinas e Quimotripsinas são geralmente substituídas por valina e treonina nas Elastases, causando uma oclusão da cavidade que justifica sua preferência por resíduos alifáticos pequenos.

Serino Proteases são capazes de atuar sobre uma ampla gama de proteínas, mas em algumas vias metabólicas, como coagulação sanguínea ou no sistema imunológico, uma Serino Protease pode ser tão específica que consiga clivar apenas uma determinada ligação em uma única proteína substrato. Uma revisão aprofundada sobre o mecanismo e a especificidade das Serino Proteases pode ser encontrada no trabalho de Hedstrom (2002).

Estudo de Caso III: Proteínas Cinases

Proteínas Cinases são enzimas que modificam a função de outras proteínas adicionando a elas grupos fosfato geralmente retirados de ATPs, ligando-os covalentemente à cadeia lateral dos aminoácidos serina, treonina ou tirosina. São uma das maiores e mais funcionalmente diversas famílias proteicas, controladoras da maioria das vias bioquímicos, desempenhando papéis-chave na regulação de processos metabólicos, diferenciação celular e na proliferação de diversos tipos celulares (Smith et al., 1997). Por meio da adição de grupos fosfato, elas direcionam a atividade, localização e função geral de muitas proteínas, e servem para orquestrar a atividade de quase todos os processos celulares.

A reação inversa à fosforilação é desempenhada pelas enzimas Fosfatases, por isso funções celulares muitas vezes são reguladas em parte pelo equilíbrio entre a atividade de Proteínas Cinases e Fosfatases (Petretti e Prigent, 2005). Dado seu papel vital na função celular normal e em doenças, Proteínas Cinases formam o segundo grupo de proteínas mais importante e são consideradas alvos prioritários pela indústria farmacêutica, que busca projetar inibidores a serem utilizados para terapia humana. Segundo Hannenhalli e Russell (2000), uma divisão principal da família é entre Serina/Treonina e Tirosina Cinases: serina e treonina são similares em tamanho e forma, enquanto a química da reação e o tamanho do substrato são substancialmente distintos para a tirosina. A maioria das Cinases atua sobre serina ou treonina, enquanto algumas são específicas de tirosina, e outras atuam sobre todas as três. Sabe-se que algumas posições conferem especificidade: no subdomínio VI, por exemplo, a sequência-consenso RDLKPEN é geralmente encontrada em Serina/Treonina Cinases, enquanto a sequência RDLAARN é típica de Tirosina Cinases.

Estudo de Caso IV: DUF849

Essa família Pfam, definida pela presença de um domínio conservado de função desconhecida, foi estudada por Bastard et al. (2014) por conter a proteína Kce, que foi de interesse para os autores por terem anteriormente descoberto uma associação inicial entre ela e uma atividade enzimática que até então era órfã, ou seja, não tinha sequências proteicas associadas. Tal atividade envolve a via de fermentação de lisina e catalisa a condensação de β-ceto-5-amino-hexanoato (KAH) e acetilcoenzima A para produzir aminobutiril-CoA e acetoacetato. Segundo os autores, como a proteína Kce pertence à DUF849, cujas proteínas não são todas de organismos capazes de fermentar lisina, isto sugere a existência de um conjunto de diversas reações bioquímicas catalisadas por diferentes membros da família. Assim, os autores consideraram a DUF849 um bom estudo de caso para a descoberta de novas atividades em uma família proteica de função desconhecida. Eles deram à família o nome "BKACE", do inglês *β-keto acid cleavage enzyme*.

Processamento

Uma vez definida uma família de proteínas de interesse no Pfam, os passos para obtenção do conjunto de proteínas submetidas ao algoritmo de agrupamento são descritos a seguir.

1. Recuperação do alinhamento de sequências completo da família no Pfam.
2. Extração dos identificadores UniProt das sequências, assim como das posições do HMM da família em cada sequência. Como discutido na Seção 2.10, cada família do Pfam é representada por um perfil HMM e, uma vez que proteínas podem apresentar múltiplos domínios, utilizamos as posições em que o HMM encontra-se em cada sequência para considerar apenas a subsequência relacionada à família em estudo.
3. Coleta do conjunto de sequências no UniProt e extração das subsequências relacionadas ao domínio que define a família Pfam.
4. Filtragem do conjunto de proteínas da família pelo tamanho das subsequências, eliminando aquelas cujo tamanho divirja mais do que um desvio-padrão da média da família. Esse passo foi adotado a fim de possibilitar a comparação com o ASMC.
5. Obtenção, no PDB, das estruturas associadas à família e separação dessas em cadeias.
6. Seleção de estruturas de referência a serem usadas para a modelagem comparativa das sequências da família e para busca de cavidades que sejam possíveis sítios ativos. São priorizadas estruturas obtidas por cristalografia de raio-x, com resoluções altas e que contenham ligantes.

7. Filtragem do conjunto de proteínas pela similaridade com as sequências das estruturas de referência, eliminando aquelas que não tenham ao menos 30% de identidade com nenhuma das estruturas. Tal limiar de identidade é um requisito do algoritmo de modelagem comparativa.

8. Modelagem das sequências resultantes com as estruturas de referência utilizando o programa Modeller (Eswar et al., 2006) e escolha do melhor modelo para cada sequência como sendo o de menor energia, como feito por Melo-Minardi et al. (2010) e Bastard et al. (2014).

Ao final desse processo, a base de dados resultante contém, para cada proteína, sua identificação no UniProt, sua sequência de aminoácidos e seu modelo estrutural.

5.2 Fontes de Dados de Similaridade

Um dos objetivos deste trabalho é analisar o modo como informação proveniente de diferentes domínios de conhecimento é capaz de auxiliar um processo não-supervisionado de agrupamento a detectar subfamílias de proteínas. Esta seção descreve os passos para obtenção dos vários tipos de dados utilizados como fontes de evidência de similaridade funcional, assim como o modo como tais bases de dados suplementares foram empregadas. Cada uma gera uma ou mais matrizes de similaridades baseada no tipo de dados em questão, e as diversas matrizes são integradas pelo sistema de programação genética (GP) para fornecerem a entrada do algoritmo de agrupamento.

Alinhamentos de sequência e estrutura

Como anteriormente discutido, em geral, alinhamentos são utilizados para inferir analogias estruturais ou funcionais entre proteínas. Neste trabalho, alinhamentos par-a-par de sequências e de estruturas foram realizados para extração de similaridades entre pares de proteínas, seguindo a hipótese de que altos valores de similaridade de sequência e, principalmente, de estrutura indicam maior evidência de similaridade funcional.

Para realização dos alinhamentos de sequência globais e locais são utilizados, respectivamente, os algoritmos de Needleman-Wunsch (Needleman e Wunsch, 1970) e Smith-Waterman (Smith e Waterman, 1981) implementados no pacote Biostrings (Pages et al., 2012) do R, um ambiente de desenvolvimento para cálculos estatísticos e gráficos. Ambos os tipos de alinhamentos foram realizados utilizando a matriz de substituição de aminoácidos BLOSUM62, e as pontuações dos alinhamentos par-a-par definem as matrizes de similaridades globais e locais de sequência. Já os alinhamentos estruturais par-a-par foram obtidos utilizando o TM-Align (Zhang e Skolnick, 2005). Três matrizes de similaridades estruturais são geradas, contendo os tamanhos dos alinhamentos, a porcentagem de identidade e o *TM-score* médio do alinhamento da proteína A com a B, e da B com a A, visto que podem haver variações conforme a ordem de entrada das estruturas no algoritmo.

Assinaturas estruturais do *Cutoff Scanning* (da Silveira et al., 2009)

O algoritmo *Cutoff Scanning* é utilizado neste trabalho para extrair o que seus autores chamam de assinatura estrutural. A estrutura tridimensional de uma proteína é representada como um histograma do número de vizinhos que um resíduo de aminoácido apresenta dentro de distâncias variáveis. Segundo da Silveira et al. (2009), famílias diferentes possuem padrões de enovelamento distintos e, consequentemente, histogramas ou assinaturas estruturais característicos.

O método CSM (do inglês *Cutoff Scanning Matrix*) (Pires et al., 2011) consiste no cálculo da distância Euclidiana em Ångströms (1Å = 10^{-10}m) entre os carbonos-alfa de todos os pares de resíduos de uma estrutura proteica. Então, é feita uma contagem dos pares cuja distância é menor que um dado limiar (*cutoff*). A variação desse limiar leva à geração de um vetor em que cada posição denota o número de pares de resíduos cujos carbonos-alfa estão dentro de uma dada distância entre si. Segundo os autores, o vetor de cada proteína é sua assinatura estrutural, representante de seu enovelamento, carregando assim informação importante para determinação de função.

Neste trabalho, foi utilizado um passo de 0,2Å, variando de 0 a 30Å como feito por Pires et al. (2011): a primeira posição do vetor corresponde ao número de pares de resíduos cuja distância máxima é de 30Å, a segunda, ao número de pares cuja distância é de até 29,8Å, e assim por diante. O exemplo a seguir mostra parte do fim da assinatura da proteína com identificador UniProt Q5SLL6, considerando a cadeia A da estrutura PDB 1ZZG. O vetor indica, por exemplo, que há um único par de resíduos cujos carbonos-alfa estão a até 3Å um do outro, 121 pares cuja distância é de até 3,8Å e 414 pares cuja distância é no máximo 4Å.

$$[\ldots\ 439\ 427\ 417\ \mathbf{414}\ \mathbf{121}\ 1\ 1\ 1\ \mathbf{1}\ 0\ 0\ \ldots]$$

As distâncias Euclidianas entre os vetores produzidos pelo CSM para cada par de proteínas são consideradas como evidência de similaridade funcional. Quanto menor a distância, maior a evidência de similaridade segundo essa fonte de dados, por isso a matriz de similaridades de assinaturas estruturais corresponde à matriz de distâncias transformada de modo que valores mínimos de distância correspondam a valores máximos de similaridade.

Contextos Genômicos

Os dados relativos ao contexto genômico foram extraídos da versão 9.1 do banco de dados STRING (Franceschini et al., 2013). Para cada par de proteínas, é armazenado sua pontuação no STRING para vizinhança genômica conservada (*neighborhood*), eventos de fusão gênica (*fusion*), coocorrência em diferentes genomas (*cooccurrence*) e coexpressão em diferentes espécies (*coexpression*), gerando assim quatro matrizes de similaridades baseadas em contexto genômico.

Propriedades das Proteínas

Segundo Dobson e Doig (2005), a composição de aminoácidos por si só contém uma quantidade surpreendente de informação relevante para a função de proteínas. Além disso, para anotar a função de uma proteína a partir de sua estrutura sem utilizar alinhamentos é preciso usar atributos estruturais que capturem informação pertinente para a diferenciação funcional. Sendo assim, diversas características proteicas que possivelmente indicam função e, consequentemente, podem auxiliar no agrupamento de proteínas em subfamílias são coletadas a partir dos softwares e banco de dados descritos a seguir. Como as propriedades são específicas de cada proteína, cada tipo de dados gera uma matriz de similaridades por meio da comparação dos valores entre pares de proteínas.

EMBOSS Pepstats (Rice et al., 2000): são coletados dados de pesos moleculares, pontos isoelétricos e porcentagens molares das classes de aminoácidos (alifáticos, aromáticos, apolares, polares, carregados, básicos e ácidos), sendo utilizadas as diferenças de valores para cada par de proteínas. Já os valores de estatísticas Dayhoff (DayhoffStat) são utilizados para calcular vetores de composição de aminoácidos, em que cada posição reflete a quantidade de um tipo de resíduo presente na proteína. Então, são calculadas as distâncias Euclideanas quadradas entre os vetores de cada par de proteínas. Considera-se que quanto menores as diferenças e distâncias, maior a evidência de similaridade entre o par de proteínas, por isso as matrizes de similaridades baseadas nesses dados são calculadas transformando as matrizes de distâncias correspondentes.

ExPASy ProtParam (Gasteiger et al., 2005): são coletados índices de instabilidade e GRAVY, utilizando as diferenças de valores para cada par de proteínas. Novamente, considera-se que quanto menores as diferenças entre esses valores, maior o indício de similaridade entre o par de proteínas, por isso as matrizes de similaridades baseadas em instabilidade e GRAVY são calculadas transformando as matrizes de distâncias.

InterPro (Mitchell et al., 2015): são coletados todos os domínios e motivos associados a cada proteína. Neste trabalho, considera-se que quanto mais domínios duas proteínas têm em comum, maior a evidência de similaridade funcional. Assim, a matriz de similaridades baseada no InterPro é calculada pelo número de anotações em comum entre cada par de proteínas.

Gene Ontology (The Gene Ontology Consortium, 2015): são coletados todos os termos GO associados a cada proteína. Assim como para os dados do InterPro, a matriz de similaridades baseada no GO é calculada pelo número de termos em comum entre cada par de proteínas, pois considera-se que quanto mais termos duas proteínas apresentam em comum, maior o indício de similaridade funcional entre elas.

Sítios ativos putativos

O sítio ativo é o elemento central para o algoritmo ASMC (Melo-Minardi et al., 2010), visto que são os atributos que descrevem as proteínas para o algoritmo de agrupamento Cobweb. Por isto, a título de comparação com essa técnica, o sítio ativo putativo foi também incluído neste trabalho utilizando o mesmo processo de obtenção, descrito a seguir. Dada uma estrutura de referência da família proteica em estudo, o Fpocket (Guilloux et al., 2009) é utilizado para detectar suas cavidades. Então, para cada proteína da família, seu modelo estrutural é sobreposto à estrutura de referência utilizando o MultiProt (Shatsky et al., 2004) visando a extrair, para cada proteína, os resíduos que alinharam com aqueles pertencentes às cavidades da estrutura de referência. Quando não há correspondência no modelo com um resíduo da estrutura de referência, aquela posição é marcada com uma lacuna (-). Assim, são obtidas as composições das cavidades em cada proteína da família. Para cada cavidade, é gerado um alinhamento múltiplo de sequências em que cada posição corresponde a um resíduo na cavidade da estrutura de referência, como exemplificado na Figura 5.1.

```
A1YQY9  ----FTCMLV-TIGDCMRVRFCLFGDVNSRES
A1ZB47  FDIVFTAVLVETI-DSMRVRYCLFGDVNSRES
A2FPM7  FDIVFTCMLMKCIGDCMRLTFEIFGPIVQQEH
A2RVE6  FDIVFTCVLVET--DAMRVRYCLFGDVNSRES
A2SW27  FGI-FTCVLVETV-DKMTIRYCLFGNVNSRET
```

Figura 5.1: Exemplo de alinhamento múltiplo dos resíduos das cavidades proteicas.

Uma vez determinadas as cavidades da estrutura de referência, é necessário escolher entre elas qual utilizar para o agrupamento. Melo-Minardi et al. (2010) executaram o ASMC utilizando todas as cavidades encontradas pelo Fpocket e apresentaram resultados para a cavidade mais conservada na família, que depois observaram ser o sítio ativo das enzimas estudas. Neste trabalho, a cavidade mais provável de ser o sítio ativo é selecionada analisando a conservação de seus resíduos na família, visto que resíduos importantes para a função tendem a ser conservados, como discutido em detalhes no Capítulo 2. O algoritmo utilizado para selecionar a cavidade mais promissora neste trabalho é simples: se existem cavidades com três ou mais resíduos conservados em pelo menos 50% da família, escolhe-se aquela com maior pontuação calculada pelo Fpocket. A quantidade de três resíduos conservados foi definida com base na observação de Bartlett et al. (2002) de que cada enzima tem uma média de 3,5 resíduos catalíticos.

O alinhamento múltiplo de sequências correspondente à cavidade escolhida como possível sítio ativo é então utilizado para comparar a composição da mesma para cada par de proteínas da família estudada. Quanto mais similares as composições do sítio ativo putativo, maior é o indício de similaridade funcional entre o par. Duas matrizes de similaridades com base em composição do sítio ativo putativo são calculadas, uma utilizando a identidade dos resíduos e outra, a pontuação do alinhamento de acordo com a matriz de substituição BLOSUM62. Além de ser fonte de indícios de similaridade funcional por meio dessas matrizes de similaridade, a composição dos sítios ativos putativos é utili-

zada para gerar padrões de conservação para os grupos obtidos (como feito no ASMC) e para avaliar a qualidade do agrupamento obtido, como será descrito na Seção 5.5.

Base de dados resultante

Foi construído um banco de dados contendo, para cada par de proteínas, todos os valores de similaridade calculados segundo os tipos de dados descritos anteriormente. A Tabela 5.1 mostra as colunas existentes no banco de dados, suas respectivas fontes e os identificadores das matrizes de similaridades correspondentes. Tais identificadores serão utilizados na apresentação dos resultados para mostrar as combinações de dados feitas pelo sistema de programação genética. É importante lembrar que as matrizes que representam diferenças de valores ou distâncias foram transformadas de modo que valores maiores indicam similaridades maiores, ou seja, diferenças e distâncias menores.

5.3 Integração de Dados via Programação Genética

Cada coluna do banco de dados criado (Tabela 5.1) é utilizada para gerar uma matriz de valores para cada par de proteínas, todas normalizadas no intervalo [0, 1] (ou [-1, 1], nos casos em que há valores negativos). Os tipos de dados em que valores menores são melhores, como é o caso das fontes em que utiliza-se diferenças de valores ou distâncias, têm seus intervalos invertidos. Assim, todas as matrizes são de similaridades e podem ser interpretadas da mesma forma: quanto maior o valor para o par de proteínas, maior a similaridade entre elas segundo a fonte de dados que gerou a matriz.

Para combinar essas matrizes primárias de similaridades em uma única matriz final a ser fornecida como entrada para o algoritmo de agrupamento, é utilizada programação genética (GP), como descrito na Seção 3.2. Assim, o **conjunto de terminais** do sistema de GP é composto por variáveis que representam cada uma das matrizes primárias. Já o **conjunto de operadores** neste trabalho é formado apenas pelo operador de adição. Assim, cada indivíduo do sistema de GP representa uma equação que soma diferentes matrizes primárias. Para cada indivíduo da população, o sistema de GP calcula a matriz de similaridades final aplicando sua equação a cada par de proteínas, e então executa o algoritmo de agrupamento espectral utilizando a matriz obtida, retornando como valor de aptidão do indivíduo a qualidade do agrupamento gerado.

A Equação 5.1 exemplifica um indivíduo do sistema de GP que calcula a similaridade s_{ij} entre cada par de proteínas (i, j) somando o número de anotações em comum entre elas no InterPro, a pontuação para vizinhança conservada do STRING e três vezes o *TM-score* do alinhamento estrutural entre elas. A matriz de similaridades é simétrica, portanto $s_{i,j} = s_{j,i}$.

$$s_{i,j} = interpro_{i,j} + neighborhood_{i,j} + 3strAliScr_{i,j} \tag{5.1}$$

Tabela 5.1: Fontes de dados e respectivos identificadores das matrizes de similaridades entre proteínas empregadas neste trabalho.

Fonte de Dados	Nome	Descrição
Pfam/UniProt	$uniprotA$	Identificador UniProt da proteína A
	$uniprotB$	Identificador UniProt da proteína B
Needleman-Wunsch	$seqAliG$	Pontuação do alinhamento de sequências global
Smith-Waterman	$seqAliL$	Pontuação do alinhamento de sequências local
TM-Align	$strAliSize$	Tamanho do alinhamento estrutural
	$strAliId$	Porcentagem de identidade do alinhamento estrutural
	$strAliScr$	TM-score do alinhamento estrutural
CSM	$csmDist$	Distância entre vetores de assinaturas estruturais
STRING	$neighborhood$	Pontuação para vizinhança genômica conservada
	$fusion$	Pontuação para eventos de fusão gênica
	$cooccurrence$	Pontuação para coocorrência em genomas
	$coexpression$	Pontuação para coexpressão em espécies
EMBOSS Pepstats	$difMolWeight$	Diferença entre pesos moleculares
	$difIsoPoint$	Diferença entre pontos isoelétricos
	$difAliphRes$	Diferença entre conteúdos de resíduos alifáticos
	$difAromRes$	Diferença entre conteúdos de resíduos aromáticos
	$difPolarRes$	Diferença entre conteúdos de resíduos polares
	$difChargedRes$	Diferença entre conteúdos de resíduos carregados
	$difBasicRes$	Diferença entre conteúdos de resíduos básicos
	$difAcidicRes$	Diferença entre conteúdos de resíduos ácidos
	$aaCompDist$	Distância entre vetores de composição de aminoácidos
ExPASy ProtParam	$difInstab$	Diferença entre índices de instabilidade
	$difGRAVY$	Diferença entre índices GRAVY
InterPro	$interpro$	Número de anotações em comum
Gene Ontology	go	Número de termos em comum
Sítio Ativo Putativo	$ASid$	Porcentagem de identidade dos sítios ativos putativos
	$ASscr$	Pontuação BLOSUM62 dos sítios ativos putativos

Os parâmetros do sistema de GP a serem ajustados são o tamanho da população, o número máximo de gerações e as taxas de cruzamento, mutação e reprodução. Ao evoluir uma população de equações que integram as diversas fontes de dados, além dos melhores agrupamentos obtidos, os resultados permitirão verificar os tipos de informação mais úteis para discriminar entre grupos de uma família proteica.

5.4 Agrupamento Espectral

Como discutido no Capítulo 3, neste trabalho a família de proteínas é agrupada utilizando agrupamento espectral que, por sua vez, utiliza o algoritmo K-Médias como passo final. Ambos requerem que o número de grupos K seja previamente definido. Os algoritmos de agrupamento K-Médias e espectral foram implementados como detalhados nos Algoritmos 3.1 e 3.3, respectivamente.

A partir da matriz de similaridades calculada pela equação envolvendo fontes de dados produzida pelo sistema de GP, é definida a matriz de adjacências do grafo de similaridades totalmente conectado. Todas as arestas desse grafo precisam ter valores não-negativos, por isso neste trabalho são testadas duas formas para produzir a matriz de adjacências a partir da matriz de similaridades recebida como entrada pelo agrupamento espectral: ignorando valores negativos ou redimensionando todos os valores para o intervalo [0, 1].

Uma vez definida a matriz de adjacências do grafo de similaridades, é calculada sua matriz Laplaciana normalizada assimétrica, assim como os autovalores e autovetores da mesma. Toma-se então os autovetores correspondentes aos K menores autovalores, cada um representando uma dimensão da nova representação do conjunto de dados. Essa nova matriz de formato $N \times K$, onde N é o número de proteínas, é fornecida como entrada para o algoritmo K-Médias, que, então, divide as proteínas em grupos.

5.5 Critérios de Avaliação

Existem diferentes tipos de índices de validação de agrupamentos, sendo divididos em validação externa, interna e relativa. Na validação externa são empregados critérios não inerentes à base de dados como, por exemplo, um comparativo com rótulos de classe de cada objeto. Métodos de validação interna aplicam critérios derivados do conjunto de dados como distâncias inter- e intragrupo, enquanto métodos de validação relativa comparam agrupamentos gerados por diferentes algoritmos ou por configurações de parâmetros distintas do mesmo algoritmo.

Como o objetivo deste trabalho é detectar subfamílias possivelmente isofuncionais em uma família de proteínas e cada grupo é descrito por um perfil construído segundo a composição dos sítios ativos putativos das proteínas componentes do grupo, são necessárias métricas de qualidade que reflitam numericamente as diferenças entre os perfis de composição dos grupos, o que caracteriza uma validação interna.

Informação Mútua

A informação mútua pontual (PMI, do inglês *Pointwise Mutual Information*) (Church e Hanks, 1990) é uma medida de associação entre dois valores x e y muito empregada na teoria da informação e na

estatística. Segundo Manning et al. (2008), a PMI mede quanto de informação a ocorrência de um valor específico x contribui para fazer a classificação correta de um objeto quanto à classe y. A PMI tem sido utilizada como base de muitos experimentos em linguística computacional para aprender associações entre palavras, além de usada para melhorar a categorização de imagens. Para um dado valor de atributo x e uma classe y, a PMI pode ser usada para decidir se a ocorrência de x é informativa ou não para aquela classe.

De acordo com Bouma (2009), a PMI é uma medida de quanto a probabilidade de uma dada coocorrência de eventos ($p(x, y)$) difere do esperado com base nas probabilidades dos eventos individuais e na suposição de independência ($p(x)p(y)$), sendo calculada pela Equação 5.2. Segundo Church e Hanks (1990), informalmente, a PMI compara a probabilidade de observar os valores x e y ao mesmo tempo com as probabilidades de observá-los independentemente. Se há uma associação genuína entre os valores das variáveis, então a probabilidade conjunta $p(x, y)$ será muito maior que $p(x)p(y)$ e, consequentemente, $PMI(x, y) \gg 0$. Se não há nenhum relacionamento interessante entre os valores x e y, então $p(x, y) \approx p(x)p(y)$ e $PMI(x, y) \approx 0$. Se x e y estão em distribuições complementares, então $p(x, y)$ será muito menor que $p(x)p(y)$, de modo que $PMI(x, y) \ll 0$.

$$PMI(x, y) = ln\frac{p(x, y)}{p(x)p(y)} \tag{5.2}$$

A PMI é chamada "pontual" porque é calculada para dois valores x e y, enquanto a informação mútua (MI, do inglês *Mutual Information*), que é calculada para duas variáveis X e Y, corresponde ao valor esperado de PMI sobre todos os possíveis valores das variáveis, como mostra a Equação 5.3. Apesar do valor de PMI poder ser negativo ou positivo, seu valor esperado sobre todos os eventos conjuntos (ou seja, a MI) é sempre positivo (Bouma, 2009) .

$$MI(X, Y) = \sum_x \sum_y p(x, y)PMI(x, y) \tag{5.3}$$

A informação mútua é utilizada neste trabalho pois mede o quanto de informação uma variável contém sobre outra, ou seja, mede a dependência ou sobreposição de informação entre duas variáveis aleatórias (Cover e Thomas, 2006; Bouma, 2009). A sobreposição é nula quando as duas variáveis são independentes, visto que $p(X)p(Y) = p(X, Y)$. Quando X e Y são perfeitamente correlacionadas, ou seja, uma determina a outra, a MI alcança seu valor máximo.

Neste trabalho, considera-se interessante um grupo que contenha resíduos (quase) exclusivos para as diferentes posições do sítio ativo putativo. Foram testadas várias métricas que combinam de diferentes formas valores de PMI ou MI, além de várias outras medidas relacionadas a entropia, visando a encontrar uma cujos valores mais altos correspondam a agrupamentos considerados interessantes. Foi calculada, por exemplo, para cada posição do sítio ativo, a informação mútua entre os

resíduos que ocorrem na mesma (variável X, em que os valores x correspondem aos resíduos) e os grupos (variável Y, em que os valores y correspondem aos grupos). No entanto, os agrupamentos que apresentavam os melhores valores dessas métricas não foram considerados bons para todas as famílias estudadas.

Sendo assim, uma vez que, nesta tese, um bom grupo é aquele em que um resíduo ocorre (quase) exclusivamente no mesmo para uma dada posição do sítio ativo putativo, cada grupo é comparado com a junção dos demais. Assim, para cada posição p_i, a importância do resíduo r_k para o grupo g_j é medida pela informação mútua pontual entre eles ($PMI_{p_i}(g_j, r_k)$), enquanto a importância do mesmo resíduo na junção dos demais grupos ($\bar{g_j}$) é dada por $PMI_{p_i}(\bar{g_j}, r_k)$. Assim, tem-se $MI_{p_i}(g_j, r_k)$, calculada pela Equação 5.4. As probabilidades $p_{p_i}(g_j, r_k)$ e $p_{p_i}(\bar{g_j}, r_k)$ são estimadas contando o número de vezes que o resíduo r_k ocorre no grupo g_j ou na junção dos demais grupos ($\bar{g_j}$) na posição p_i. Como $PMI_{p_i}(g_j, r_k)$ e $PMI_{p_i}(\bar{g_j}, r_k)$ têm sempre valores com sinais opostos e são considerados importantes para um grupo apenas os resíduos de maior ocorrência no mesmo do que nos demais grupos, apenas os resíduos para os quais $PMI_{p_i}(g_j, r_k) > 0$ são avaliados. Nesse caso, se o resíduo ocorrer nos demais grupos, então $PMI_{p_i}(\bar{g_j}, r_k) < 0$, e a soma dos dois valores reduzirá a importância do mesmo para o grupo g_j. Se $PMI_p(g_j, r_k) \leq 0$, então considera-se $MI_{p_i}(g_j, r_k) = 0$.

$$MI_{p_i}(g_j, r_k) = p_{p_i}(g_j, r_k) PMI_{p_i}(g_j, r_k) + p_{p_i}(\bar{g_j}, r_k) PMI_{p_i}(\bar{g_j}, r_k) \tag{5.4}$$

Dado que em cada grupo pode haver mais de um resíduo em uma dada posição, é calculado o valor de $MI_{p_i}(g_j, r_k)$ para todos os resíduos e feita a média ponderada dos mesmos conforme a frequência de ocorrência do resíduo no grupo. Como lacunas (*gaps*) correspondem a posições que não foram alinhadas na sobreposição estrutural do modelo com a estrutura de referência da família, elas não são consideradas nesse cálculo. Assim, sendo f_k a frequência do resíduo r_k no grupo g_j normalizada pelo tamanho do grupo, tem-se que $MI_{p_i}(g_j) = \sum_k f_k MI_{p_i}(g_j, r_k)$. Finalmente, a métrica para o agrupamento como um todo é dada pela média global calculada pela Equação 5.5, onde P é o número total de posições no sítio ativo putativo, e G é o número de grupos. Esse é o valor de aptidão que o sistema de GP utiliza para determinar quais indivíduos, ou seja, quais agrupamentos, são melhores. Dessa forma, o sistema de GP busca maximizar a informação mútua entre resíduos do sítio ativo putativo e grupos, o que equivale a buscar por grupos que apresentem perfis de composição do sítio ativo característicos.

$$MI = \frac{1}{P}\frac{1}{G}\sum_i \sum_j MI_{p_i}(g_j) \tag{5.5}$$

Capítulo 6
Resultados e Discussão

Neste capítulo são apresentados os principais resultados obtidos aplicando-se a metodologia proposta nesta tese às quatro famílias proteicas descritas: Nucleotidil Ciclases, Serino Proteases, Proteínas Cinases e a família de proteínas com função desconhecida DUF849. Para as três primeiras, são utilizados os rótulos de subfamílias empregados por Melo-Minardi et al. (2010). No entanto, as famílias foram atualizadas em relação àquelas utilizadas por eles, removendo-se as proteínas que foram excluídas do UniProt desde então, sendo 75 Nucleotidil Ciclases, 140 Serino Proteases e 314 Proteínas Cinases. A qualidade dos agrupamentos gerados é avaliada segundo seus respectivos valores de informação mútua (MI), apresentados no Apêndice A e calculados conforme descrito na Seção 5.5. São considerados melhores aqueles agrupamentos que obtêm os maiores valores para essa métrica.

Inicialmente, é feita uma comparação das duas formas consideradas neste trabalho para construir, a partir da matriz de similaridades produzida pelo sistema de programação genética (GP), o grafo de similaridades utilizado pelo algoritmo de agrupamento espectral. Em seguida, são comparados diferentes valores de taxas de cruzamento, mutação e reprodução para, então, determinar a configuração de parâmetros do sistema de GP que leva aos melhores resultados. Então, para cada família proteica, são avaliados os agrupamentos obtidos utilizando as combinações de tipos de dados produzidas pelo sistema de GP, além dos resíduos mais importantes para discriminar diferentes grupos segundo a MI. Os melhores resultados obtidos são comparados aos grupos gerados pelo ASMC (Melo-Minardi et al., 2010) para as famílias Nucleotidil Ciclases, Serino Proteases e Proteínas Cinases, e com aqueles obtidos por Bastard et al. (2014) para a família DUF849. Para permitir uma comparação direta dos resultados, são utilizados os mesmos modelos estruturais e sítios ativos que os autores desses dois trabalhos usaram. O ASMC não mostrou-se muito estável em relação aos grupos produzidos, pois, como mostrado adiante, quando elimina-se dos conjuntos de proteínas usadas por Melo-Minardi et al. (2010) aquelas que foram excluídas do UniProt, os grupos obtidos são bastante diferentes para os mesmos parâmetros de execução.

No restante deste capítulo, as matrizes de similaridades primárias, combinadas pelo sistema de GP para produzir a matriz de similaridades final fornecida ao algoritmo de agrupamento, são representadas pelos respectivos identificadores previamente apresentados na Tabela 5.1. Os logotipos que representam o perfil de composição dos sítios ativos de cada grupo são gerados pelo WebLogo (Crooks et al., 2004). O esquema de cores dos resíduos nos logotipos representa as características químicas dos aminoácidos: **verde** para polares, **roxo** para neutros, **azul** para básicos, **vermelho** para

ácidos e **preto** para hidrofóbicos. Cada coluna no logotipo corresponde a uma posição no sítio ativo putativo, e colunas mais finas denotam a ocorrência de lacunas, que, por sua vez, indicam que não há, no modelo estrutural da proteína em questão, correspondência com aquela posição do sítio ativo da estrutura de referência da família. Já a altura dos aminoácidos no logotipo é proporcional à sua frequência de ocorrência no grupo correspondente.

6.1 Construção do Grafo de Similaridades

O grafo de similaridades utilizado pelo agrupamento espectral é construído a partir da matriz de similaridades calculada pelo sistema de GP. Como descrito na Seção 5.4, neste trabalho é usado o grafo totalmente conectado, cujas arestas precisam ter pesos não-negativos. Por isso, são testadas duas formas de produzir a matriz de adjacências do grafo a partir da matriz de similaridades: considerando somente os valores positivos (e zerando os demais) ou utilizando todos os valores, redimensionando-os para o intervalo [0, 1].

A fim de determinar a melhor forma de construir o grafo de similaridades para o cenário de aplicação desta tese, é feita uma comparação de amostras pareadas entre os resultados obtidos empregando cada construção. Segundo Jain (1991), a análise de amostras pareadas pode ser empregada para comparar dois sistemas quando há uma correspondência um-para-um entre o i-ésimo experimento no sistema A e o i-ésimo experimento no sistema B, como é o caso aqui, em que deseja-se comparar o emprego do grafo de similaridades construído apenas com valores positivos da matriz de similaridades (Sistema A) com o uso do grafo de similaridades que considera todos os valores (Sistema B). Para realizar essa análise, são executados, para cada família de proteínas, pares de experimentos com os mesmos valores de parâmetros, variando apenas a forma de construção do grafo de similaridades, estabelecendo assim a correspondência necessária entre os experimentos. Então, são calculadas as diferenças entre os valores de MI obtidos para cada par de experimentos, além do intervalo de 95% de confiança dessas diferenças, que indica que, com 95% de probabilidade, a diferença entre os valores de MI estará dentro desse intervalo. Ainda segundo Jain (1991), caso esse intervalo inclua zero, então não há diferença estatisticamente significativa entre as qualidades dos agrupamentos obtidos ao empregar cada método de construção do grafo de similaridades, uma vez que a diferença entre valores de MI pode ser nula. Caso não inclua zero, o intervalo de confiança indica qual das formas de construção levou a resultados significativamente melhores.

Os intervalos de confiança das diferenças entre os valores de MI, obtidos empregando cada uma das formas de construção do grafo de similaridades, são apresentados na Tabela 6.1, para cada família de proteínas e cada quantidade de grupos, as quais serão justificadas posteriormente. Intervalos em negrito correspondem àqueles em que há, com 95% de confiança, diferença estatisticamente significativa entre os resultados obtidos empregando as diferentes construções do grafo de similaridades. Como mencionado anteriormente, os valores de MI para cada execução do sistema de GP são apre-

sentados no Apêndice A. As diferenças são calculadas subtraindo, do valor de MI do agrupamento obtido com o grafo gerado apenas com valores positivos (Sistema A), o valor de MI para o agrupamento gerado com o grafo contendo todos os valores (Sistema B). Sendo assim, diferenças positivas indicam que o Sistema A leva a resultados melhores (valores maiores de MI), enquanto diferenças negativas indicam que os resultados são melhores para o Sistema B.

Tabela 6.1: Intervalos de confiança das diferenças entre valores de MI dos agrupamentos gerados empregando cada forma de construção do grafo de similaridades no agrupamento espectral.

Família Proteica	**Grupos**	**Intervalo de Confiança**
Nucleotidil Ciclases	2	[0,00, 0,01]
	3	[-0,01, 0,05]
	4	**[1,77, 2,73]**
	5	**[0,14, 0,32]**
	6	[-0,01, 0,05]
Serino Proteases	3	**[3,04, 3,37]**
	4	[-0,03, 0,89]
	5	**[0,33, 0,95]**
	6	**[0,18, 0,41]**
	7	[-0,04, 0,31]
	8	[-0,09, 0,21]
	9	**[0,25, 0,63]**
	10	[-0,17, 0,23]
	11	**[0,08, 0,31]**
	12	[-0,07, 0,15]
	13	[-0,07, 0,11]
Proteínas Cinases	2	**[0,04, 0,21]**
	3	[-0,19, 0,06]
	4	**[0,66 1,03]**
	5	**[0,06, 0,67]**
	6	**[-0,84, -0,23]**
	7	[-0,28, 0,26]
DUF849	7	**[0,02, 0,63]**
	9	**[0,11, 0,68]**
	32	**[0,10, 0,23]**
	84	[-0,01, 0,05]

Pode-se ver na Tabela 6.1 que, nos casos em que o intervalo de confiança não inclui zero, ou seja, quando o emprego de cada construção do grafo de similaridades leva a resultados significativamente distintos, o intervalo de confiança sempre tem valores positivos. Isso indica que, quando há diferença, são melhores os resultados obtidos ao empregar o grafo que usa apenas os valores positivos da matriz de similaridades. Portanto, essa será a forma de construção do grafo de similaridades adotada no restante deste capítulo.

6.2 Parâmetros do Sistema de Programação Genética

Segundo Poli et al. (2008), a programação genética é, na prática, robusta, de modo que é provável que vários valores de parâmetros funcionem bem. Ainda de acordo com os autores, em muitos casos o principal limitador do tamanho da população é o tempo necessário para avaliar as aptidões dos indivíduos, como é o caso deste trabalho. A metodologia aqui desenvolvida, como apresentado no Capítulo 5, possui cinco passos principais: definição da família proteica a ser estudada; coleta de evidências de similaridade; integração de dados; agrupamento; e avaliação de experimentos. Os passos de integração de dados, agrupamento e parte da avaliação são realizados durante a execução do sistema de GP.

Como discutido anteriormente, cada indivíduo do sistema de GP representa uma equação que combina as diferentes matrizes primárias de similaridades em uma única matriz final fornecida como entrada ao algoritmo de agrupamento. O valor de aptidão que o sistema de GP utiliza para determinar quais indivíduos são bons é a MI do agrupamento resultante. Assim, para cada indivíduo, ocorre o seguinte processamento:

1. cálculo da matriz de similaridades final segundo a equação que o indivíduo representa;
2. execução do algoritmo de agrupamento espectral com a matriz de similaridades calculada; e
3. cálculo do valor de MI do agrupamento resultante.

Devido à carga de processamento necessária à avaliação das aptidões, principalmente no caso de famílias proteicas maiores, em que o passo de agrupamento é mais demorado, neste trabalho considera-se populações de cinquenta indivíduos evoluindo por cinco gerações. Entende-se que se um bom agrupamento de famílias de proteínas não for encontrado até a quinta geração, é pouco provável que seja encontrado em um tempo de execução aceitável.

Como discutido na Seção 3.2, o sistema de GP utiliza os operadores genéticos de cruzamento, mutação e reprodução para criar indivíduos para uma nova população a partir da recombinação de partes de indivíduos da geração anterior. Em geral, a taxa de cruzamento é de cerca de 90%, enquanto a taxa de mutação tipicamente fica em torno de 1% (Poli et al., 2008). No entanto, devido à limitação do tamanho da população e considerando a observação de Luke e Spector (1998) de que há uma tendência geral de que a mutação é mais bem sucedida em populações menores, são utilizadas neste trabalho taxas de mutação maiores do que 1%. Se a soma das taxas de cruzamento e mutação for menor que 100%, é aplicado o operador de reprodução com taxa complementar.

Para cada família de proteínas e cada quantidade de grupos, o sistema de GP foi executado cinco vezes para cada configuração de parâmetros e cada um dos métodos de construção do grafo de similaridades discutidos anteriormente. A Tabela 6.2 apresenta as configurações das taxas dos operadores genéticos empregadas neste trabalho, assim como o número de vezes que cada configuração aparece

entre os experimentos que produziram os maiores valores de MI para cada família de proteínas e cada quantidade de grupos, respeitando multiplicidades no caso de empate.

Tabela 6.2: Configurações de taxas dos operadores do sistema de GP e número de ocorrências de cada uma entre os melhores resultados para cada família proteica e cada quantidade de grupos.

Cruzamento	**Reprodução**	**Mutação**	**Ocorrências**
70%	10%	20%	12
70%	20%	10%	7
80%	5%	15%	15
80%	15%	5%	6
80%	20%	0%	8
85%	5%	10%	6
85%	10%	5%	9
90%	5%	5%	8
90%	10%	0%	11

Pode-se observar na Tabela 6.2 que a configuração de parâmetros do sistema de GP que mais ocorreu entre os melhores resultados para cada família de proteínas e cada quantidade de grupos foi utilizando 80% de cruzamento, 5% de reprodução e 15% de mutação. Por isso, os resultados apresentados no restante deste capítulo correspondem àqueles obtidos com essa configuração de parâmetros e empregando o grafo de similaridades construído a partir dos valores positivos da matriz de similaridades, como anteriormente mencionado.

6.3 Estudo de Caso I: Nucleotidil Ciclases

Após a remoção, do conjunto de proteínas empregado por Melo-Minardi et al. (2010), de 75 Nucleotidil Ciclases que tornaram-se obsoletas no UniProt, essa família ficou com 461 proteínas, sendo 186 Adenilato Ciclases e 275 Guanilato Ciclases, de acordo com os rótulos de subfamílias empregados pelos autores. Por isso, o sistema de GP foi executado para dividir a família em dois grupos. Com os mesmos parâmetros usados pelos autores para o conjunto original de proteínas, o ASMC produziu, para o conjunto atualizado, um agrupamento hierárquico no qual o primeiro nível dividiu essa família em três grupos, e o segundo nível, em seis. Por essas razões, o sistema de GP foi executado com dois a seis grupos. Os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados nas Tabelas A.1 a A.5 do Apêndice A.

A Figura 6.1 apresenta o logotipo que ilustra a composição do sítio ativo putativo das 461 Nucleotidil Ciclases, no qual pode-se ver que a família apresenta resíduos bastante conservados para a maioria das posições. A título de ilustração da dificuldade em encontrar o agrupamento ótimo, há mais de cem mil formas de combinar as proteínas dessa família em dois grupos, mais de dezesseis milhões de formas de combiná-las em três, e quase treze bilhões de formas de combiná-las em seis

grupos. Como é infactível testar todas essas combinações para definir a melhor delas, são necessários métodos heurísticos para chegar a um agrupamento razoavelmente bom, como feito neste trabalho.

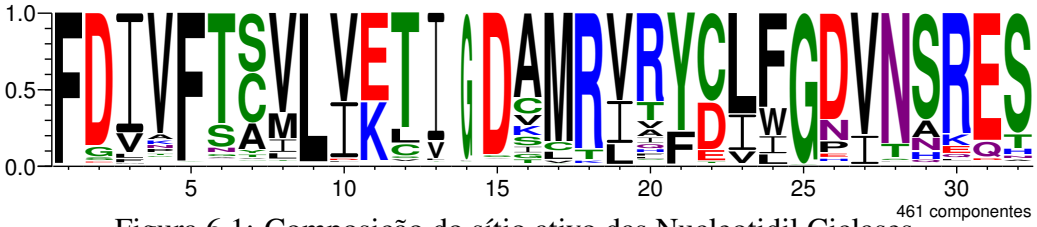


Figura 6.1: Composição do sítio ativo das Nucleotidil Ciclases.

São apresentadas na Tabela 6.3 as combinações de dados produzidas pelo sistema de GP que levaram aos melhores resultados para as Nucleotidil Ciclases, empregando o grafo de similaridades construído com apenas os valores positivos das matrizes de similaridades calculadas por essas combinações, e utilizando 80% de cruzamento, 5% de reprodução e 15% de mutação. Uma vez que a métrica de qualidade, a MI, é baseada na composição do sítio ativo, era esperado que as matrizes de similaridades primárias nele baseadas teriam participação nas combinações de dados que produziram os melhores resultados. Das dez equações apresentadas na Tabela 6.3, nove contêm a identidade dos sítios ativos ($ASid$), enquanto cinco contêm as pontuações do alinhamento par-a-par dos mesmos ($ASscr$). Outros tipos de dados que destacaram-se para essa família de proteínas foram as pontuações dos alinhamentos de sequências global ($seqAliG$) e local ($seqAliL$), que aparecem, respectivamente, em sete e cinco equações, assim como a identidade do alinhamento estrutural ($strAliId$) e a diferença na composição de resíduos alifáticos ($difAliphRes$), que aparecem, cada uma, em seis equações. Destaca-se a grande quantidade de tipos de dados que foram utilizados pelo sistema de GP a fim de particionar essa família em dois grupos.

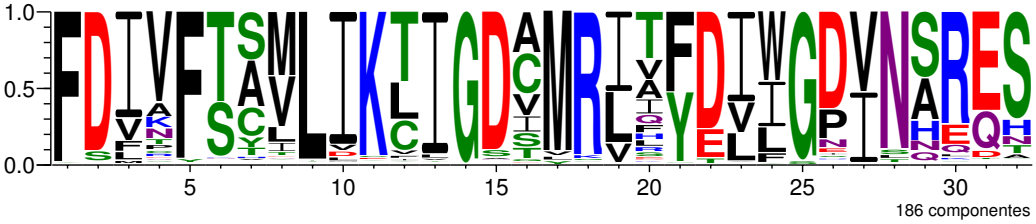
Divisão das Nucleotidil Ciclases em dois grupos

Quando o sistema de GP é executado para dividir a família em duas, as quatro equações apresentadas na Tabela 6.3 levam à formação dos mesmos grupos, cujos logotipos e composições são apresentados na Figura 6.2. Observa-se que esses grupos são quase idênticos aos rótulos de subfamílias utilizados por Melo-Minardi et al. (2010), exceto por duas discordâncias.

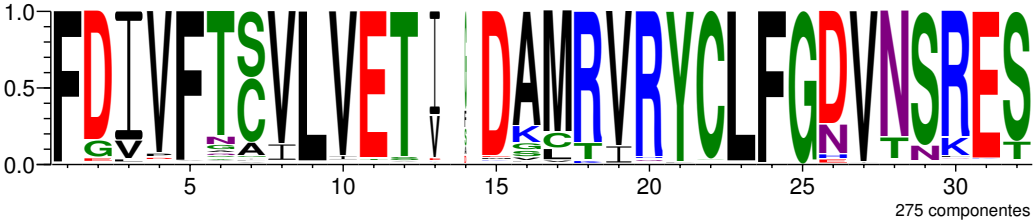
A proteína rotulada como Guanilato Ciclase inserida no Grupo I, em que predominam as Adenilato Ciclases, tem identificador UniProt Q5UFR4. Apesar de não ter sido manualmente curada, essa proteína foi anotada com o termo *adenylate cyclase activity* da ontologia de função molecular do GO, e o termo *cAMP biosynthetic process* da ontologia de processo biológico. Isso sugere que o agrupamento produzido pelo sistema de GP atribuiu essa proteína corretamente ao grupo de Adenilato Ciclases, divergindo do rótulo de subfamília usado por Melo-Minardi et al. (2010).

Tabela 6.3: Combinações de dados que levaram à obtenção dos melhores agrupamentos para a família de Nucleotidil Ciclases para o grafo de similaridades construído apenas com valores positivos e com 80% de cruzamento, 5% de reprodução e 15% de mutação.

Grupos	Repetição	Equação
2	1	$4ASid + ASscr + cooccurence + 4csmDist + difAcidicRes + 3difAliphRes + difAromRes + difBasicRes + difChargedRes + 2difInstab + difIsoPoint + 2difPolarRes + interpro + neighborhood + 3seqAliG + 2strAliId + 2strAliSize$
	2	$ASid + 3ASscr + aaCompDist + 2coexpression + 2cooccurence + csmDist + 2difAcidicRes + difAliphRes + difBasicRes + difChargedRes + 2difGRAVY + 3difInstab + 2difIsoPoint + 2difMolWeight + 2go + interpro + 5seqAliG + strAliId + strAliScr$
	3	$ASid + 2ASscr + aaCompDist + cooccurence + difAcidicRes + difAliphRes + difAromRes + difBasicRes + difChargedRes + difIsoPoint + difMolWeight + difPolarRes + 2seqAliL + strAliId + 4strAliScr$
	4	$4ASid + 4ASscr + 6coexpression + cooccurence + 5csmDist + difAcidicRes + 4difAliphRes + 3difAromRes + difBasicRes + 4difChargedRes + difGRAVY + 3difInstab + 2difIsoPoint + difMolWeight + 5difPolarRes + 2go + 2interpro + 4neighborhood + seqAliG + 2seqAliL + 2strAliId + 3strAliScr + 3strAliSize$
3	1	$ASscr + difMolWeight + seqAliG$
4	2	$ASid + aaCompDist + neighborhood$
5	1	$ASid + seqAliG + strAliSize$
	3	$2ASid + difAliphRes + go + seqAliG + 2seqAliL + 2strAliId$
	4	$ASid + seqAliL$
6	3	$5ASid + difAliphRes + 2go + seqAliG + 3seqAliL + 2strAliId$



(a) Grupo I: 185 Adenilato Ciclases e 1 Guanilato Ciclase (Tabela 6.4)



(b) Grupo II: 274 Guanilato Ciclases e 1 Adenilato Ciclase (Tabela 6.5)

Figura 6.2: Divisão das Nucleotidil Ciclases em dois grupos pelo sistema de GP.

Tabela 6.4: Exceção no Grupo I obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em dois grupos.

Proteína	**Anotações**
Q5UFR4	Não-revisada. Apresenta anotação dos termos GO *adenylate cyclase activity* e *cAMP biosynthetic process*.

Tabela 6.5: Exceção no Grupo II obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em dois grupos.

Proteína	**Anotações**
Q7RKA2	Não-revisada. Não apresenta anotação. Nome sugerido: *Guanylyl cyclase enzyme-related*.

Já a proteína rotulada como Adenilato Ciclase que foi inserida no Grupo II, das Guanilato Ciclases, tem identificador Q7RKA2. Analisando sua entrada no UniProt, observa-se que ela também não foi manualmente curada e que não há qualquer anotação específica de uma ou de outra subfamília. A única informação existente relacionada a subfamília é que o nome submetido dessa proteína é *Guanylyl cyclase enzyme-related*, que é um tipo fraco de anotação, mas que sugere que o rótulo de subfamília pode estar novamente equivocado e que o sistema de GP pode tê-la posicionado corretamente no grupo de Guanilato Ciclases.

Neste trabalho, considera-se interessante um grupo que contenha, em determinadas posições do sítio ativo, resíduos que são exclusivos ou quase exclusivos às proteínas que o compõem. Essas são posições determinantes de especificidade (SDPs, do inglês *Specificity Determining Positions*). Como apresentado no Capítulo 5, o cálculo da MI de um agrupamento é construído somando os valores de MI para cada resíduo, em cada posição, em cada grupo. Esses valores parciais permitem avaliar numericamente quais resíduos, em quais posições, mais diferenciam um grupo dos demais. Os dez resíduos mais importantes para discriminar entre os dois grupos gerados pelo sistema de GP para as Nucleotidil Ciclases são listados a seguir, em ordem decrescente do valor de MI do resíduo na sua respectiva posição. Os resíduos em negrito correspondem àqueles que conhecidamente alteram a especificidade de substrato da família, como descrito a seguir.

- ◇ **Grupo I: K (11)**, I (10), G (14), **D (22)**, I (23), F (21), I (13), W (24), I (27) e M (8); e
- ◇ **Grupo II: C (22)**, **E (11)**, F (24), V (10), R (20), L (23), V (19), 8 (V), 21 (Y) e A (16).

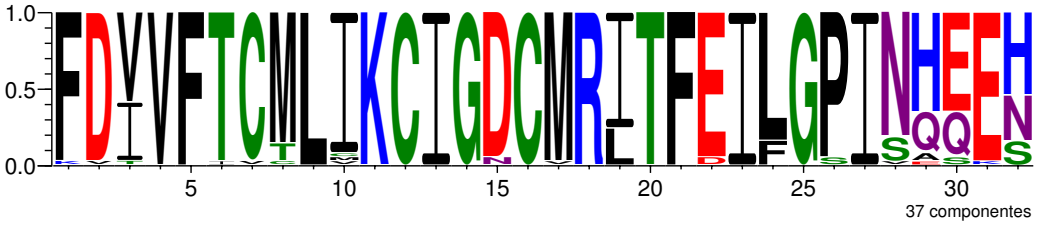
Segundo Melo-Minardi et al. (2010) e Hannenhalli e Russell (2000), a mutação de apenas dois resíduos é suficiente para alterar a especificidade de uma proteína de Guanilato para Adenilato Ciclase. Considerando a cadeia A da estrutura com identificador PDB 3ET6, essas mutações são a troca do glutamato (E) na posição 523 e da cisteína (C) na posição 592 nas Guanilato Ciclases, respectivamente, por uma lisina (K) e um aspartato (D), nas Adenilato Ciclases. A mutação da cisteína para

aspartato favorece a alteração de especificidade de GTP (Guanilato Ciclases) para ATP (Adenilato Ciclases) porque impede a ligação de guanidina por criar uma repulsão eletrostática entre o aspartato e a guanina do substrato, além de estabilizar a ligação de adenina. Já a mutação de glutamato para lisina cria uma ponte de hidrogênio com adenina, favorecendo a ligação do ATP em detrimento do GTP. As posições 523 e 592 na estrutura 3ET6 correspondem às posições 11 e 22 no sítio ativo apresentado. De fato, K na posição 11 e D na posição 22 estão entre os resíduos considerados mais importantes para diferenciar o Grupo I (Adenilato Ciclases), conforme o valor de MI, enquanto E na posição 11 e C na posição 22 são os dois resíduos mais importantes para diferenciar o Grupo II (Guanilato Ciclases). Isso mostra que o sistema de GP foi capaz de gerar grupos cujas posições que mais os diferenciam correspondem àquelas que conhecidamente definem a especificidade das subfamílias.

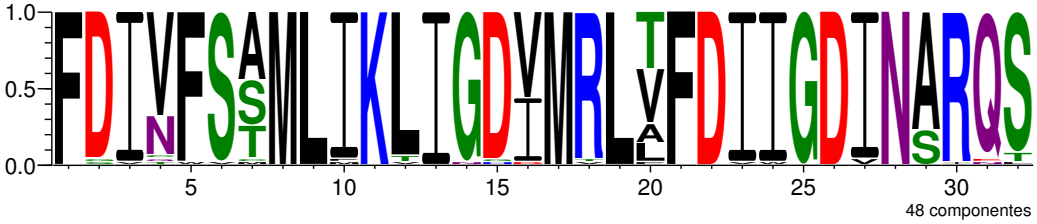
Ainda segundo os autores, outros resíduos conservados nas diferentes subfamílias descritos na literatura são, para as posições 590, 593 e 594 da estrutura 3ET6, arginina (R), leucina (L) e fenilalanina (F) nas Guanilato Ciclases, substituídas, respectivamente, por glicina (G), isoleucina (I) e triptofano (W) nas Adenilato Ciclases. No entanto, ainda não foram identificados os papéis desses aminoácidos na especificidade de substrato. Essas posições da estrutura correspondem, respectivamente, às posições 20, 23 e 24 no sítio ativo apresentado. Pode-se observar que, exceto pela glicina na posição 20 das Adenilato Ciclases, todas as demais estão entre os resíduos mais importantes para discriminar cada família: I (23) e W (24) para o Grupo I, das Adenilato Ciclases, e R (20), L (23) e F (24) para o Grupo II, das Guanilato Ciclases.

Divisão das Nucleotidil Ciclases em três grupos

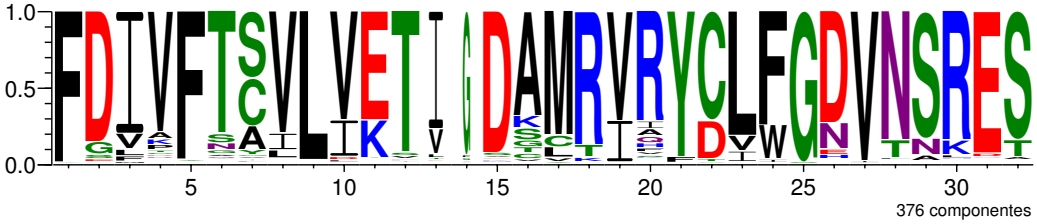
Apesar de já separar com sucesso a família de Nucleotidil Ciclases em suas duas subfamílias, o sistema de GP foi executado dividindo-a em três grupos a título de comparação com o ASMC, que, no primeiro nível do agrupamento hierárquico, divide a família em três grupos, cujas composições e logotipos são apresentados na Figura 6.3. Pode-se observar que, ao invés de encontrar grupos relacionados às duas subfamílias existentes, o ASMC priorizou subgrupos das Adenilato Ciclases, enquanto a maior parte da família continua em um grupo grande em relação aos demais contendo todas as Guanilato Ciclases e a maioria das Adenilato Ciclases. Esse resultado é radicalmente diferente do obtido com o conjunto original de proteínas usado por Melo-Minardi et al. (2010), cujos logotipos e composições são apresentados na Figura B.1 do Apêndice B. O agrupamento gerado com o conjunto original de proteínas muito assemelha-se aos rótulos de subfamílias e ao resultado obtido com o sistema de GP para dois grupos, apresentado anteriormente. Conclui-se, então, que o ASMC não é muito estável em termos dos agrupamentos obtidos, uma vez que a eliminação de apenas 14% da família que tornou-se obsoleta no UniProt levou a um agrupamento de qualidade muito inferior.



(a) Grupo I: 37 Adenilato Ciclases



(b) Grupo II: 48 Adenilato Ciclases



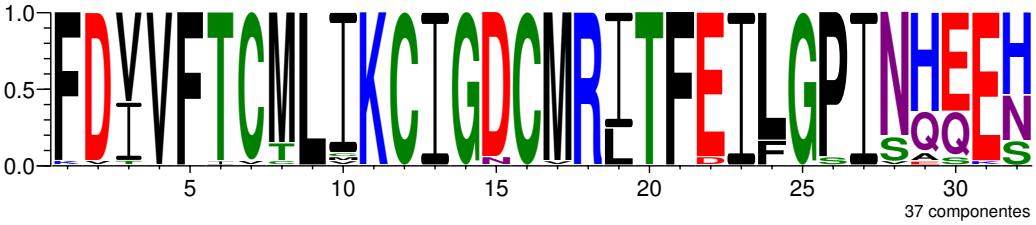
(c) Grupo III: 275 Guanilato Ciclases e 101 Adenilato Ciclases

Figura 6.3: Divisão das Nucleotidil Ciclases em três grupos no primeiro nível do agrupamento hierárquico do ASMC.

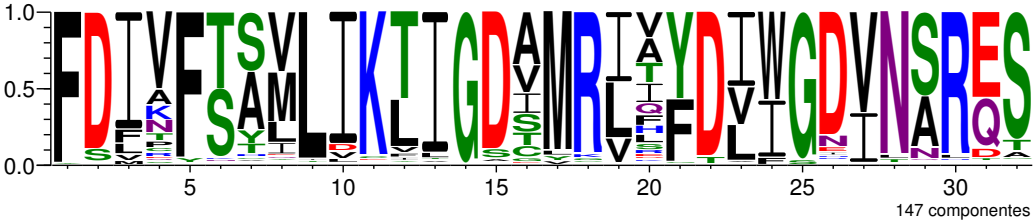
O melhor agrupamento produzido pelo sistema de GP para três grupos utiliza a matriz de similaridades calculada pela combinação da pontuação do alinhamento dos sítios ativos, da diferença entre pesos moleculares e da pontuação do alinhamento global de sequências, como anteriormente apresentado na Tabela 6.3. Nos logotipos dos grupos, apresentados na Figura 6.4, observa-se que é obtido o mesmo subgrupo de 37 Adenilato Ciclases encontrado pelo ASMC, mantendo, no entanto, a divisão das subfamílias nos outros dois grupos.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

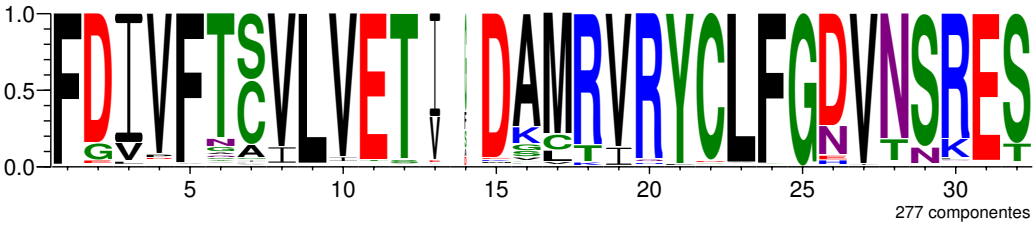
- ◇ **Grupo I:** C (12), E (22), P (26), C (16), L (24), T (20), I (27), H (29), E (30) e F (21);
- ◇ **Grupo II:** D (22), K (11), G (14), I (10), W (24), S (6), I (13), I (24), L (12) e Q (31); e
- ◇ **Grupo III:** C (22), E (11), R (20), V (10), F (24), L (23), V (19), V (8), Y (21) e T (12).



(a) Grupo I: 37 Adenilato Ciclases



(b) Grupo II: 146 Adenilato Ciclases e 1 Guanilato Ciclase (Tabela 6.6)



(c) Grupo III: 274 Guanilato Ciclases e 3 Adenilato Ciclases (Tabela 6.7)

Figura 6.4: Divisão das Nucleotidil Ciclases em três grupos pelo sistema de GP.

Tabela 6.6: Exceção no Grupo II obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em três grupos.

Proteína	**Anotações**
Q5UFR4	Não-revisada. Apresenta anotação dos termos GO *adenylate cyclase activity* e *cAMP biosynthetic process*.

Tabela 6.7: Exceções no Grupo III obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em três grupos.

Proteína	**Anotações**
A0CM46	Não-revisada. Não apresenta anotação.
A0DT50	Não-revisada. Não apresenta anotação.
A0DT51	Não-revisada. Não apresenta anotação.

Novamente, a proteína rotulada como Guanilato Ciclase que foi inserida no grupo cuja maioria é de Adenilato Ciclases é a de identificador UniProt Q5UFR4 que, como mostrado anteriormente, foi equivocadamente rotulada. Já as três Adenilato Ciclases que foram inseridas no grupo das Guanilato Ciclases têm identificadores UniProt A0DT50, A0CM46 e A0DT51. Nenhuma das três foi manualmente revisada, nem apresenta qualquer anotação que remeta a uma das duas subfamílias. Por

Tabela 6.8: Exceções no Grupo III obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em seis grupos.

Proteína	Anotações
Q4XVA2	Não-revisada. Não apresenta anotação.
Q4YH96	Não-revisada. Não apresenta anotação.
Q5UFR4	Não-revisada. Apresenta anotação dos termos GO *adenylate cyclase activity* e *cAMP biosynthetic process*.

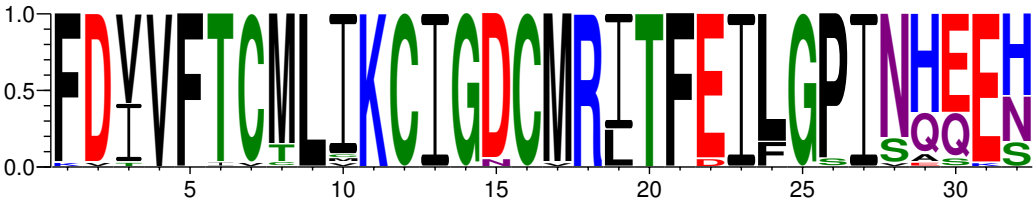
isso, não é possível afirmar se o rótulo de subfamília adotado por Melo-Minardi et al. (2010) ou o agrupamento produzido pelo sistema de GP é correto para essas proteínas.

Na divisão da família em dois grupos pelo sistema de GP, a proteína de identificador Q7RKA2, rotulada como Adenilato Ciclase, foi incluída junto às Guanilato Ciclases. No entanto, na execução do sistema de GP com três grupos, ela é agrupada com as Adenilato Ciclases. Isso mostra uma vantagem do agrupamento de particionamento utilizado no sistema de GP em relação ao agrupamento hierárquico usado pelo ASMC: uma vez que um nó da árvore que representa o agrupamento hierárquico é subdividido, não é possível que uma proteína mude de um ramo a outro da árvore. Então, caso uma subdivisão seja realizada equivocadamente durante o processo, o erro será propagado aos outros níveis da hierarquia e não poderá ser consertado. Já no caso de um agrupamento de particionamento, em que define-se *a priori* o número de grupos buscados, proteínas podem migrar para um grupo mais adequado conforme o número de grupos aumenta, o que seria equivalente a uma mudança de ramo na árvore para reparar um erro.

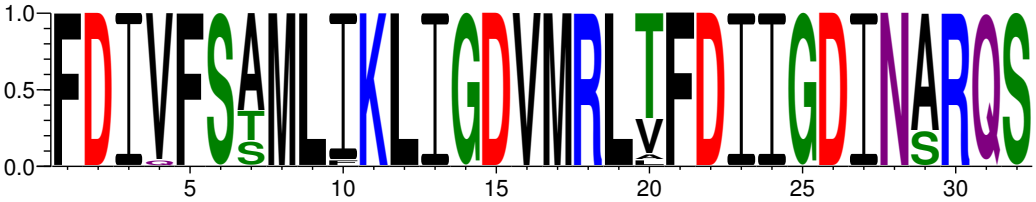
Divisão das Nucleotidil Ciclases em seis grupos

As subfamílias das Nucleotidil Ciclases somente foram separadas no segundo nível do agrupamento hierárquico gerado pelo ASMC. Em relação ao primeiro nível, o grupo de 37 Adenilato Ciclases foi mantido, mas os demais foram subdivididos, como mostra a Figura 6.5. Pode-se observar que, para gerar um grupo com a subfamília de Guanilato Ciclases, o ASMC precisou fragmentar as Adenilato Ciclases em cinco subgrupos, um dos quais contém apenas duas proteínas, mesmo não havendo grande variabilidade nessa subfamília.

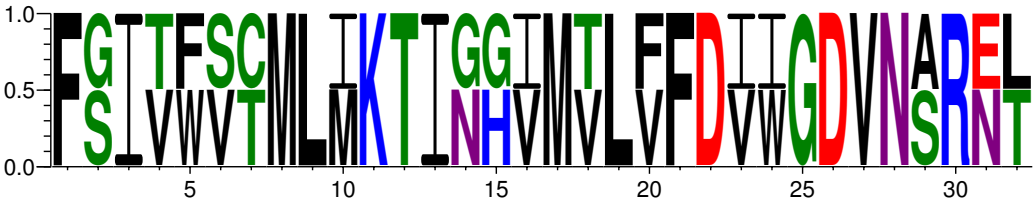
A título de comparação, o sistema de GP foi executado para dividir a família de Nucleotidil Ciclases em seis grupos. Conforme mostrado na Tabela 6.3, a equação que obteve o melhor resultado combina seis tipos de dados, com ênfase na identidade do sítio ativo putativo, como era esperado, uma vez que a métrica de qualidade é baseada no mesmo. Os logotipos e composições dos grupos são apresentados na Figura 6.6, onde pode-se observar a divisão de cada subfamília em três subgrupos.



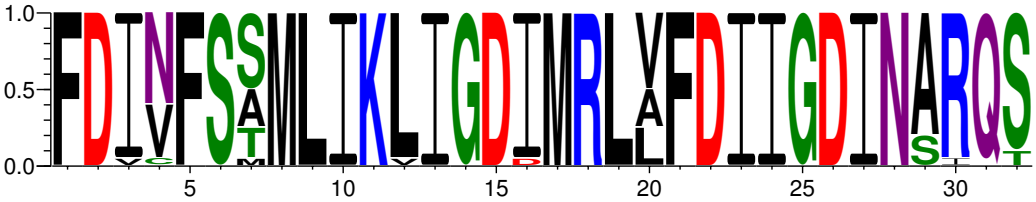
(a) Grupo I: 37 Adenilato Ciclases



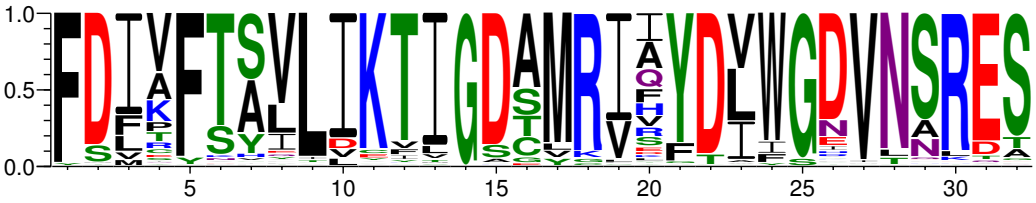
(b) Grupo II.A: 26 Adenilato Ciclases



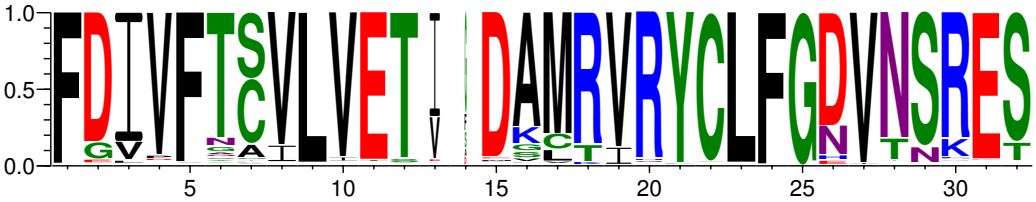
(c) Grupo II.B: 2 Adenilato Ciclases



(d) Grupo II.C: 20 Adenilato Ciclases

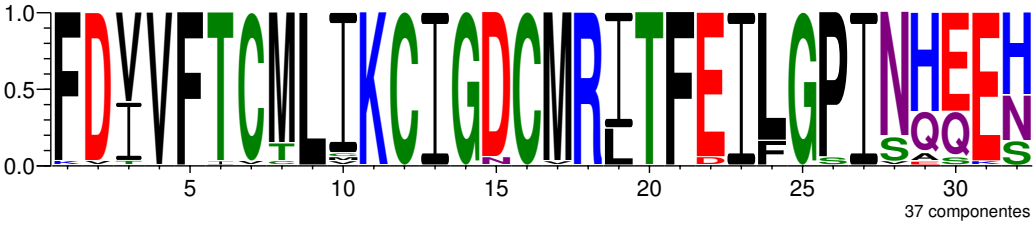


(e) Grupo III.A: 99 Adenilato Ciclases e 1 Guanilato Ciclase

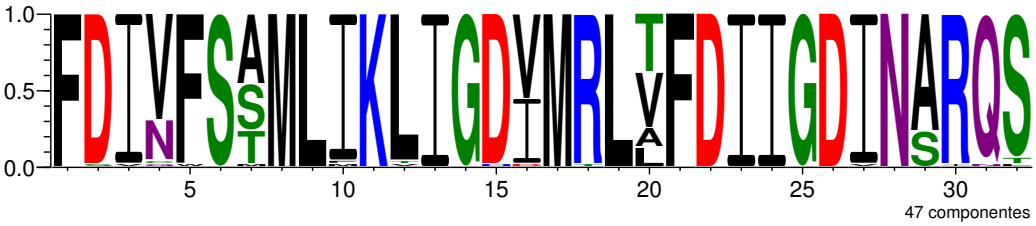


(f) Grupo III.B: 274 Guanilato Ciclases e 2 Adenilato Ciclases

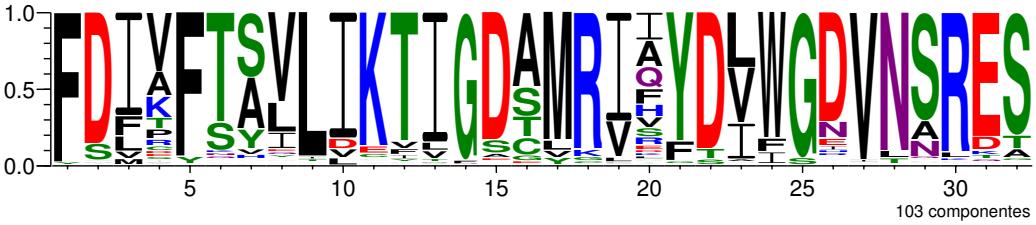
Figura 6.5: Divisão das Nucleotidil Ciclases em seis grupos no segundo nível do agrupamento hierárquico do ASMC.



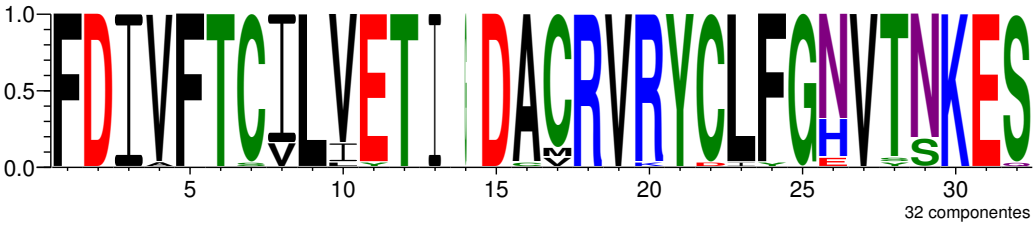
(a) Grupo I: 37 Adenilato Ciclases



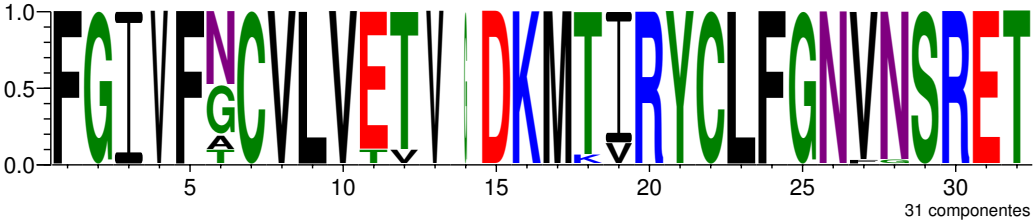
(b) Grupo II: 47 Adenilato Ciclases



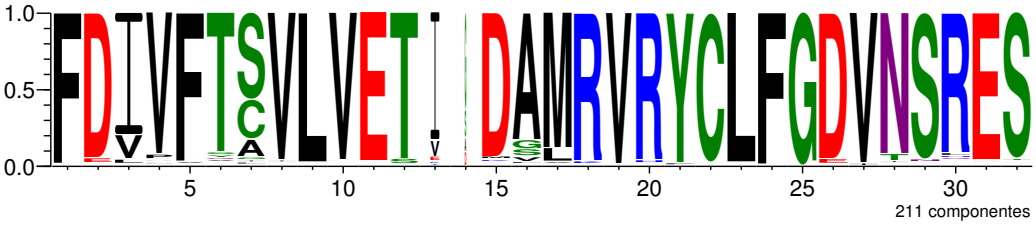
(c) Grupo III: 100 Adenilato Ciclases e 3 Guanilato Ciciases (Tabela 6.8)



(d) Grupo IV: 31 Guanilato Ciclases e 1 Adenilato Ciclase (Tabela 6.9)



(e) Grupo V: 31 Guanilato Ciclases



(f) Grupo VI: 210 Guanilato Ciclases e 1 Adenilato Ciclase (Tabela 6.10)

Figura 6.6: Divisão das Nucleotidil Ciclases em seis grupos pelo sistema de GP.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

- ◇ **Grupo I:** C (12), E (22), P (26), C (16), L (24), T (20), I (27), H (29), E (30) e F (21);
- ◇ **Grupo II:** L (12), Q (31), I (24), L (19), S (6), M (8), I (27), F (21), I (23) e A (29);
- ◇ **Grupo III:** W (24), D (22), K (11), I (10), I (19), G (14), V (23), I (13), L (8) e I (20);
- ◇ **Grupo IV:** K (30), C (17), T (28), I (8), N (29), C (7), N (26), V (19), A (16) e R (20);
- ◇ **Grupo V:** K (16), G (2), T (18), T (32), N (26), V (13), N (6), C (7), I (19) e G (6); e
- ◇ **Grupo VI:** C (22), V (10), E (11), V (19), F (24), R (20), V (8), L (23), A (16) e S (29).

No Grupo III, cuja maioria é de Adenilato Ciclases, há três proteínas rotuladas como Guanilato Ciclases, cujos identificadores UniProt são Q4XVA2, Q4YH96 e Q5UFR4. Nenhuma das três entradas do UniProt foi manualmente revisada. Anteriormente, concluiu-se que o rótulo da proteína Q5UFR4 está equivocado. Quanto às demais, nenhuma delas têm qualquer anotação que remeta a uma das duas subfamílias. Elas possuem apenas o nome sugerido de *guanylyl cyclase, putative*, o que é uma forma muito fraca de anotação. Por isso, não é possível afirmar se o que está correto são os rótulos de subfamílias adotados por Melo-Minardi et al. (2010) ou o grupo de Adenilato Ciclases em que foram inseridas pelo sistema de GP.

Nos Grupos IV e VI, que são compostos em maioria por Guanilato Ciclases, há uma proteína rotulada como Adenilato Ciclase em cada um, cujos identificadores UniProt são, respectivamente, A0DT51 e A7RY93. Nenhuma das duas apresenta qualquer anotação que remeta a uma das subfamílias, por isso, novamente, não é possível afirmar se o rótulo está errado ou se o sistema de GP as atribuiu aos grupos errados.

Tabela 6.9: Exceção no Grupo IV obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em seis grupos.

Proteína	**Anotações**
A0DT51	Não-revisada. Não apresenta anotação.

Tabela 6.10: Exceção no Grupo VI obtido pelo sistema de GP para a divisão das Nucleotidil Ciclases em seis grupos.

Proteína	**Anotações**
A7RY93	Não-revisada. Não apresenta anotação.

O estudo de caso para a família de Nucleotidil Ciclases mostrou que os grupos obtidos pelo sistema de GP apresentam maior concordância com a divisão da família em suas duas subfamílias do que os grupos produzidos pelo ASMC. Além disso, quando há uma quantidade maior de grupos do que o número de subfamílias existente, o sistema de GP tende a produzir grupos mais contrastantes do que o ASMC.

6.4 Estudo de Caso II: Serino Proteases

Após a eliminação de 140 sequências que tornaram-se obsoletas no UniProt desde sua utilização por Melo-Minardi et al. (2010), essa família passou a ter 1.533 proteínas, sendo 43 anotadas como Elastases, 26 como Quimotripsinas e 1.464 como Tripsinas, segundo os rótulos de subfamílias empregados pelos autores. Adicionalmente, eles reportaram um subgrupo de treze Tripsinas que correspondem a Calicreínas. Por isso, o sistema de GP foi executado dividindo a família em três e quatro grupos. O grande desbalanceamento é um desafio para a separação dessa família em suas subfamílias.

Quando executado para o conjunto atualizado de proteínas com os mesmos parâmetros utilizados por Melo-Minardi et al. (2010) para o conjunto original (-A 1,0 e -C 0,25), o ASMC não dividiu essa família. Então, o principal parâmetro (-C) foi reduzido de 0,05 em 0,05 até que fosse encontrado um valor que a dividisse. Esse valor foi de 0,15, que produziu um agrupamento hierárquico no qual o primeiro nível dividiu a família em quatro grupos, e o segundo, em onze grupos. Para ilustrar a dificuldade do problema, há quase 600 milhões de formas de combinar as 1.533 proteínas da família em três grupos e $2{,}6 \times 10^{27}$ formas de combiná-las em onze grupos. Por essas razões, e considerando a grande variabilidade de resíduos em algumas posições do sítio ativo putativo mostrado na Figura 6.7, além do imenso desbalanceamento entre o número de proteínas em cada subfamília, o sistema de GP foi executado para essa família com três a treze grupos. Os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados nas Tabelas A.6 a A.16 do Apêndice A.

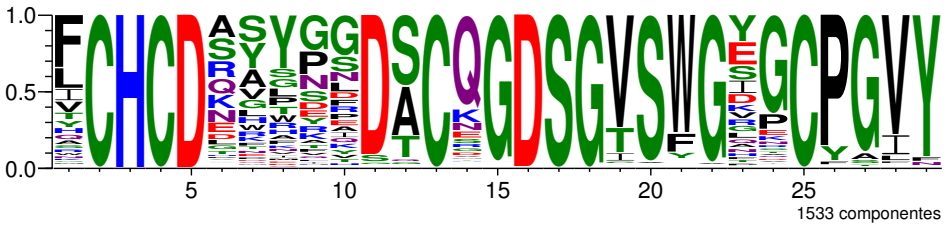


Figura 6.7: Composição do sítio ativo das Serino Proteases.

A Tabela 6.11 apresenta as combinações de dados produzidas pelo sistema de GP que levaram aos melhores resultados para as Serino Proteases empregando o grafo de similaridades construído apenas com valores positivos das matrizes de similaridades calculadas por essas equações, e aplicando as taxas de 80% de cruzamento, 5% de reprodução e 15% de mutação. Observa-se, novamente, a

grande utilidade das matrizes de similaridades primárias baseadas na identidade e na pontuação dos sítios ativos, uma vez que a métrica de qualidade baseia-se na composição do mesmo. Entre as onze equações, dez envolvem a identidade do sítio ativo ($ASid$), enquanto cinco utilizam a pontuação do seu alinhamento ($ASscr$). Outros tipos de dados que destacaram-se foram as pontuações do alinhamento de sequências global ($seqAliG$) e a identidade do alinhamento estrutural ($strAliId$).

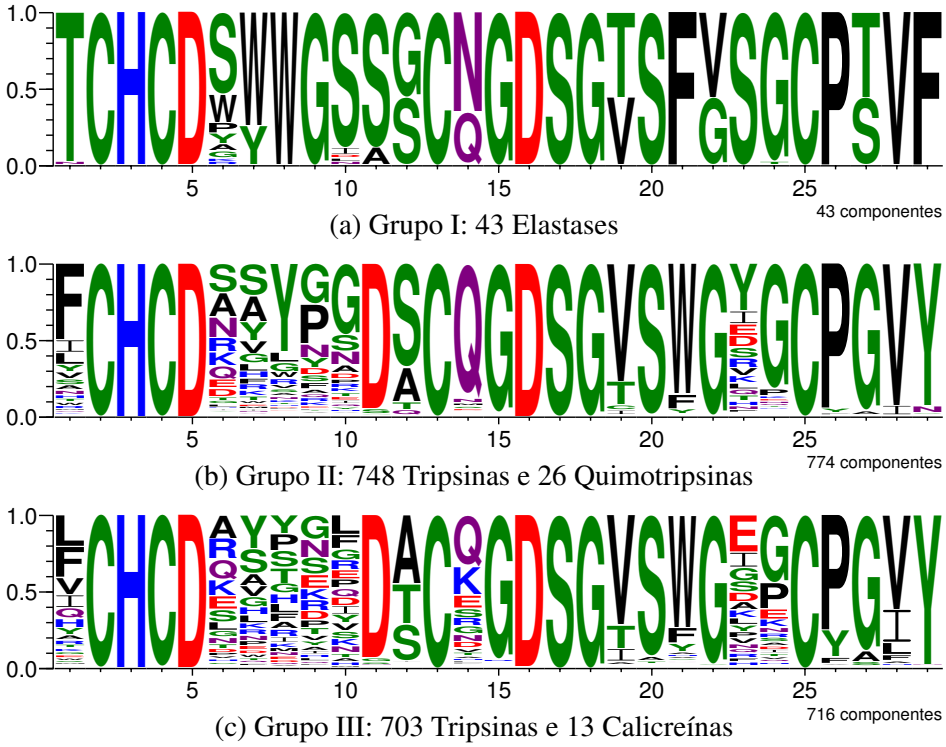
Tabela 6.11: Combinações de dados que levaram à obtenção dos melhores agrupamentos para a família de Serino Proteases para o grafo de similaridades construído apenas com valores positivos e com 80% de cruzamento, 5% de reprodução e 15% de mutação.

Grupos	**Repetição**	**Equação**
3	1 a 5	$ASid$
4	1	$ASid + ASscr + seqAliG$
5	2	$3ASscr + seqAliG$
6	2	$2ASid + strAliId$
7	4	$2ASid + strAliId$
8	5	$ASid + strAliId$
9	1	$2ASid + ASscr + seqAliG$
10	3	$ASid + 3strAliScr$
11	1	$2ASid + ASscr + seqAliG$
12	1	$3ASid + 2ASscr + go$
13	1	$ASid + cooccurence$

Divisão das Serino Proteases em três grupos

Quando o sistema de GP é executado para dividir a família em três grupos visando à identificação das três principais subfamílias (Elastases, Quimotripsinas e Tripsinas), todas as cinco repetições obtêm o mesmo agrupamento empregando apenas a identidade do sítio ativo, como mostrado na Tabela 6.11. A Figura 6.8 apresenta os logotipos e composições dos grupos produzidos pelo sistema de GP. Pode-se observar que apenas a subfamília de Elastases foi separada das demais. A grande discrepância na quantidade de Tripsinas, em relação às demais subfamílias, torna necessária a utilização de uma quantidade maior de grupos para detectá-las, uma vez que a variabilidade entre as Tripsinas dificulta a identificação das subfamílias pequenas.

Pode-se observar que apenas a composição do Grupo I é homogênea, com pouquíssima variabilidade. Já os outros grupos mostram a heterogeneidade da composição das Tripsinas. Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição. Os resíduos em negrito correspondem àqueles que conhecidamente alteram a especificidade de substrato da família, como descrito em seguida.



(a) Grupo I: 43 Elastases

(b) Grupo II: 748 Tripsinas e 26 Quimotripsinas

(c) Grupo III: 703 Tripsinas e 13 Calicreínas

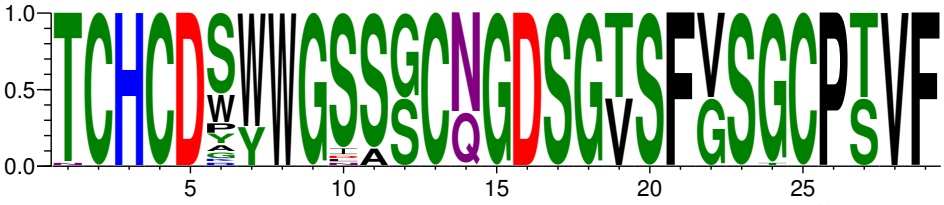
Figura 6.8: Divisão das Serino Proteases em três grupos pelo sistema de GP.

- **Grupo I:** F (29), **W (8)**, T (1), **S (11)**, S (23), F (21), W (7), **V (22)**, S (10) e **T (27)**;
- **Grupo II:** Q (14), Y (8), G (10), S (12), G (24), F (1), Y (23), V (28), P (9) e P (26); e
- **Grupo III:** T (12), A (12), K(14), E (23), Y (26), I (28), L (10), P (24), L (1) e P (8).

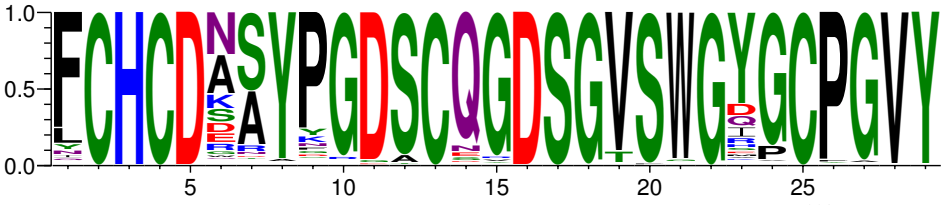
Segundo Melo-Minardi et al. (2010) e Hannenhalli e Russell (2000), os resíduos determinantes de especificidade para as Serino Proteases estão, considerando as posições dos resíduos na cadeia A da estrutura PDB 5PTP, nas posições 172, 189, 216 e 226, que correspondem às posições 8, 11, 22 e 27 do sítio ativo apresentado. O resíduo de aspartato (D) na posição 189 das Tripsinas é responsável pela preferência delas por atuar sobre resíduos polares, sendo trocado para serina (S) nas Quimotripsinas e Elastases, o que causa a propriedade hidrofóbica. Há oclusão do sítio ativo das Elastases por uma valina (V) na posição 216 e uma treonina (T) na posição 226, o que determina sua preferência para atuar sobre resíduos alifáticos pequenos; nessas posições, há glicinas (G) nas Quimotripsinas e Tripsinas. Finalmente, o resíduo de tirosina (Y) na posição 172 das Tripsinas é trocado para triptofano (W) nas Quimotripsinas e Elastases. Essa troca é descrita na literatura como determinante da conversão entre Tripsina e Quimotripsina. De fato, pode-se observar que os resíduos de triptofano (W), serina (S), valina (V) e treonina (T) que ocorrem, respectivamente, nas posições 8, 11, 22 e 27 para as Elastases, estão entre os resíduos mais importantes para diferenciar o Grupo I obtido pelo sistema de GP. Como os demais grupos apresentam mistura de subfamílias, os resíduos mais importantes para diferenciá-los dos demais não correspondem às posições determinantes de especificidade da família.

Divisão das Serino Proteases em quatro grupos

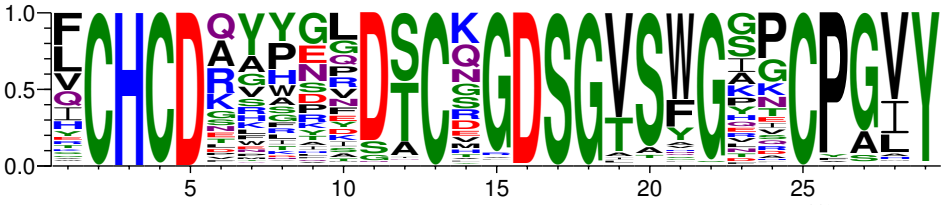
Como anteriormente mencionado, com os parâmetros originalmente utilizados por Melo-Minardi et al. (2010), o ASMC não divide o conjunto atualizado de Serino Proteases. Reduzindo o valor do parâmetro -C de 0,25 para 0,15, o ASMC produz um agrupamento hierárquico que divide a família, no primeiro nível, em quatro grupos, cujos logotipos e composições são apresentados na Figura 6.9. Pode-se observar que o ASMC também foi capaz de separar apenas a subfamília de Elastases, enquanto Quimotripsinas e Calicreínas, que representam uma porcentagem muito baixa da família, ficam misturadas em um dos três grupos de Tripsinas.



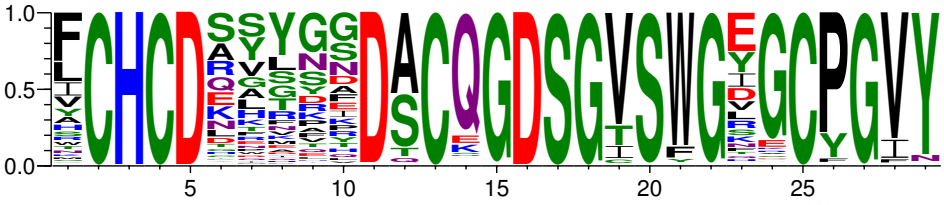
(a) Grupo I: 43 Elastases



(b) Grupo II: 221 Tripsinas



(c) Grupo III: 401 Tripsinas, 26 Quimotripsinas e 13 Calicreínas

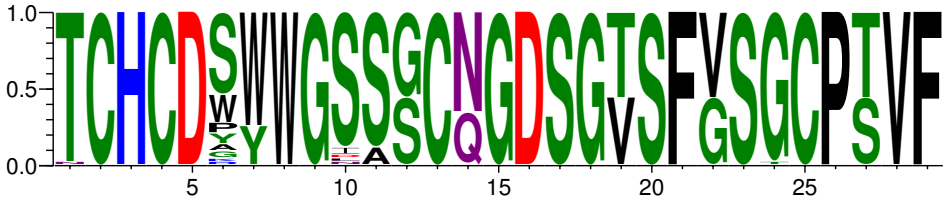


(d) Grupo IV: 829 Tripsinas

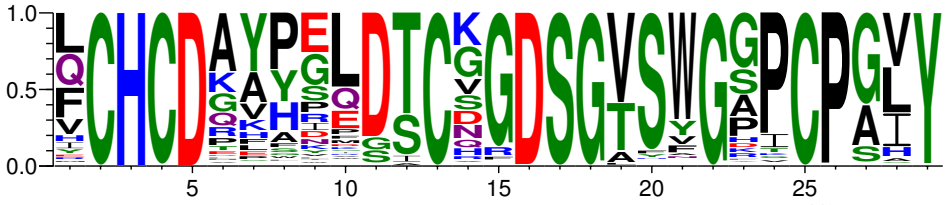
Figura 6.9: Divisão das Serino Proteases em quatro grupos no primeiro nível do agrupamento hierárquico do ASMC.

Quando o sistema de GP é executado para separar a família em quatro grupos, o melhor resultado alcançado com a configuração em questão utiliza três tipos de dados combinados com o mesmo peso: a identidade e pontuação dos sítios ativos, e a pontuação do alinhamento de sequências global, como apresentado na Tabela 6.11. Os logotipos e composições dos grupos obtidos são apresentados

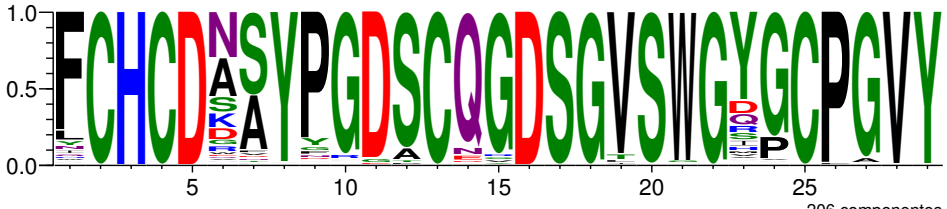
na Figura 6.10. O valor de MI desse agrupamento é de 17,71, enquanto o agrupamento produzido pelo ASMC tem MI = 16,58.



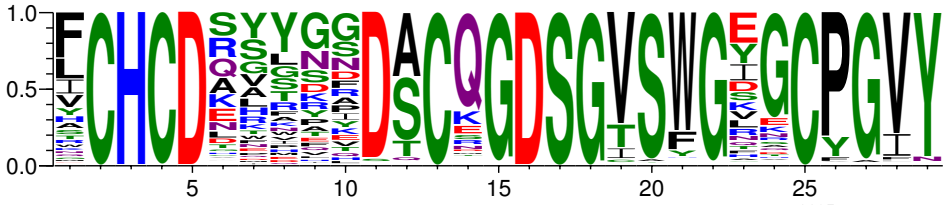
(a) Grupo I: 43 Elastases



(b) Grupo II: 136 Tripsinas e 13 Calicreínas



(c) Grupo III: 206 Tripsinas



(d) Grupo IV: 1.109 Tripsinas e 26 Quimotripsinas

Figura 6.10: Divisão das Serino Proteases em quatro grupos pelo sistema de GP.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

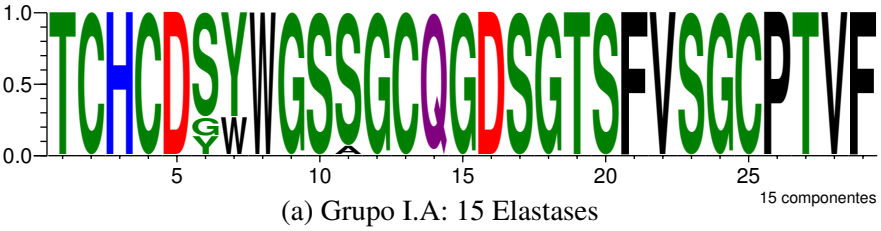
- ◇ **Grupo I:** F (29), W (8), T (1), S (11), S (23), F (21), W (7), V (22), S (10) e T (27);
- ◇ **Grupo II:** P (24), T (12), L (10), G (23), P (8), L (28), A (27), E (9), G (14) e Q (1);
- ◇ **Grupo III:** P (9), G (10), Y (8), Y (23), F (1), S (7), S (12), A (7), N (6) e Q (14); e
- ◇ **Grupo IV:** A (12), E (23), G (27), Y (26), G (24), D (11), Q (14), N (9), I (23) e G (7).

Pode-se observar que o sistema de GP foi capaz de concentrar uma maior quantidade de Tripsinas em um mesmo grupo do que o ASMC, além de inserir Quimotripsinas e Calicreínas em grupos

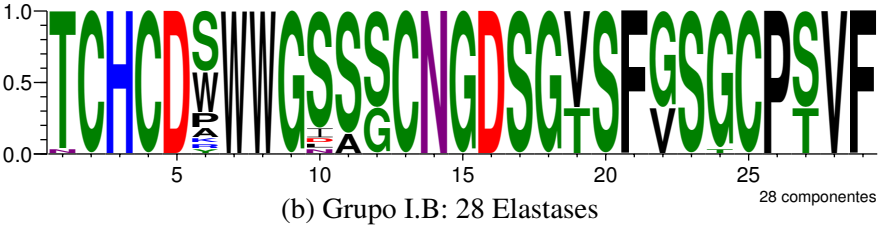
distintos. Ainda assim, essas subfamílias continuam misturadas a subgrupos das Tripsinas, que ocorrem em quantidade muito maior na família e, por isso, a dominam.

Divisão das Serino Proteases em onze grupos

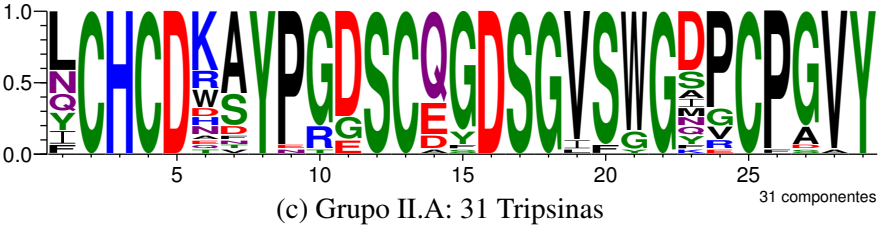
Há onze grupos no segundo nível do agrupamento hierárquico produzido pelo ASMC, cujos logotipos e composições são apresentados na Figura 6.11. Observa-se que todos os grupos do primeiro nível foram subdivididos, assim, o ASMC conseguiu isolar as Quimotripsinas em um grupo próprio e quase isolou as Calicreínas em outro, compartilhado apenas com cinco Tripsinas. No entanto, o grupo de Elastases, que já era bastante homogêneo, foi quebrado em dois. Isso indica que os grupos são subdivididos independente do nível de variabilidade de resíduos que apresentem, podendo produzir, dessa forma, subgrupos de menor relevância.



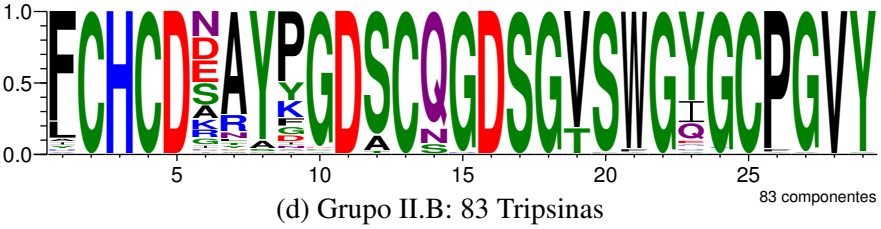
(a) Grupo I.A: 15 Elastases



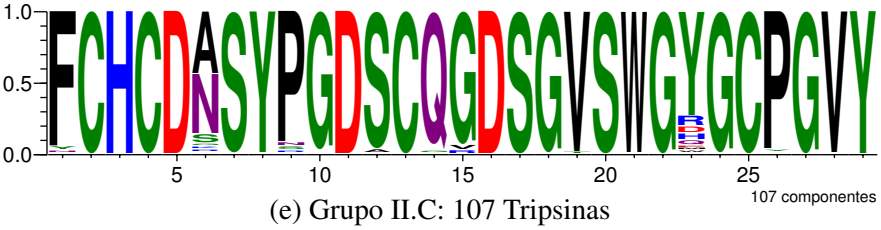
(b) Grupo I.B: 28 Elastases



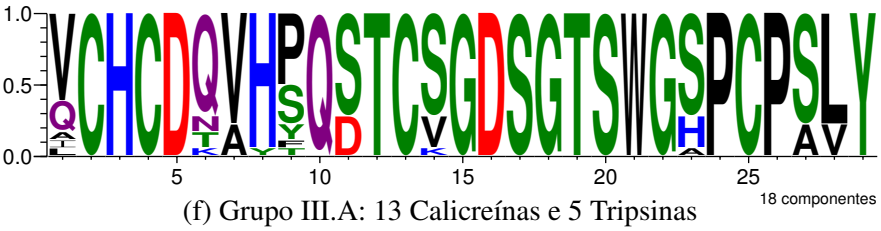
(c) Grupo II.A: 31 Tripsinas



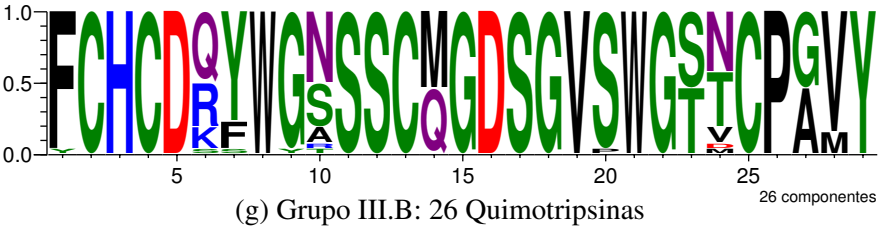
(d) Grupo II.B: 83 Tripsinas



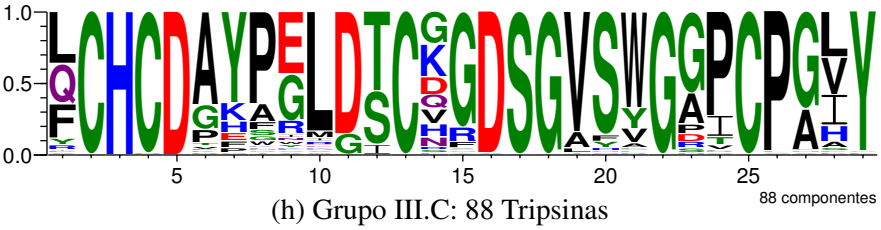
(e) Grupo II.C: 107 Tripsinas



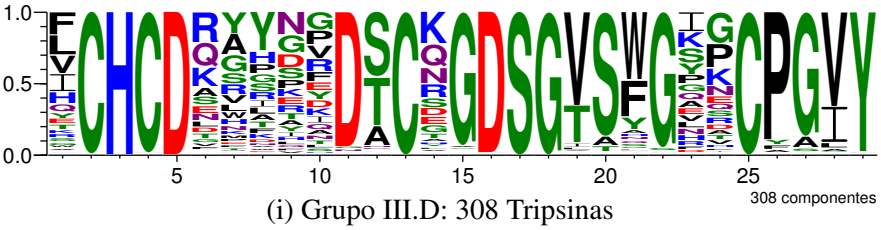
(f) Grupo III.A: 13 Calicreínas e 5 Tripsinas



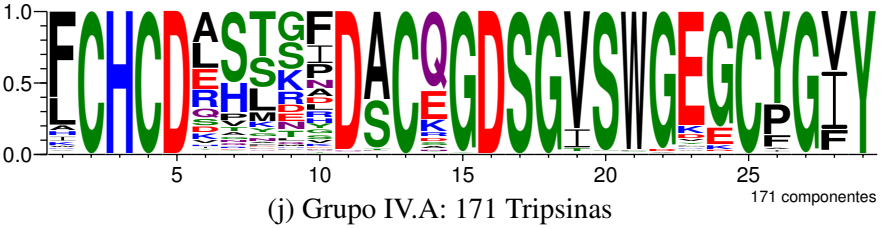
(g) Grupo III.B: 26 Quimotripsinas



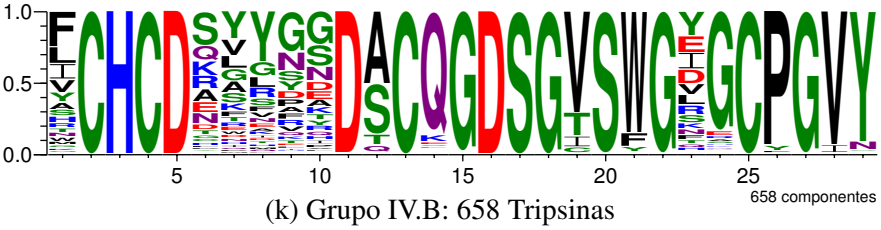
(h) Grupo III.C: 88 Tripsinas



(i) Grupo III.D: 308 Tripsinas



(j) Grupo IV.A: 171 Tripsinas

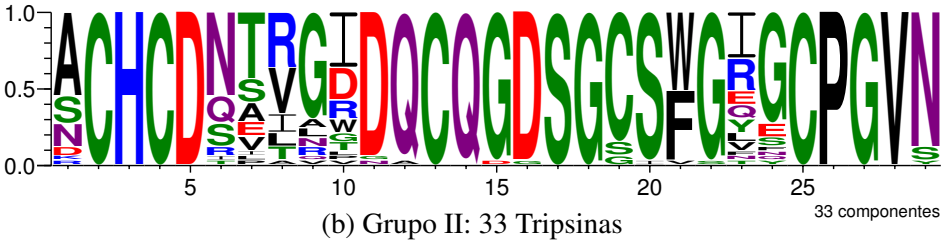


(k) Grupo IV.B: 658 Tripsinas

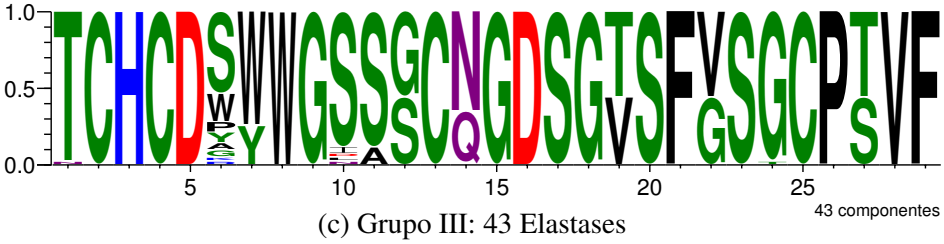
Figura 6.11: Divisão das Serino Proteases em onze grupos no segundo nível do agrupamento hierárquico do ASMC.

Executando o sistema de GP com onze grupos a fim de comparar os agrupamentos obtidos entre as duas técnicas, o melhor resultado obtido combina os mesmos três tipos de dados que o resultado com quatro grupos, apenas dobrando o peso da identidade do sítio ativo. A Figura 6.12 apresenta os logotipos e composições dos grupos gerados, que apresentam MI = 12,49, contra MI = 10,59 do resultado do ASMC.

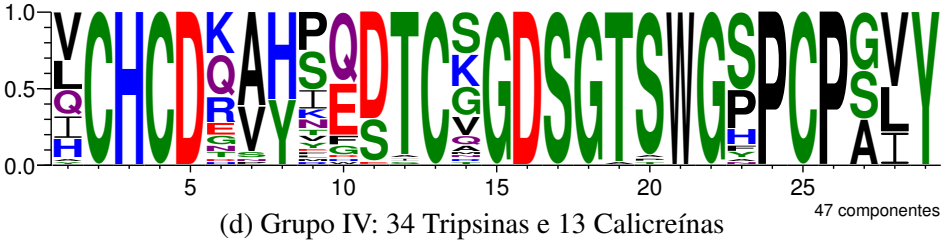
(a) Grupo I: 27 Tripsinas



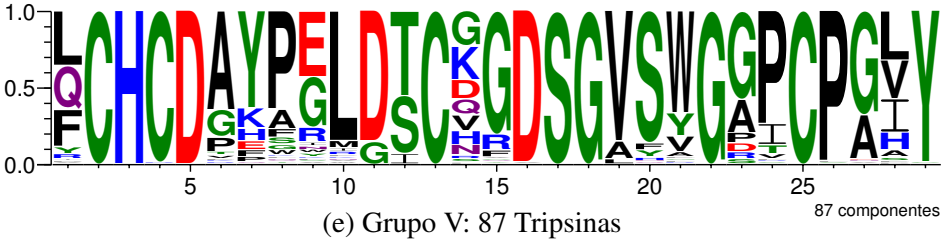
(b) Grupo II: 33 Tripsinas



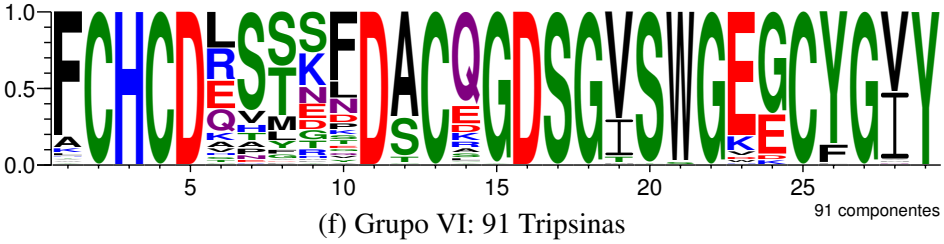
(c) Grupo III: 43 Elastases



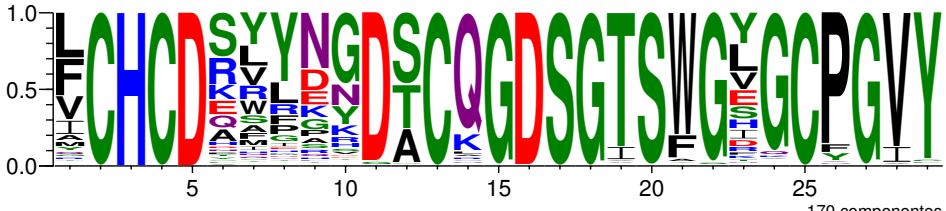
(d) Grupo IV: 34 Tripsinas e 13 Calicreínas



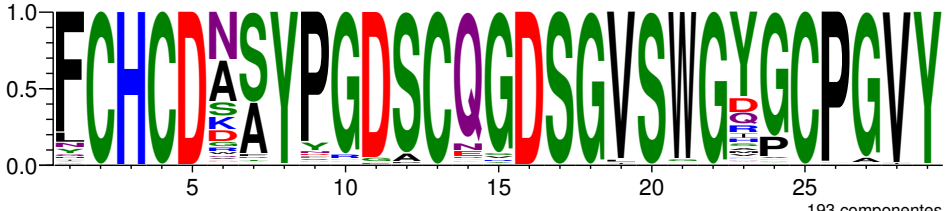
(e) Grupo V: 87 Tripsinas



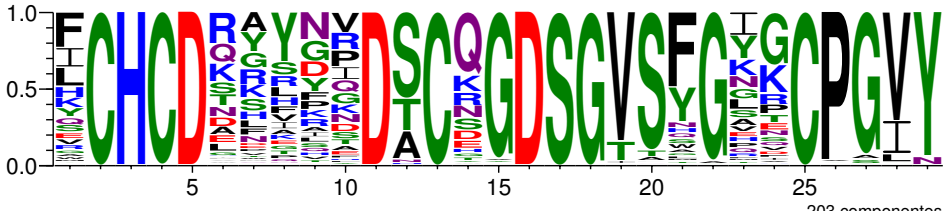
(f) Grupo VI: 91 Tripsinas



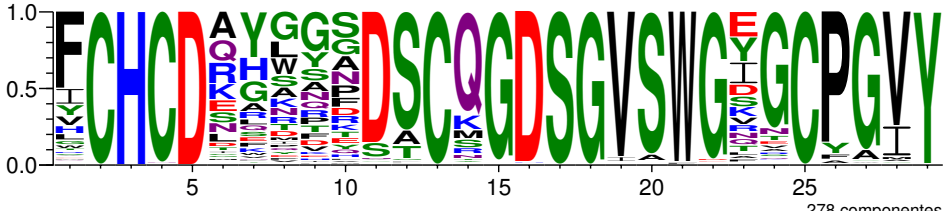
(g) Grupo VII: 170 Tripsinas



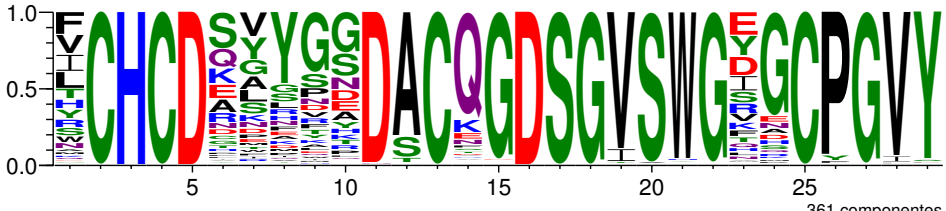
(h) Grupo VIII: 193 Tripsinas



(i) Grupo IX: 203 Tripsinas



(j) Grupo X: 252 Tripsinas e 26 Quimotripsinas



(k) Grupo XI: 361 Tripsinas

Figura 6.12: Divisão das Serino Proteases em onze grupos pelo sistema de GP.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

- ◇ **Grupo I:** F (28), E (14), T (8), I (10), Y (26), L (1), E (23), S (7), A (12) e R (9);
- ◇ **Grupo II:** Q (12), C (19), N (29), A (1), T (7), N (6), G (9), I (10), R (8) e V (8);
- ◇ **Grupo III:** F (29), W (8), T (1), S (11), F (21), W (7), V (22), S (10) e T (27);
- ◇ **Grupo IV:** P (24), T (19), T (12), H (8), Q (10), A (7), S (23), S (27), E (10) e P (23);
- ◇ **Grupo V:** L (10), P (8), P (24), G (23), A (6), E (9), Y (7), T (12), L (28) e Q (1);
- ◇ **Grupo VI:** Y (26), E (23), F (10), S (7), F (1), S (8), A (12), T (8), I (28) e E (24);
- ◇ **Grupo VII:** T (19), N (9), G (24), G (10), L (23), L (1), Q (14), S (6), T (12) e Y (10);
- ◇ **Grupo VIII:** P (9), G (10), Y (8), S (7), F (1), S (12), Y (23), N (6), A (7) e Q (14);
- ◇ **Grupo IX:** F (21), Y (21), K (24), V (10), P (26), V (19), S (12), N (9), R (6) e I (28);
- ◇ **Grupo X:** S (12), V (19), W (21), G (8), F (1), H (7), Y (7), G (24), G (7) e M (14); e
- ◇ **Grupo XI:** A (12), W (21), G (9), V (19), V (28), Y (8), V (7), S (6), G (27) e Q (14).

Observa-se na Figura 6.12 que os onze grupos gerados pelo sistema de GP não isolam as subfamílias menores. No entanto, analisando os logotipos pode-se notar que os grupos são mais expressivos que aqueles produzidos pelo ASMC, por terem maiores diferenças entre si, enquanto o ASMC subdivide grupos que já eram bastante homogêneos em relação aos demais.

De fato, ao analisar as entradas no UniProt das proteínas em cada grupo, observações interessantes podem ser feitas. No Grupo I, por exemplo, há 27 proteínas rotuladas como Tripsinas. Sete são entradas manualmente revisadas no UniProt, seis das quais são anotadas como Protrombinas e uma como Fator de Coagulação VII, cujo nome alternativo é *serum prothrombin conversion accelerator*. Entre as vinte cujas entradas no UniProt não foram manualmente revisadas, dez foram nomeadas Protrombinas, sete, Trombinas e uma, Fator de Coagulação VII. As duas últimas, com identificadores UniProt Q4SUA7 e Q6GNK4, não têm nomes relacionados aos anteriores, mas apresentam os domínios do InterPro IPR003966 (*Prothrombin/thrombin*) e IPR018992 (*Thrombin light chain*).

Já entre as 34 proteínas rotuladas como Tripsinas inseridas no Grupo IV junto às treze Calicreínas, vinte foram manualmente revisadas no UniProt, todas anotadas como Calicreínas. Considerando as quatorze entradas que não foram manualmente revisadas, treze têm nomes sugeridos de Calicreína como *Prostatic kallikrein 2*, *Glandular kallikrein* e *Kallikrein*. A última, de identificador UniProt Q8C232, não tem nenhuma anotação relacionada a Calicreínas, mas corresponde ao gene Klk1b1,

o mesmo gene codificador de várias Calicreínas manualmente anotadas. Isso mostra que os rótulos de subfamílias usados por Melo-Minardi et al. (2010) estão incoerentes e o sistema de GP foi capaz de agrupar corretamente as Calicreínas em um mesmo grupo, o que não ocorreu no agrupamento produzido pelo ASMC.

Divisão das Serino Proteases em doze grupos

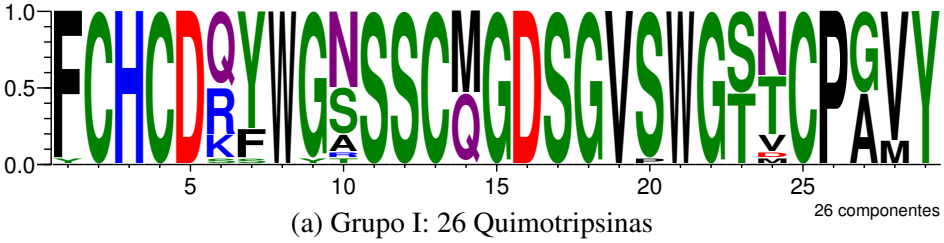
Trabalhando com onze grupos, o sistema de GP não foi capaz de criar um grupo de Quimotripsinas. No entanto, quando executado para doze grupos, o sistema de GP mantém as Elastases em um único grupo, mantém os grupos interessantes de Calicreínas e Protrombinas mencionados anteriormente e consegue isolar as Quimotripsinas em um grupo próprio, como mostra a Figura 6.13.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

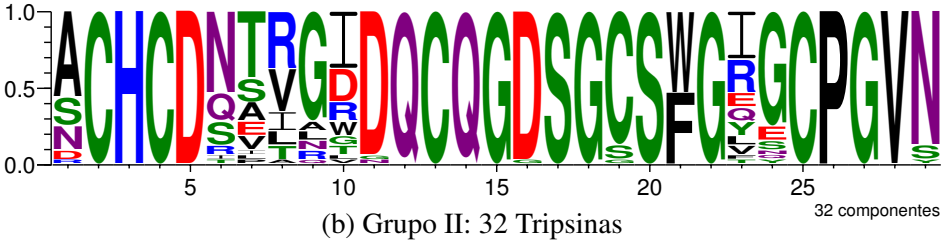
- ◇ **Grupo I:** S (11), W (8), M (14), G (9), T (23), Y (7), N (24), F (1), T (24) e A (27);
- ◇ **Grupo II:** Q (12), C (19), N (29), A (1), T (7), N (6), G (9), I (10), R (8) e V (8);
- ◇ **Grupo III:** F (29), W (8), T (1), S (11),S (23), F (21), W (7), V (22), S (10) e T (27);
- ◇ **Grupo IV:** F (28), E (14), E (23), Y (26), T (8), I (10), L (1), A (12), S (7) e G (24);
- ◇ **Grupo V:** P (24), T (12), T (19), H (8), Q (10), A (7), S (23), S (27), A (27) e E (10);
- ◇ **Grupo VI:** Y (26), E (23), F (10), S (7), F (1), S (8), I (28), T (8), E (24) e L (6);
- ◇ **Grupo VII:** L (10), P (8), P (24), G (23), A (6), E (9), Y (7), T (12), L (28) e Q (1);
- ◇ **Grupo VIII:** I (28), L (8), H (7), R (6), F (1), T (19), N (9), T (12), S (12) e L (23);
- ◇ **Grupo IX:** Y (8), V (7), S (6), G (9), I (19), T (19), V (28), W (21), Q (14) e L (1);
- ◇ **Grupo X:** P (9), G (10), Y (8), F (1), Y (23), S (7), S (12), N (6), A (7) e Q (14);
- ◇ **Grupo XI:** S (12), F (21), V (19), V (28), G (7), I (23), G (8), A (9), K (24) e Q (6); e
- ◇ **Grupo XII:** A (12), V (28), V (19), W (21), G (24), I (1), G (27), K (6), V (1) e Y (10).

Devido ao imenso desbalanceamento entre a quantidade de proteínas em cada subfamília, as Serino Proteases mostraram-se a família mais difícil de separar em subfamílias, tanto para o sistema de GP quanto para o ASMC. Isso deve-se à grande variabilidade de resíduos existente entre as Tripsinas, que representam 94,7% da família, o que leva os algoritmos a encontrarem subgrupos de Tripsinas mais facilmente do que as outras subfamílias, uma vez que são mais volumosos que elas. Por essa

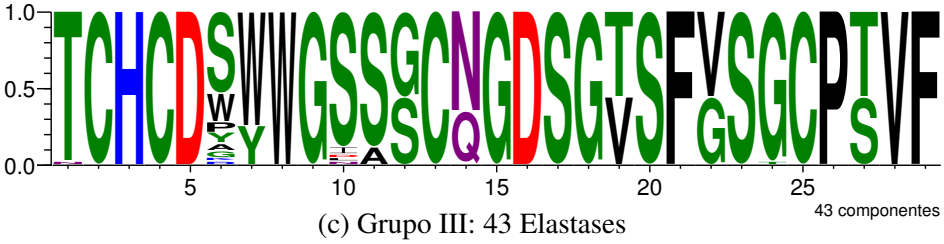
razão, ambas as técnicas somente conseguiram grupos próprios para Calicreínas e Quimotripsinas após quebrarem as Tripsinas em vários subgrupos. Ainda assim, os subgrupos obtidos pelo sistema de GP para as Tripsinas mostraram-se relevantes, tendo sido encontrados grupos de Protrombinas, por exemplo, e um grupo de Calicreínas maior do que o observado por Melo-Minardi et al. (2010).



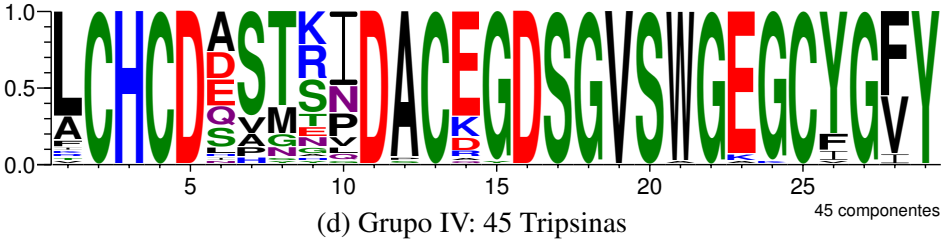
(a) Grupo I: 26 Quimotripsinas



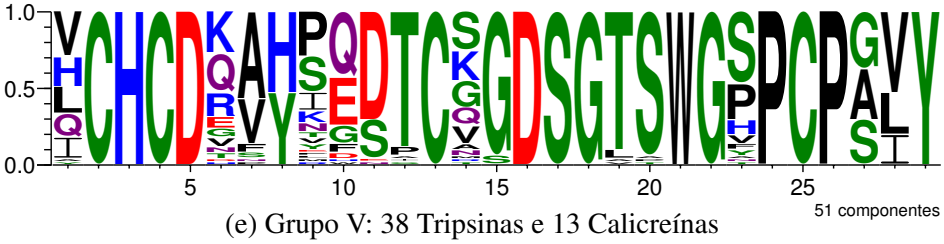
(b) Grupo II: 32 Tripsinas



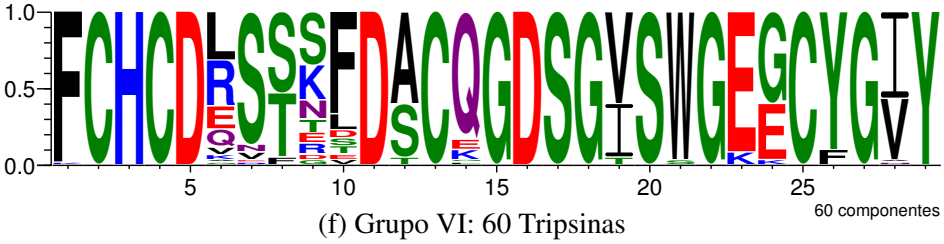
(c) Grupo III: 43 Elastases



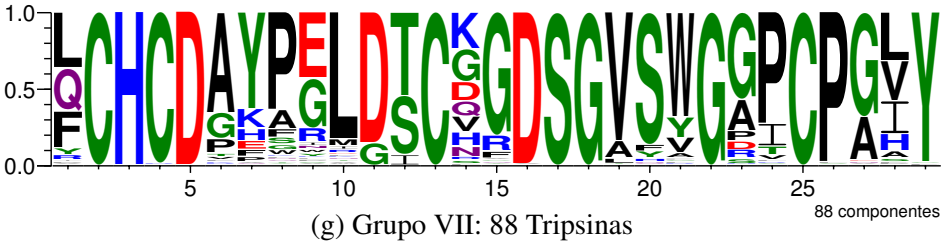
(d) Grupo IV: 45 Tripsinas



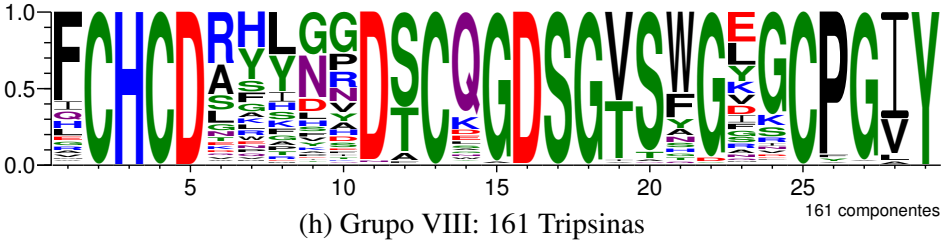
(e) Grupo V: 38 Tripsinas e 13 Calicreínas



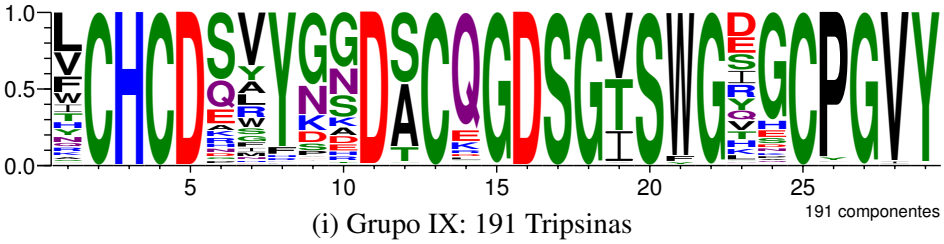
(f) Grupo VI: 60 Tripsinas



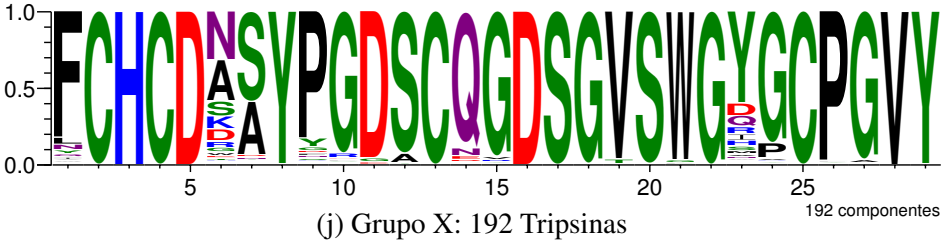
(g) Grupo VII: 88 Tripsinas



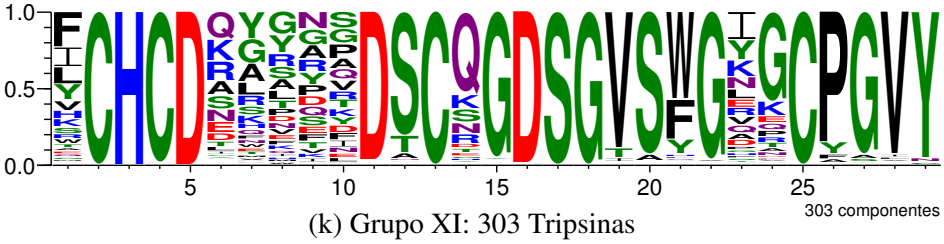
(h) Grupo VIII: 161 Tripsinas



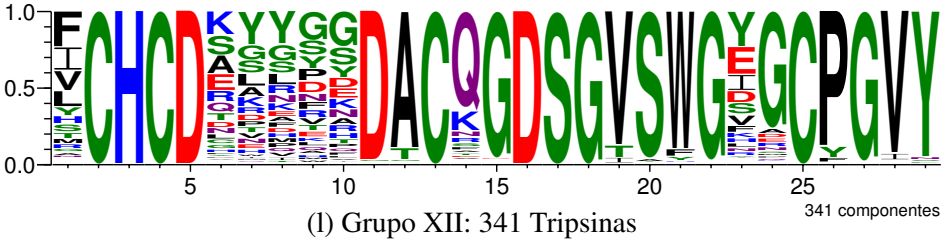
(i) Grupo IX: 191 Tripsinas



(j) Grupo X: 192 Tripsinas



(k) Grupo XI: 303 Tripsinas



(l) Grupo XII: 341 Tripsinas

Figura 6.13: Divisão das Serino Proteases em doze grupos pelo sistema de GP.

6.5 Estudo de Caso III: Proteínas Cinases

Após remover, do conjunto de proteínas usado por Melo-Minardi et al. (2010), 314 sequências que tornaram-se obsoletas no UniProt desde então, essa família passou a ter 3.087 proteínas, sendo 2.044 rotuladas como Serina/Treonina Cinases e 1.043, como Tirosina Cinases. Os autores relataram também um subgrupo de 235 Tirosina Cinases rotuladas como EGFRs (do inglês *Epidermal Growth Factor Receptor*). Para a família atualizada, o ASMC produziu, com os mesmos parâmetros usados pelos autores para o conjunto de proteínas original, um agrupamento hierárquico no qual o primeiro nível dividiu essa família em três grupos e o segundo nível, em sete. Por essas razões, o sistema de GP foi executado com dois a sete grupos. Novamente ilustrando a dificuldade do problema, existem mais de quatro milhões de formas de combinar as 3.087 proteínas em dois grupos, mais de quatro bilhões de formas de combiná-las em três grupos e $5{,}2 \times 10^{20}$ formas de combiná-las em sete grupos. Os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados nas Tabelas A.17 a A.22 do Apêndice A. A Figura 6.14 apresenta o logotipo que ilustra a composição do sítio ativo putativo das 3.087 Proteínas Cinases, no qual pode-se se observar uma grande variabilidade de resíduos.

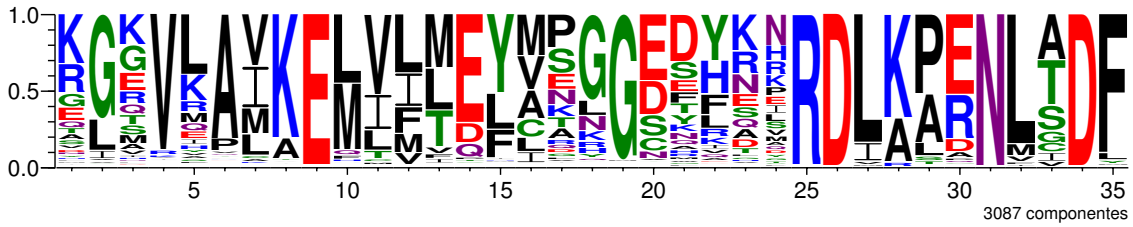


Figura 6.14: Composição do sítio ativo das Proteínas Cinases.

A Tabela 6.12 apresenta as combinações de dados produzidas pelo sistema de GP que levaram aos melhores resultados para as Proteínas Cinases, empregando o grafo de similaridades construído apenas com valores positivos das matrizes de similaridades calculadas por essas equações, e aplicando as taxas de 80% de cruzamento, 5% de reprodução e 15% de mutação, como nos outros estudos de caso. Novamente, observa-se a presença dos dados relacionados ao sítio ativo na maioria das equações, o que é esperado devido à métrica de qualidade empregada. Outros tipos de dados que destacaram-se foram a identidade do alinhamento estrutural ($strAliId$) e as similaridades de anotações do InterPro ($interpro$) e de termos do GO (go).

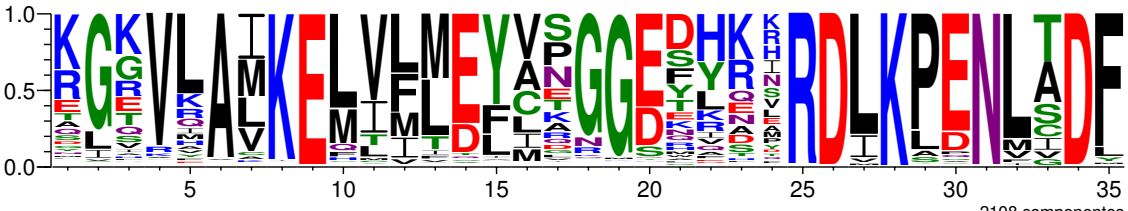
Divisão das Proteínas Cinases em dois grupos

Quando o sistema de GP é executado para dividir essa família em dois grupos, visando à identificação das duas principais subfamílias (Serina/Treonina Cinases e Tirosina Cinases), o melhor resultado obtido com a configuração de parâmetros em questão envolve três tipos de dados: identidade do sítio ativo, similaridade de termos do GO e similaridade de anotações do InterPro, como mostrado na Ta-

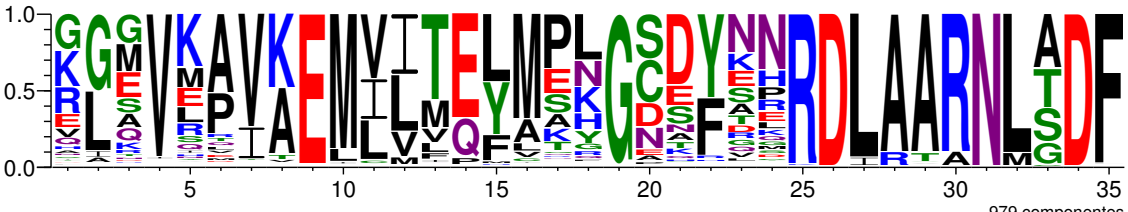
Tabela 6.12: Combinações de dados que levaram à obtenção dos melhores agrupamentos para a família de Proteínas Cinases para o grafo de similaridades construído apenas com valores positivos e com 80% de cruzamento, 5% de reprodução e 15% de mutação.

Grupos	Repetição	Equação
2	1	$ASid + go + interpro$
3	2 e 4	$ASid + strAliId$
4	4	$interpro + 2strAliId$
5	2	$ASid + ASscr$
6	5	$5ASid + 2ASscr + 2aaCompDist + cooccurence + 3go + 3interpro + strAliId + 3strAliScr + strAliSize$
7	3	$coexpression + 2go + 2strAliId + strAliSize$

bela 6.12. A Figura 6.15 apresenta os logotipos e as composições dos grupos produzidos pelo sistema de GP. Pode-se observar que os grupos apresentam claramente as sequências-consenso RDLKPEN para Serina/Treonina Cinases e RDLAARN típica de Tirosina Cinases, como descrito na Seção 5.1.



(a) Grupo I: 2.044 Serina/Treonina Cinases e 64 Tirosina Cinases



(b) Grupo II: 744 Tirosina Cinases e 235 EGFRs

Figura 6.15: Divisão das Proteínas Cinases em dois grupos pelo sistema de GP.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição. Os resíduos em negrito correspondem àqueles que conhecidamente alteram a especificidade de substrato da família, como descrito a seguir.

- **Grupo I: K (28)**, **E (30)**, **P (29)**, G (18), E (20), L (10), K (8), L (5), A (6) e H (22); e
- **Grupo II: A (28)**, **R (30)**, **A (29)**, M (16), M (10), V (7), A (8), T (13), F (22) e L (18).

Segundo Melo-Minardi et al. (2010) e Hannenhalli e Russell (2000), sabe-se que os resíduos envolvidos na diferenciação da especificidade de substrato das Proteínas Cinases são a troca das

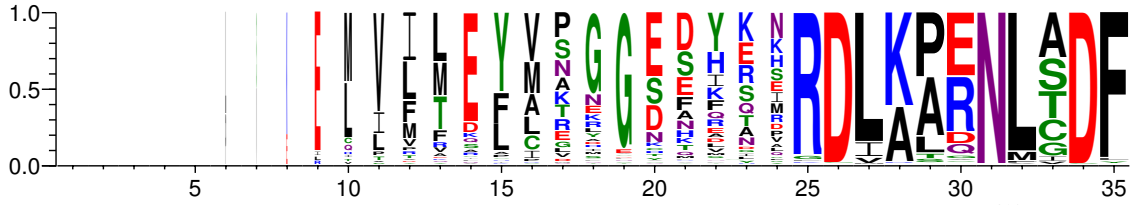
alaninas (A) das posições 254 e 255 e da arginina (R) da posição 256 nas Tirosina Cinases para, respectivamente, resíduos de lisina (K), prolina (P) e glutamato (E) nas Serina/Treonina Cinases. Essas são as posições que esses resíduos ocorrem na cadeia A da estrutura PDB 1U46, usada por Melo-Minardi et al. (2010) para modelar as Tirosina Cinases, e correspondem às posições 28, 29 e 30, respectivamente, do sítio ativo apresentado. Pode-se observar que esses são os resíduos considerados pelo sistema de GP os mais importantes para diferenciar cada subfamília, devido aos valores mais altos de MI: K (28), P (29) e E (30) para o Grupo I, de Serina/Treonina Cinases, e A (28), A (29) e R (30) para o Grupo II, de Tirosina Cinases. Isso mostra que o sistema de GP foi capaz de gerar grupos cujas posições que mais os diferenciam correspondem àquelas que conhecidamente definem a especificidade dessas subfamílias.

Analisando as entradas no UniProt das 64 proteínas rotuladas como Tirosina Cinases que foram inseridas no grupo das Serina/Treonina Cinases, observa-se que nenhuma foi manualmente curada. Entre elas, 57 apresentam o domínio InterPro IPR008271, que corresponde ao sítio ativo de Serina/Treonina Cinases. Entre as outras sete, três (identificadores Q5RAR7, A0JN96 e Q4T0K5) apresentam o domínio InterPro IPR002290 (*Serine/threonine/dual specificity protein kinase, catalytic domain*), encontrado em Serina/Treonina Cinases ou em Proteínas Cinases com dupla especificidade, enquanto três (A0JN96, Q4T0K5 e Q4R8A9) apresentam o termo GO *protein serine/threonine kinase activity*. As últimas três, de identificadores UniProt A1INL8, Q6K3D4 e Q56WL1, não apresentam qualquer anotação específica de uma ou de outra subfamília, de modo que não é possível afirmar se o que está certo são os rótulos utilizados por Melo-Minardi et al. (2010) ou a inserção dessas proteínas no grupo de Serina/Treonina Cinases feita pelo sistema de GP. A composição desse conjunto de 64 proteínas parece sugerir que os rótulos de subfamílias estão incorretos e que o sistema de GP as agrupou corretamente como Serina/Treonina Cinases, ou que essas proteínas podem ser duais, desempenhando ambas as funções.

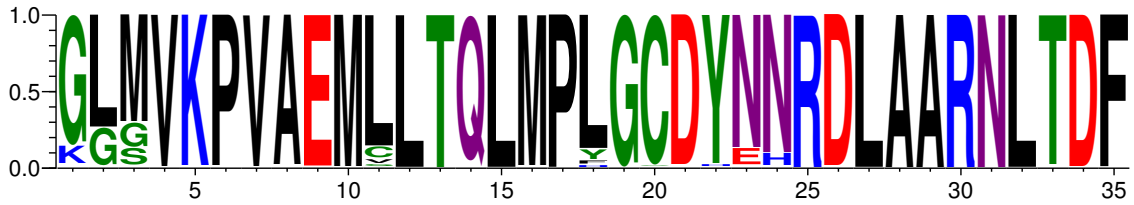
Divisão das Proteínas Cinases em três grupos

No primeiro nível do agrupamento hierárquico gerado pelo ASMC, a família de Proteínas Cinases foi dividida em três grupos, cujos logotipos e composições são apresentados na Figura 6.16. Essa família apresenta algumas proteínas cujos modelos estruturais não alinharam bem com o sítio ativo da estrutura de referência, por isso ocorrem muitas lacunas nas posições que não estão envolvidas nas sequências-consenso RDLKPEN e RDLAARN. Pode-se observar que o ASMC gerou um grupo contendo as proteínas cujos sítios ativos contêm muitas lacunas. No entanto, tais lacunas não são indicativas de função. Adicionalmente, foi gerado um grupo contendo as EGFRs, mas mais de 87% da família foi inserida no Grupo III, que mistura as duas principais subfamílias.

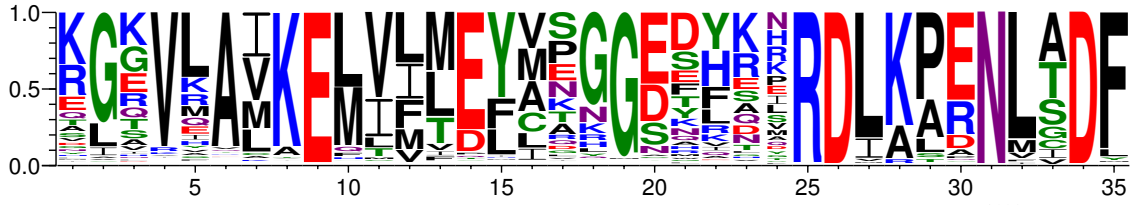
Novamente, o ASMC não mostrou-se muito estável em relação aos grupos gerados, uma vez que a remoção de apenas 9,2% das proteínas levou a um agrupamento completamente diferente da-



(a) Grupo I: 94 Serina/Treonina Cinases, 67 Tirosina Cinases e 2 EGFRs



(b) Grupo II: 233 EGFRs e 3 Tirosina Cinases



(c) Grupo III: 1.950 Serina/Treonina Cinases e 738 Tirosina Cinases

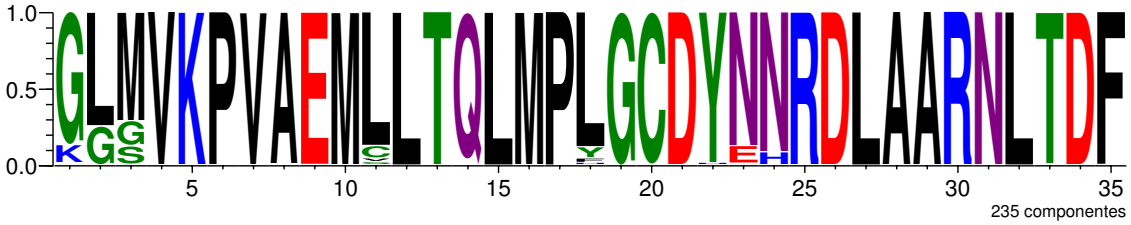
Figura 6.16: Divisão das Proteínas Cinases em três grupos no primeiro nível do agrupamento hierárquico do ASMC.

quele produzido para o conjunto original de proteínas utilizado por Melo-Minardi et al. (2010), cujos logotipos e composições são apresentados na Figura B.4 do Apêndice B. O agrupamento produzido para a família original tem grande correspondência com os rótulos de subfamílias e com aquele produzido pelo sistema de GP.

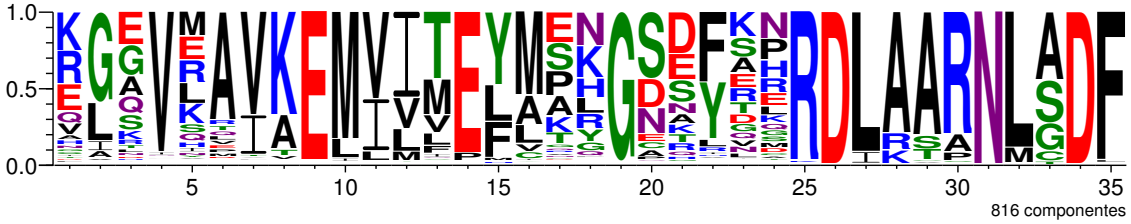
Ao executar o sistema de GP para dividir essa família em três grupos, os melhores resultados, para a configuração de parâmetros em questão, são obtidos combinando dois tipos de dados: as identidades dos sítios ativos e do alinhamento estrutural, como mostrado na Tabela 6.12. Os logotipos e as composições dos grupos gerados são apresentados na Figura 6.17, na qual observa-se que todas as proteínas rotuladas como EGFRs foram inseridas em um mesmo grupo homogêneo, e que os demais grupos respeitam quase totalmente os rótulos de classe, assim como as sequências-consenso típicas de cada subfamília.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

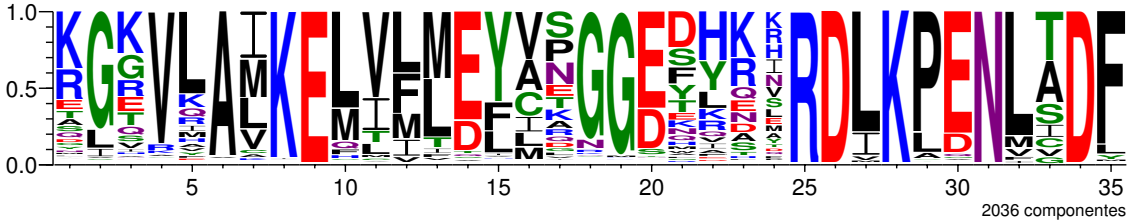
- **Grupo I:** P (6), C (20), Q (14), L (18), G (1), A (8), M (3), N (23), K (5) e N (24);
- **Grupo II:** R (30), A (28), A (29), F (22), M (10), I (12), S (20), V (7), M (16) e T (13); e
- **Grupo III:** K (28), E (30), P (29), G (18), E (20), L (10), L (5), K (8), A (6) e H (22).



(a) Grupo I: 235 EGFRs



(b) Grupo II: 805 Tirosina Cinases e 11 Serina/Treonina Cinases (Tabela 6.13)



(c) Grupo III: 2.033 Serina/Treonina Cinases e 3 Tirosina Cinases (Tabela 6.14)

Figura 6.17: Divisão das Proteínas Cinases em três grupos pelo sistema de GP.

Quando o sistema de GP buscou por dois grupos, um conjunto de 64 Tirosina Cinases foi inserido junto às Serina/Treonina Cinases. Já considerando três grupos, esse conjunto foi agrupado com as Tirosina Cinases, conforme seus rótulos de subfamília. Isso reforça a ideia de que essas proteínas podem ser duais, ou seja, podem desempenhar ambas as funções. Além disso, mostra a capacidade, já citada anteriormente, do agrupamento de particionamento utilizado pelo sistema de GP reparar possíveis erros anteriores, algo que não é possível com o agrupamento hierárquico empregado pelo ASMC.

As três proteínas rotuladas como Tirosina Cinases inseridas no Grupo III, de maioria Serina/Treonina Cinases, têm identificadores UniProt A2DGV6, Q6K3D4 e O42291. Nenhuma delas é uma entrada manualmente curada. A proteína A2DGV6 foi anotada com o termo GO *protein serine/threonine kinase activity*, o que sugere que seu rótulo esteja equivocado e que o sistema de GP a tenha agrupado corretamente entre as Serina/Treonina Cinases. A proteína Q6K3D4 não tem qualquer anotação que remeta a nenhuma das duas subfamílias, então não se pode afirmar se o rótulo esteja ou não correto. Já a proteína O42291 têm muitas anotações relacionadas às Tirosina Cinases, incluindo o domínio InterPro IPR008266, que corresponde ao sítio ativo das Tirosina Cinases, e o termo GO *protein tyrosine kinase activity*, que, combinados, são um forte indicador de que o sistema de GP incluiu essa proteína erroneamente junto às Serina/Treonina Cinases.

Tabela 6.13: Exceções no Grupo II obtido pelo sistema de GP para a divisão das Proteínas Cinases em três grupos.

Proteína	**Anotações**
A7J1T0	Revisada. Apresenta anotação de Serina/Treonina Cinase.
A7J1T1	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
A7J1T3	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
A7SFG8	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q0DE32	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases e com o termo GO *serine/threonine kinase activity*.
Q4RX00	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q4S5N1	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q54TM7	Revisada. Apresenta anotação de Serina/Treonina Cinase.
Q5RCD1	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases e com o termo GO *serine/threonine kinase activity*.
Q5SMJ0	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q6NYW1	Não-revisada. Anotada com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases e com o termo GO *serine/threonine kinase activity*.

Tabela 6.14: Exceções no Grupo III obtido pelo sistema de GP para a divisão das Proteínas Cinases em três grupos.

Proteína	**Anotações**
A2DGV6	Não-revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity*.
O42291	Não-revisada. Apresenta anotação com termo GO *protein tyrosine kinase activity* e com o domínio InterPro IPR008266, correspondente ao sítio ativo das Tirosina Cinases.
Q6K3D4	Não-revisada. Não apresenta anotação.

Considerando as onze proteínas rotuladas como Serina/Treonina Cinases inseridas no Grupo II, em que predominam as Tirosina Cinases, duas são entradas manualmente curadas no UniProt (identificadores Q54TM7 e A7J1T0) e anotadas como Serina/Treonina Cinases, indicando que o sistema de GP as incluiu erroneamente entre as Tirosina Cinases. Entre as nove que não foram manualmente revisadas todas foram automaticamente anotadas com o domínio InterPro IPR008271, que corresponde ao sítio ativo das Serina/Treonina Cinases. Adicionalmente, três (Q0DE32, Q5RCD1 e Q6NYW1) foram anotadas com o termo GO *serine/threonine kinase activity*. Essas anotações indicam que o sistema de GP incluiu erroneamente essas onze proteínas entre as Tirosina Cinases.

Apesar de alguns erros em 0,45% das proteínas da família, o sistema de GP conseguiu criar grupos relevantes e que apresentam correspondência quase total com as subfamílias existentes entre as Proteínas Cinases, diferente do agrupamento produzido pelo ASMC, que acabou por priorizar um grupo de proteínas contendo muitas lacunas.

Divisão das Proteínas Cinases em sete grupos

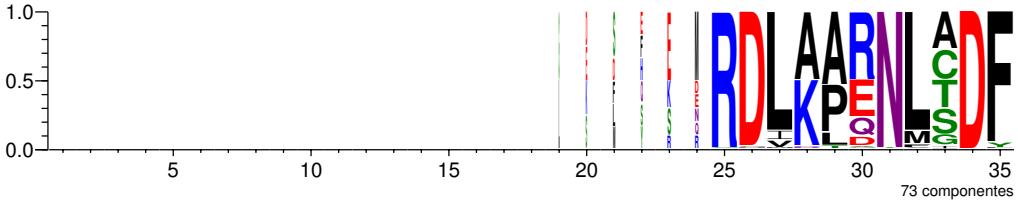
Uma vez que o ASMC não foi capaz de separar Serina/Treonina Cinases e Tirosina Cinases, é necessário considerar o segundo nível do seu agrupamento hierárquico. Os grupos produzidos são apresentados na Figura 6.18, na qual pode-se observar que o Grupo II, das EGFRs, é mantido, mas os demais são subdivididos, gerando assim sete grupos.

Pode-se observar que, mesmo aumentando a quantidade de grupos, as subfamílias Serina/Treonina Cinases e Tirosina Cinases continuaram misturadas, além de terem sido gerados grupos cujas diferenças são biologicamente irrelevantes, como os Grupos I.A e I.B. Apenas as EGFRs foram, em maioria, isoladas das demais subfamílias, o que já havia sido feito no nível anterior da hierarquia. A utilidade do segundo nível parece ter limitado-se apenas à obtenção de alguns subgrupos relativamente homogêneos de Tirosina Cinases.

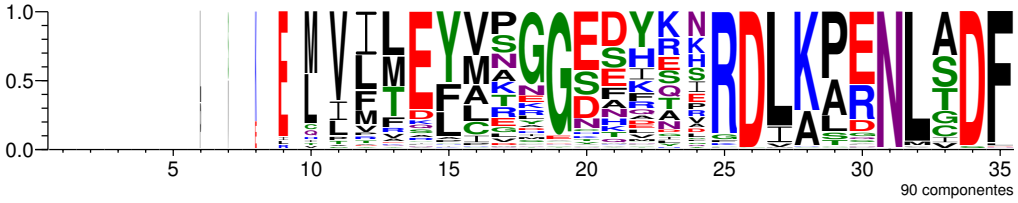
Apesar de o sistema de GP já ter separado as subfamílias com sucesso para dois e três grupos, foi executado para sete grupos a fim de comparação com o segundo nível do agrupamento hierárquico produzido pelo ASMC. Inevitavelmente, isso provocou a quebra das subfamílias em subgrupos. Como apresentado na Tabela 6.12, o melhor resultado obtido para a configuração de parâmetros em questão envolve quatro tipos de dados: a similaridade de termos do GO, a identidade e o tamanho do alinhamento estrutural, e o nível de coexpressão. Os logotipos e composições dos grupos obtidos são apresentados na Figura 6.19.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

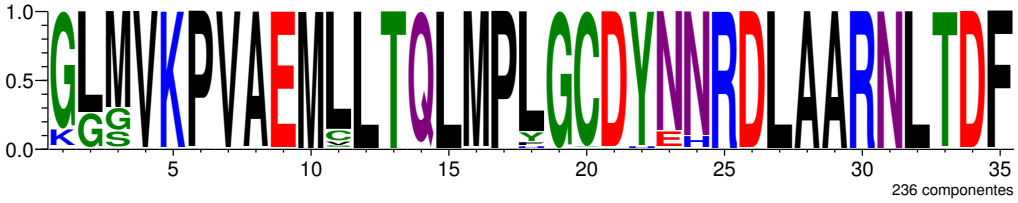
- ◇ **Grupo I:** V (13), N (20), V (12), K (18), E (21), P (24), A (16), I (11), A (28) e R (30);
- ◇ **Grupo II:** R (28), A (30), I (12), S (20), M (16), F (22), T (13), A (33), A (29) e M (10);
- ◇ **Grupo III:** L (29), D (30), F (12), D (20), H (22), N (17), F (10), T (11), F (21) e Y (35);
- ◇ **Grupo IV:** C (20), P (6), Q (14), G (1), L (18), A (8), M (3), N (23), K (5) e N (24);
- ◇ **Grupo V:** R (30), A (28), A (29), V (7), F (22), I (12), S (20), M (10), S (33) e M (16);
- ◇ **Grupo VI:** E (30), K (28), E (20), P (29), D (21), E (14), A (16), I (7), K (22) e G (2); e
- ◇ **Grupo VII:** P (29), L (10), G (18), E (30), K (28), D (14), H (22), T (33), M (7) e E (20).



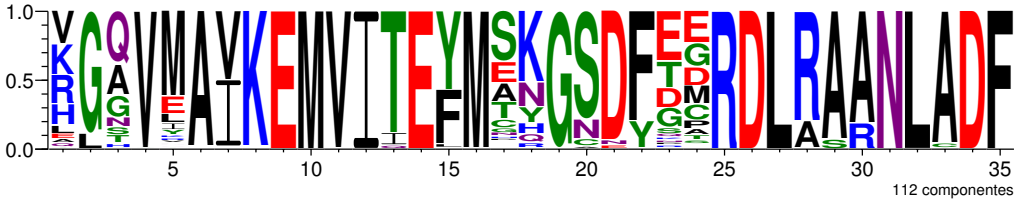
(a) Grupo I.A: 38 Tirosina Cinases e 35 Serina/Treonina Cinases



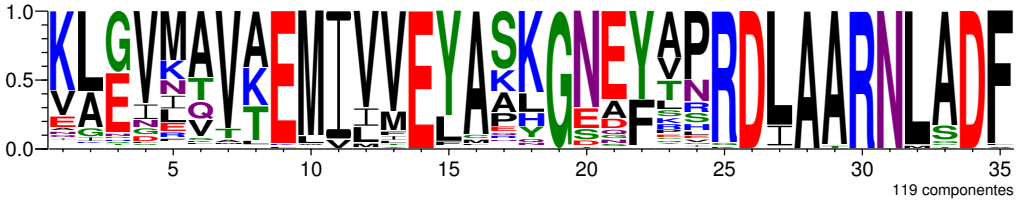
(b) Grupo I.B: 59 Serina/Treonina Cinases, 29 Tirosina Cinases e 2 EGFRs



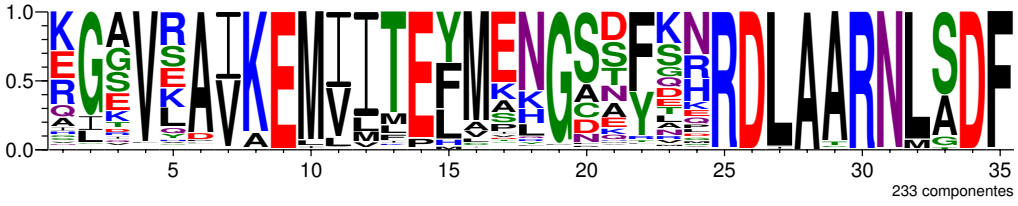
(c) Grupo II: 233 EGFRs e 3 Tirosina Cinases



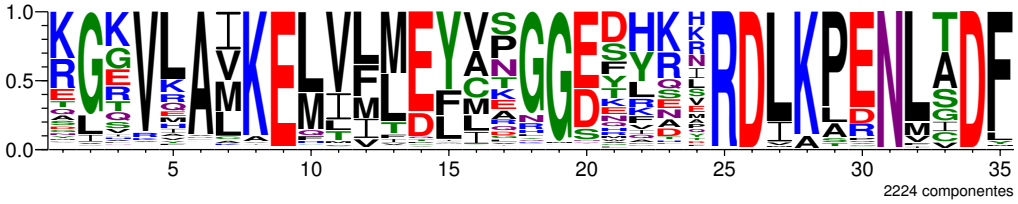
(d) Grupo III.A: 112 Tirosina Cinases



(e) Grupo III.B: 119 Tirosina Cinases

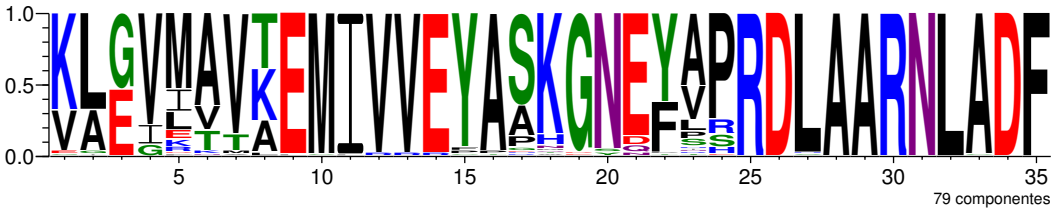


(f) Grupo III.C: 233 Tirosina Cinases

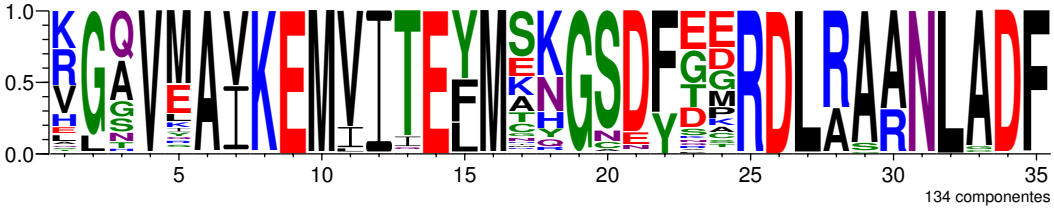


(g) Grupo III.D: 1.950 Serina/Treonina Cinases e 274 Tirosina Cinases

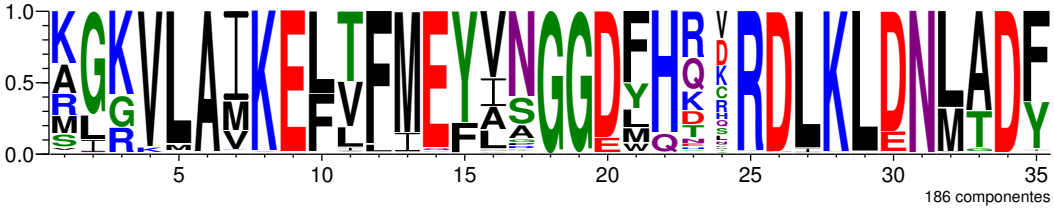
Figura 6.18: Divisão das Proteínas Cinases em sete grupos no segundo nível do agrupamento hierárquico do ASMC.



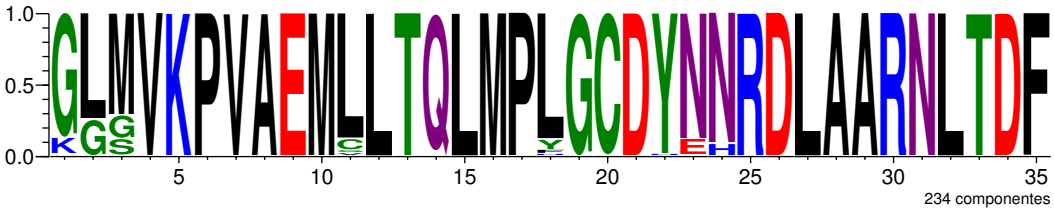
(a) Grupo I: 79 Tirosina Cinases



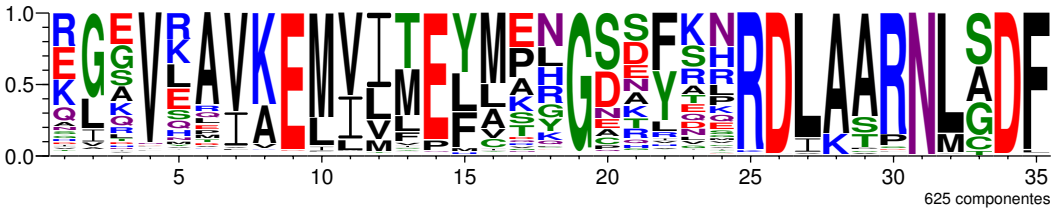
(b) Grupo II: 134 Tirosina Cinases



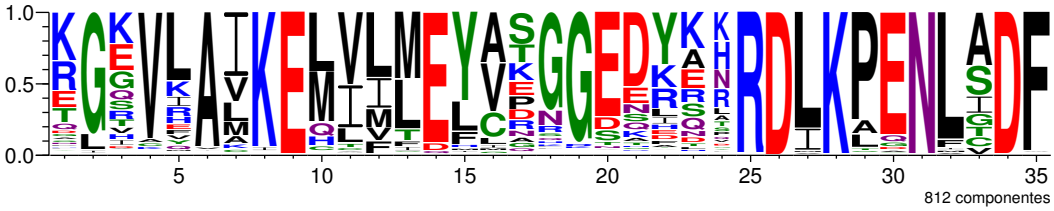
(c) Grupo III: 186 Serina/Treonina Cinases



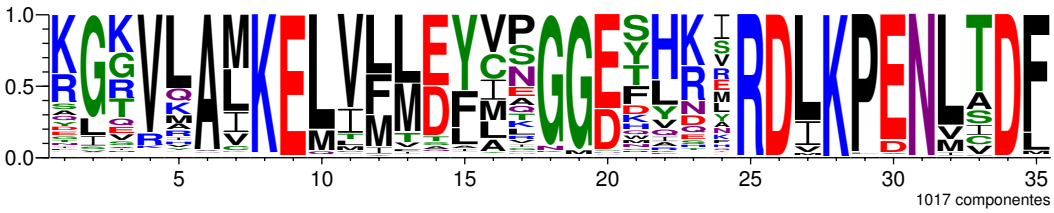
(d) Grupo IV: 234 EGFRs



(e) Grupo V: 589 Tirosina Cinases, 35 Serina/Treonina Cinases e 1 EGFR (Tabela 6.15)



(f) Grupo VI: 807 Serina/Treonina Cinases e 5 Tirosina Cinases (Tabela 6.16)



(g) Grupo VII: 1.016 Serina/Treonina Cinases e 1 Tirosina Cinase (Tabela 6.17)

Figura 6.19: Divisão das Proteínas Cinases em sete grupos pelo sistema de GP.

Tabela 6.15: Exceções no Grupo V obtido pelo sistema de GP para a divisão das Proteínas Cinases em sete grupos.

Proteína	Anotações
A7J1T0	Revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q2HZD7	Não-revisada. Apresenta anotação com o domínio InterPro IPR016245 (*tyrosine protein kinase, EGF/ERP/XmrK receptor*).
Q54TM7	Revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q5AUJ7	Não-revisada. Apresenta anotação com o domínio IPR002290, presente em Serina/Treonina Cinases ou em proteínas com dupla especificidade.
32 proteínas	Não-revisadas. Apresentam anotação com o domínio InterPro IPR008271, correspondente ao sítio ativo das Serina/Treonina Cinases.

Tabela 6.16: Exceções no Grupo VI obtido pelo sistema de GP para a divisão das Proteínas Cinases em sete grupos.

Proteína	Anotações
A0JN96	Não-revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com o domínio InterPro IPR002290, presente em Serina/Treonina Cinases ou em proteínas de dupla especificidade.
A2DGV6	Não-revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com o domínio InterPro IPR0008271, correspondente ao sítio ativo das Serina/Treonina Cinases.
Q4T0K5	Não-revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com o domínio InterPro IPR002290, presente em Serina/Treonina Cinases ou em proteínas de dupla especificidade.
Q5RAR7	Não-revisada. Apresenta anotação com domínio InterPro IPR002290, presente em Serina/Treonina Cinases ou em proteínas de dupla especificidade.
Q69U56	Não-revisada. Apresenta anotação com o termo GO *protein serine/threonine kinase activity* e com o domínio InterPro IPR0008271, correspondente ao sítio ativo das Serina/Treonina Cinases.

Tabela 6.17: Exceção no Grupo VII obtido pelo sistema de GP para a divisão das Proteínas Cinases em sete grupos.

Proteína	Anotações
Q6K3D4	Não-revisada. Não apresenta anotação.

No Grupo V, cujas proteínas são em maioria rotuladas como Tirosina Cinases, há 35 proteínas rotuladas como Serina/Treonina Cinases e uma como EGFR, de identificador UniProt Q2HZD7. Essa última não é uma entrada que foi manualmente revisada, mas está anotada com o domínio InterPro IPR016245 (*tyrosine protein kinase, EGF/ERP/XmrK receptor*), ou seja, apesar de ser uma Tirosina

Cinase como toda EGFR, ela deveria ter sido agrupada junto às EGFRs. Já entre as 35 rotuladas como Serina/Treonina Cinases, há duas manualmente curadas no UniProt: Q54TM7 e A7J1T0, ambas anotadas com o domínio InterPro IPR008271, que corresponde ao sítio ativo das Serina/Treonina Cinases, além do termo GO *protein serine/threonine kinase activity*. Das outras 33, apenas uma (Q5AUJ7) não apresenta o domínio InterPro que corresponde ao sítio ativo de Serina/Treonina Cinases, mas ela contém o domínio IPR002290, que está presente em proteínas que ou são Serina/Treonina Cinases, ou têm dupla especificidade. Essas anotações indicam que o sistema de GP incluiu erroneamente essas 36 proteínas em um grupo de Tirosina Cinases.

No Grupo VI, cuja maioria são Serina/Treonina Cinases, há cinco proteínas rotuladas como Tirosina Cinases, nenhuma das quais foi manualmente curada. Duas delas, de identificadores UniProt Q69U56 e A2DGV6, apresentam o domínio InterPro IPR008271, referente ao sítio ativo das Serina/Treonina Cinases, enquanto as outras três (Q5RAR7, Q4T0K5 e A0JN96) apresentam o domínio IPR002290, que ocorre em proteínas que são Serina/Treonina Cinases ou têm dupla especificidade. Além disso, entre as cinco, apenas a proteína Q5RAR7 não foi anotada com o termo GO *protein serine/threonine kinase activity*. Portanto, a análise dessas proteínas sugere que seus rótulos de subfamílias estão incorretos, e que elas foram corretamente inseridas no grupo de Serina/Treonina Cinases pelo sistema de GP desenvolvido neste trabalho.

Por último, no Grupo VII há uma proteína rotulada como Tirosina Cinase em meio a mais de mil Serina/Treonina Cinases. Analisando sua entrada no UniProt (identificador Q6K3D4), observa-se que ela não tem qualquer anotação que remeta a uma dessas subfamílias. Portanto, não é possível afirmar se o rótulo está incorreto ou se o erro foi do sistema de GP. Apesar de ter errado o posicionamento de 36 proteínas, o que corresponde a menos de 1,2% da família, os grupos gerados pelo sistema de GP mostraram-se muito mais relevantes em relação aos logotipos e à concordância com os rótulos de subfamílias do que aqueles produzidos pelo ASMC.

6.6 Estudo de Caso IV: DUF849

Bastard et al. (2014) analisaram 725 proteínas da família de função desconhecida DUF849 com uma abordagem de *ensemble clustering* para integrar resultados de diferentes estratégias de agrupamento, uma vez que cada método individual levou a grupos diferentes. Manualmente, foram atribuídos pesos a cada um dos agrupamentos primários para dar-lhes maior ou menor influência. O agrupamento final escolhido pelos autores contém 32 subfamílias consenso, com tamanhos variando de 3 a 130 proteínas. Segundo eles, o ASMC foi o principal método de agrupamento utilizado na abordagem de *ensemble clustering*. Sua execução automática produziu uma árvore com 84 grupos-folhas, em que apenas 31 continham mais de uma proteína. No entanto, o principal resultado apresentado pelos autores foram sete grupos obtidos manipulando manualmente a árvore produzida pelo ASMC, cortando-a como função da raiz, exceto por um ramo, que foi dividido em dois grupos. Essa manipulação per-

mite considerar, para cada ramo, níveis diferentes do agrupamento hierárquico, alterando a saída do algoritmo para construir grupos conforme convier.

Por essas razões, o sistema de GP foi executado para produzir 7, 32 e 84 grupos para a família DUF849, e os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados, respectivamente, nas Tabelas A.23, A.24 e A.25 do Apêndice A. A Figura 6.20 apresenta o logotipo que ilustra a composição do sítio ativo putativo das 725 proteínas da família DUF849. Destacando a dificuldade do problema, há 2×10^{16} possíveis combinações das proteínas dessa família em sete grupos.

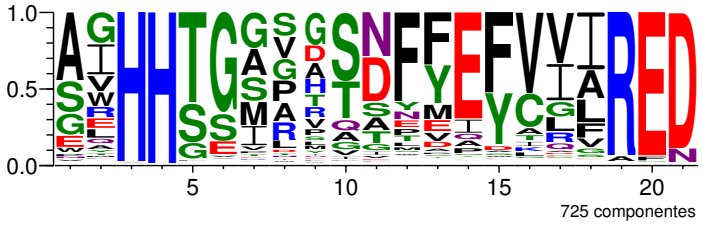


Figura 6.20: Composição do sítio ativo da família DUF849.

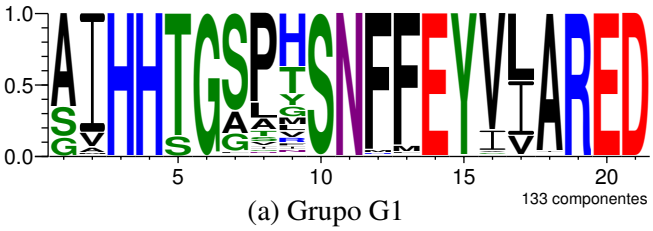
A Tabela 6.18 apresenta as combinações de dados produzidas pelo sistema de GP que levaram aos melhores resultados para essa família, empregando o grafo de similaridades construído apenas com os valores positivos das matrizes de similaridades calculadas por essas equações, e aplicando as taxas de 80% de cruzamento, 5% de reprodução e 15% de mutação, como já mencionado nos demais estudos de caso. Uma vez que a métrica de qualidade dos agrupamentos baseia-se no sítio ativo, era esperado que os melhores resultados fossem obtidos utilizando os tipos de dados relacionados a ele: $ASid$ e $ASscr$. Pode-se observar que diferentes tipos de dados estão envolvidos na equação obtida para cada quantidade de grupos.

Tabela 6.18: Combinações de dados que levaram à obtenção dos melhores agrupamentos para a família DUF849 para o grafo de similaridades construído apenas com valores positivos e com 80% de cruzamento, 5% de reprodução e 15% de mutação.

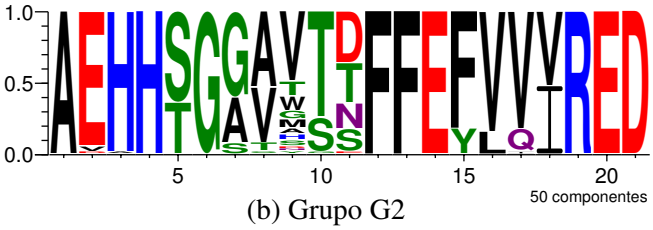
Grupos	Repetição	Equação
7	1	$2ASid + csmDist + 2neighborhood + 2seqAliL + strAliId$
32	4	$ASid + ASscr + difAliphRes$
84	4	$2ASid + difGRAVY + seqAliG$

Divisão da família DUF849 em sete grupos

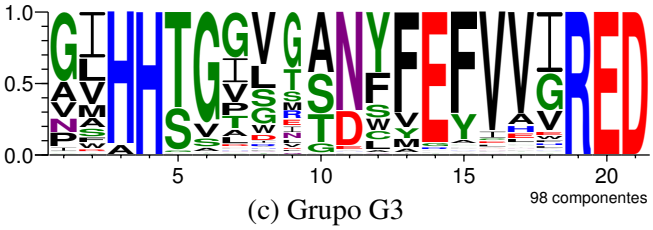
Os logotipos dos sete grupos principais reportados por Bastard et al. (2014), definidos por meio da manipulação manual do agrupamento hierárquico gerado pelo ASMC, são apresentados na Figura 6.21.



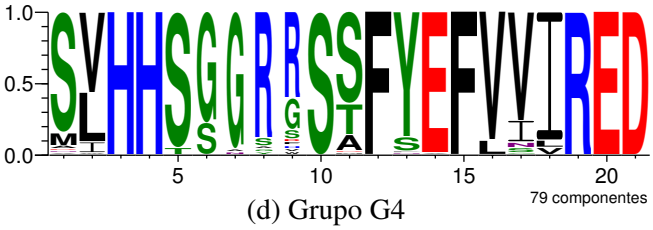
(a) Grupo G1



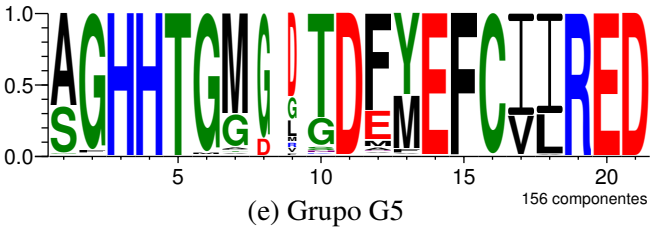
(b) Grupo G2



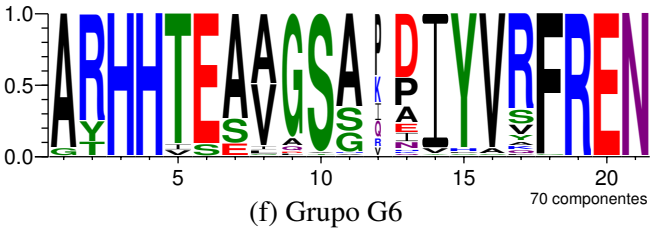
(c) Grupo G3



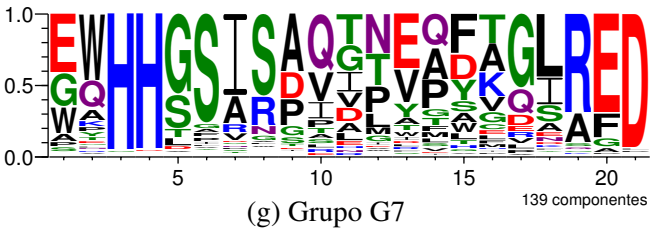
(d) Grupo G4



(e) Grupo G5



(f) Grupo G6



(g) Grupo G7

Figura 6.21: Divisão da família DUF849 em sete grupos produzidos por Bastard et al. (2014) pela manipulação manual do agrupamento hierárquico gerada pelo ASMC .

Segundo Bastard et al. (2014), que deram à família DUF849 o nome de "BKACE", do inglês β-*keto acid cleavage enzyme*, esses sete grupos possuem alta correlação com a natureza dos compostos transformados pelas proteínas que a eles pertencem, listadas a seguir.

- ◇ **Grupo G1:** substratos hidrofóbicos e polares não-carregados.
- ◇ **Grupo G2:** substratos KAH.
- ◇ **Grupo G3:** contém cinco subgrupos, sendo um de proteínas que não atuam como BKACE; um de BKACEs mistas; um que atua sobre β-cetoadipatos; um que atua sobre substratos hidrofóbicos e polares; e um que atua sobre benzoilacetato e β-cetohexanoato.
- ◇ **Grupo G4:** substratos negativamente carregados.
- ◇ **Grupo G5:** substratos positivamente carregados.
- ◇ **Grupo G6:** não são BKACEs, apresentando atividade de descarboxilação.
- ◇ **Grupo G7:** não são BKACEs.

No entanto, a divisão das atividades enzimáticas nesses grupos não é tão clara quanto essa lista sugere, uma vez que pode-se observar proteínas que têm atividade para substratos relacionados a outros grupos. Considerando os substratos para os quais Bastard et al. (2014) testaram atividade, a Tabela 6.19 mostra quantas vezes foi detectada uma atividade para cada grupo, considerando que há duas repetições de cada teste. Esses dados foram publicados pelos autores como material suplementar. Segundo eles, *_R* denota testes de atividade para reações reversas. A distribuição, entre os grupos manualmente produzidos por Bastard et al. (2014), do número de enzimas consideradas ativas para cada substrato mostra a complexidade de agrupar-se essa família em subfamílias isofuncionais devido à promiscuidade que ela apresenta.

A fim de comparar um agrupamento produzido de forma totalmente automática com os grupos manualmente definidos por Bastard et al. (2014), o sistema de GP foi executado para dividir as proteínas da família DUF849 em sete grupos. A Figura 6.22 apresenta os logotipos dos grupos obtidos com a configuração do sistema de GP em questão. Por tratar-se de uma família de proteínas de função desconhecida, a comparação dos resultados do sistema de GP com aqueles obtidos pelos autores é difícil. No entanto, pode-se observar uma grande correspondência entre os dois agrupamentos.

Os Grupos I e II obtidos pelo sistema de GP, contendo, respectivamente, 24 e 70 proteínas, estão relacionados ao Grupo G7 definido por Bastard et al. (2014), que contém 139 proteínas. Em nenhum deles há qualquer proteína com alguma atividade para os substratos considerados. O Grupo III obtido pelo sistema de GP corresponde exatamente ao Grupo G6 definido pelos autores, com 70 proteínas, com dezesseis atividades de β-cetoglutarato e uma atividade 4-hidroxibenzoilacetato.

Tabela 6.19: Distribuição de atividades enzimáticas nos sete grupos obtidos por Bastard et al. (2014) com a manipulação manual do agrupamento gerado pelo ASMC.

Natureza	**Substrato**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**
Catiônicos	(S)-KAH	-	8	-	-	11	-	-
	dehydrocarnitine	1	-	1	-	52	-	-
Aniônicos	*β-ketoadipate*	-	-	4	25	-	-	-
	β-ketoadipate_R	-	-	-	18	-	-	-
	β-ketoglutarate	-	-	-	15	-	16	-
	β-ketoglutarate_R	-	-	-	18	-	-	-
Não-iônicos polares	*3,5-dioxohexanoate*	10	-	4	-	-	-	-
	5-hydroxy-β-ketohexanoate	37	-	2	-	-	-	-
	6-acetamido-β-ketohexanoate	14	-	4	8	21	-	-
Apolares	*β-ketopentanoate*	1	1	-	-	-	-	-
	β-ketopentanoate_R	37	-	4	-		-	-
	β-ketoisocaproate	28	-	4	2	13	-	-
	β-ketoisocaproate_R	24	-	2	-	-	-	-
	(E)-β-ketohex-4-enoate_R	30	-	2	-	-	-	-
	β-ketohexanoate	7	-	-	-	-	-	-
	β-ketohexanoate_R	53	-	8	1	5	-	-
	7-methyl-β-ketooct-6-enoate	6	-	2	6	2	-	-
	β-ketooctanoate_R	25	-	4	-	2	-	-
	β-ketododecanoate	27	-	2	2	-	-	-
	benzoylacetate_R	-	-	4	-	-	-	-
	4-hydroxybenzoylacetate	2	1	1	2	3	1	-
	2-formamidobenzoylacetate	1	-	-	2	2	-	-

O Grupo IV obtido pelo sistema de GP, contendo 74 proteínas, está relacionado ao Grupo G4 definido pelos autores, que contém 79 proteínas. Não há diferença entre as atividades presentes entre esses dois grupos. O Grupo V obtido pelo sistema de GP está relacionado ao Grupo G1. Ambos apresentam 133 proteínas, no entanto o Grupo V tem uma proteína a mais com atividade. Há atividades para dezesseis substratos nesses grupos, e a diferença entre eles é que no Grupo G1 há duas atividades para 4-hidroxibenzoilacetato, enquanto para o Grupo V há três.

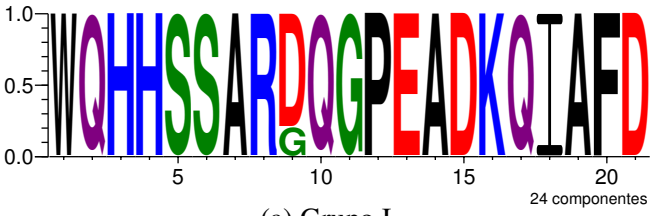
O Grupo VI obtido pelo sistema de GP, com 155 proteínas, está relacionado ao Grupo G5 definido pelos autores, com 156 proteínas. Não há diferença entre as atividades presentes em cada grupo. Por último, o Grupo VII obtido pelo sistema de GP, contendo 199 proteínas, está relacionado aos Grupos G2 e G3, que contêm, respectivamente, 50 e 98 proteínas. De fato, eles reportaram ter

manualmente dividido um dos ramos de árvore gerada pelo ASMC para produzir esses dois grupos. A diferença entre atividades consiste de uma atividade 4-hidroxibenzoilacetato no Grupo G2 que não está presente no Grupo VII, pois a enzima correspondente foi inserida no Grupo V, concentrando melhor as enzimas ativas para esse substrato.

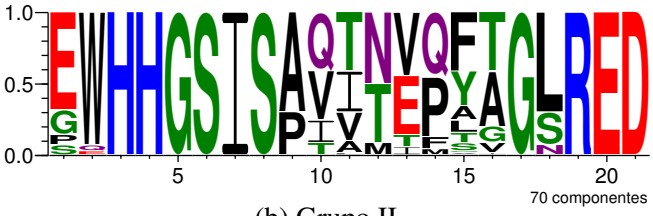
Apesar de tratar-se de uma técnica completamente automática, ainda assim o sistema de GP foi capaz de obter grupos muito similares àqueles produzidos manualmente, até mesmo concentrando um pouco melhor as enzimas ativas para 4-hidroxibenzoilacetato. O único problema nesse estudo de caso foi que o sistema de GP julgou mais relevante, em termos de composição de sítio ativo, quebrar o Grupo G7 em dois grupos distintos bastante homogêneos, enquanto a manipulação manual do agrupamento hierárquico produzido pelo ASMC manteve esse grupo, uma vez que trata-se de enzimas sem atividade para os substratos estudados. Assim, fica demonstrada a capacidade e utilidade da metodologia aqui desenvolvida para detectar subfamílias possivelmente isofuncionais também em famílias proteicas de função desconhecida.

Os dez resíduos mais importantes para discriminar entre esses grupos gerados pelo sistema de GP são listados a seguir, em ordem da importância definida pelo valor de MI do resíduo na sua respectiva posição.

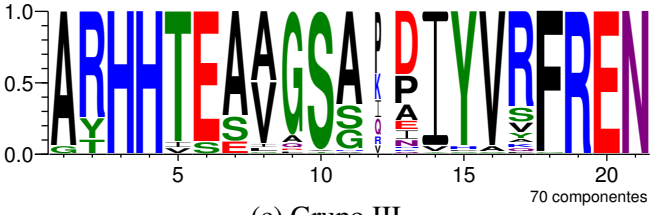
- ◇ **Grupo I:** K (16), F (20), D (15), W (1), Q (2), A (19), P (12), A (14), Q (17) e G (11).
- ◇ **Grupo II:** G (17), G (5), W (2), I (7), S (8), S (6), A (9), E (1), N (12) e Q (14);
- ◇ **Grupo III:** N (21), F (18), I (14), E (6), R (2), R (17), A (11), A (7), Y (15) e G (9);
- ◇ **Grupo IV:** S (5), S (11), S (1), Y (13), V (2), R (8), I (18), S (10), V (17) e R (9);
- ◇ **Grupo V:** A (18), N (11), P (8), I (2), S (7), Y (15), F (13), L (17), S (10) e H (9);
- ◇ **Grupo VI:** G (2), C (16), D (11), M (7), I (17), T (10), F (15), G (8), T (5) e M (13); e
- ◇ **Grupo VII:** F (13), V (17), G (1), E (2), A (10), Y (12), V (18), V (8), G (18) e V (16);



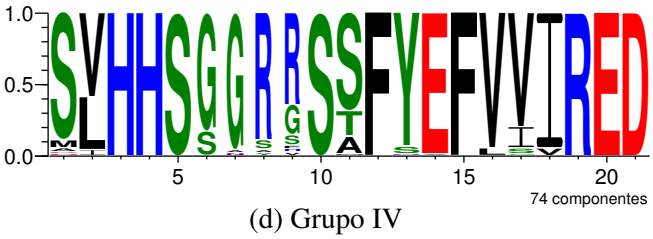
(a) Grupo I



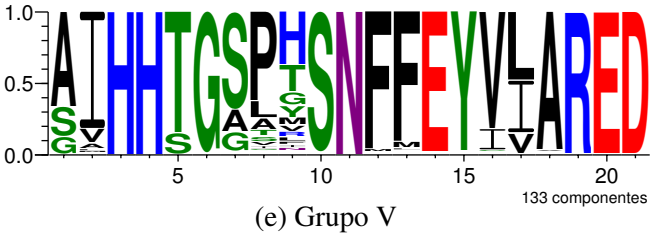
(b) Grupo II



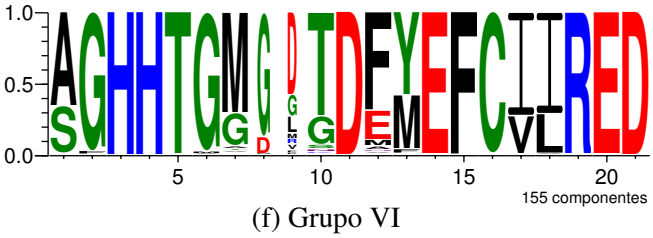
(c) Grupo III



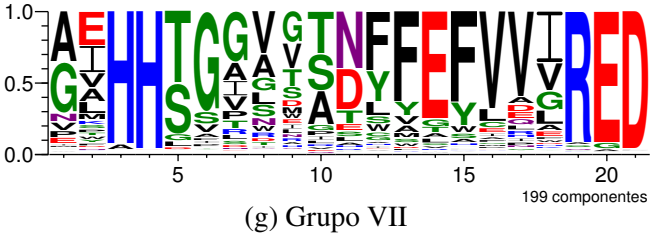
(d) Grupo IV



(e) Grupo V



(f) Grupo VI



(g) Grupo VII

Figura 6.22: Agrupamento da família DUF849 em sete grupos pelo sistema de GP.

6.7 Síntese

A metodologia desenvolvida nesta tese pode ser aplicada a qualquer família de proteínas, mesmo as de função desconhecida, uma vez que a técnica de agrupamento empregada independe da existência de rótulos de subfamília. No entanto, para avaliar seu desempenho e compará-la ao ASMC (Melo-Minardi et al., 2010), uma técnica similar da literatura, foram realizados quatro estudos de caso com famílias de proteínas às quais o ASMC foi anteriormente aplicado: Nucleotidil Ciclases, Serino Proteases, Proteínas Cinases e a família de proteínas de função desconhecida DUF849. Para uma comparação direta, foram estudados os mesmos conjuntos de proteínas utilizados pelos autores, atualizando-os apenas para eliminar aquelas que foram removidas do UniProt desde então.

Inicialmente, foi feita uma comparação das duas formas consideradas neste trabalho para construir, a partir da combinação de dados produzida pelo sistema de GP, o grafo de similaridades utilizado pelo algoritmo de agrupamento espectral: considerar somente valores positivos ou utilizar todos, redimensionando-os para o intervalo [0, 1]. Concluiu-se que, quando há diferença significativa, são melhores os resultados obtidos ao empregar o grafo de similaridades que usa apenas os valores positivos da matriz de similaridades.

Em seguida, foram comparadas diferentes configurações de valores das taxas de cruzamento, mutação e reprodução do sistema de GP, a fim de determinar a configuração de parâmetros que obtém os melhores agrupamentos. Concluiu-se que a configuração que mais ocorreu entre os melhores resultados foi utilizando as taxas de 80% de cruzamento, 5% de reprodução e 15% de mutação. Então, os melhores resultados obtidos pelo sistema de GP, empregando o grafo de similaridades construído apenas com valores positivos da matriz de similaridades e essa configuração de parâmetros, foram analisados e comparados aos agrupamentos produzidos pelo ASMC (Melo-Minardi et al., 2010) para as famílias Nucleotidil Ciclases, Serino Proteases e Proteínas Cinases, e com aqueles gerados por Bastard et al. (2014) para a família DUF849. Além disso, foram analisados os resíduos mais importantes para discriminar entre os diferentes grupos produzidos pelo sistema de GP.

Observa-se nos trabalhos de Melo-Minardi et al. (2010) e Bastard et al. (2014) que o ASMC costuma ser utilizado para fornecer um agrupamento hierárquico inicial das proteínas, que depois é manualmente manipulado a fim de obter grupos que os autores julgam mais interessantes. Essa manipulação permite considerar níveis diferentes da hierarquia para cada ramo da árvore gerada, alterando o resultado do algoritmo. No entanto, a fim de comparação com a metodologia aqui desenvolvida, foram considerados, para as famílias Nucleotidil Ciclases, Serino Proteases e Proteínas Cinases, os dois primeiros níveis do agrupamento hierárquico produzido pelo ASMC, por isso o passo de agrupamento do algoritmo foi reexecutado para os conjuntos atualizados de proteínas. Já considerando a família de proteínas de função desconhecida DUF849, os resultados do sistema de GP aqui desenvolvido foram comparados aos sete grupos criados por Bastard et al. (2014) pela manipulação manual do agrupamento produzido pelo ASMC.

Os resultados mostraram que o sistema de GP separou com sucesso a família de Nucleotidil Ciclases em suas duas subfamílias. Já o ASMC, ao invés de encontrar grupos relacionados às duas subfamílias existentes, priorizou a criação de subgrupos das Adenilato Ciclases, enquanto a maior parte da família foi inserida em um grupo grande em relação aos demais contendo todas as Guanilato Ciclases e a maioria das Adenilato Ciclases. Apenas no segundo nível da hierarquia o ASMC gerou um grupo específico de Guanilato Ciclases, mas fragmentou as Adenilato Ciclases em cinco subgrupos, mesmo não havendo grande variabilidade nessa subfamília. Esse estudo de caso mostrou que os grupos obtidos pelo sistema de GP apresentam maior concordância com a divisão da família em suas duas subfamílias do que aqueles produzidos pelo ASMC. Além disso, quando há uma quantidade maior de grupos do que o número de subfamílias existente, o sistema de GP tende a produzir grupos que apresentam maiores diferenças entre si do que aqueles gerados pelo ASMC, que subdividiu grupos que já eram bastante homogêneos.

No estudo de caso com as Serino Proteases, observou-se que essa foi a família mais difícil de separar em subfamílias tanto para o sistema de GP quanto para o ASMC, devido ao imenso desbalanceamento entre as subfamílias. A grande variabilidade de resíduos existente entre as Tripsinas, que representam 94,7% da família, levou as técnicas a encontrarem subgrupos de Tripsinas mais facilmente do que as outras subfamílias, uma vez que eles são mais volumosos que elas. Por essa razão, ambas as técnicas somente conseguiram grupos próprios para as subfamílias pequenas de Calicreínas e Quimotripsinas após quebrarem as Tripsinas em vários subgrupos. Ainda assim, os subgrupos obtidos pelo sistema de GP para as Tripsinas mostraram-se relevantes, tendo sido encontrados grupos de Protrombinas, algumas das quais foram manualmente curadas, por exemplo, e um grupo de Calicreínas maior do que o observado por Melo-Minardi et al. (2010), incluindo várias Calicreínas com anotações revisadas que estavam rotuladas como Tripsinas.

Para as Proteínas Cinases, o sistema de GP conseguiu produzir grupos relevantes e cuja correspondência com as subfamílias existentes foi quase total, apesar de alguns equívocos no agrupamento em menos de 2% das proteínas da família. Já o ASMC acabou por priorizar a produção de um grupo de proteínas contendo muitas lacunas, o que não está relacionado à função. Além disso, mesmo aumentando a quantidade de grupos ao considerar o segundo nível do agrupamento hierárquico do ASMC, as duas subfamílias continuaram misturadas.

O último estudo de caso envolveu a família de proteínas de função desconhecida DUF849 estudada por Bastard et al. (2014). O principal resultado apresentado pelos autores foram sete grupos obtidos manipulando manualmente a árvore produzida pelo ASMC. A distribuição, entre esses sete grupos, do número de enzimas consideradas ativas para cada substrato testado pelos autores, mostrou a complexidade de agrupar essa família em subfamílias isofuncionais devido à promiscuidade que ela apresenta. Ainda assim, o agrupamento produzido pela metodologia totalmente automática aqui desenvolvida tem uma grande correspondência com os grupos produzidos manualmente por Bastard

et al. (2014). Assim, ficou demonstrada a capacidade e utilidade desta metodologia para detectar subfamílias possivelmente isofuncionais também em famílias de função desconhecida.

Dado o objetivo de detectar subfamílias isofuncionais em uma família de proteínas, considera-se interessante um grupo que contenha resíduos exclusivos ou quase exclusivos para as diferentes posições do sítio ativo putativo. A métrica de qualidade dos agrupamentos aqui empregada, baseada na informação mútua, reflete esse objetivo, uma vez que ela é uma medida da associação entre um resíduo em determinada posição e um dado grupo. Os resultados apresentados mostram que a utilização dessa métrica pelo sistema de GP como função de aptidão dos indivíduos fez com que os agrupamentos obtidos realmente envolvessem grupos cujas composições são as mais diferentes possível uns dos outros, conforme observado nos logotipos, obtendo assim agrupamentos com maior concordância com as subfamílias existentes que aqueles produzidos pelo ASMC. A Tabela 6.20 apresenta uma comparação dos valores de MI para os agrupamentos produzidos pelos dois níveis considerados da hierarquia do ASMC e aqueles produzidos pelo sistema de GP para a mesma quantidade de grupos, considerando a configuração de 80% de cruzamento, 5% de reprodução e 10% de mutação. Pode-se observar que os valores de MI são maiores para o sistema de GP, o que indica que a métrica está de acordo com o objetivo de encontrar grupos que contenham resíduos (quase) exclusivos para determinadas posições, como havia sido observado nos logotipos dos grupos.

Tabela 6.20: Valores de MI para os agrupamentos produzidos pelo ASMC e pelo sistema de GP para cada família e cada quantidade de grupos.

Família	**Grupos**	**ASMC**	**Sistema de GP**
Nucleotidil Ciclases	3	22,16	22,35
	6	14,11	16,13
Serino Proteases	4	16,58	17,71
	11	10,59	12,09
Proteínas Cinases	3	67,46	102,94
	7	45,99	50,70

Uma prática comum na literatura é a realização de múltiplos agrupamentos de um mesmo conjunto de dados com diferentes números de grupos, escolhendo o número "ótimo" de grupos como aquele com o maior valor para uma dada métrica de qualidade. Várias métricas foram testadas nesta tese visando a tratar o problema de identificar o número ideal de grupos em uma família proteica, tais como coeficiente de silhueta, BetaCV, *Normalized Cut*, índice de Dunn, informação mútua (pontual) entre grupos, entropia relativa, verossimilhança, assim como diversas variações destas. No entanto, como esta tese trabalha com famílias proteicas, toda proteína tem algum grau de similaridade com todas as demais. Sendo assim, as métricas de qualidade de agrupamento tendem a ser melhores quando (quase) todas as proteínas são atribuídas a um mesmo grupo, levando assim a agrupamentos em que

um único grupo consiste da maior parte da família, enquanto outros contêm muito poucas proteínas. Embora se saiba que subfamílias existem nas famílias estudadas, as métricas de qualidade testadas não refletem isso. Assim, apesar dos melhores esforços, não foi encontrada uma métrica de qualidade que permitiria comparar agrupamentos contendo diferentes números de grupos para determinar o número ideal de grupos em uma família do Pfam. Deste modo, no momento utiliza-se a MI para comparar agrupamentos contendo o mesmo número de grupos, enquanto inspeciona-se visual e manualmente os logotipos e composições dos grupos a fim de comparar agrupamentos com diferentes números de grupos, como feio por Melo-Minardi et al. (2010) e Bastard et al. (2014).

A Tabela 6.21 apresenta o número de vezes que cada tipo de dados aparece entre as combinações que levaram aos melhores agrupamentos considerando 80% de cruzamento, 5% de reprodução e 15% de mutação, além da média de pesos com que ocorrem entre essas equações, que foram apresentadas previamente nas Tabelas 6.3, 6.11, 6.12 e 6.18. Uma vez que a métrica de qualidade dos agrupamentos, a informação mútua (MI), é calculada com base na composição do sítio ativo, era esperado que os dados de similaridade dele derivados ($ASid$ e $ASscr$) teriam participação nas combinações de dados que produziram os melhores resultados, como pode ser observado na Tabela 6.21. No entanto, sua combinação com outros tipos de dados contribuiu para a melhoria da qualidade dos agrupamentos obtidos. Entre esses outros tipos, destacaram-se as identidade dos alinhamentos estruturais ($strAliId$), as pontuações dos alinhamentos globais de sequências ($seqAliG$) e a similaridade de anotações de termos do GO (go). Esses dados são comumente empregados, separadamente, por métodos de anotação de função baseados em homologia, de modo que sua presença entre as combinações de dados mais úteis encontradas pelo sistema de GP deve-se à correspondência, embora imperfeita, que a similaridade em relação a esses dados tem com a similaridade funcional.

Foi possível notar que o sistema de GP encontra agrupamentos semelhantes com combinações de dados muito diferentes. Isso provavelmente deve-se ao fato de que os tipos de dados estudados não são independentes. A redundância existente entre eles impossibilitou que se chegasse a uma conclusão sobre a semântica das equações obtidas pelo sistema de GP. No entanto, foi possível observar a tendência geral de que usar mais tipos de informações leva à melhoria dos resultados. Uma vez que os melhores resultados obtidos pelo sistema de GP envolvem a combinação de múltiplos tipos de dados, confirmou-se a hipótese inicial desta tese de que similaridades entre proteínas em relação a diferentes domínios do conhecimento podem ser interpretadas como evidências de similaridade funcional.

A separação de famílias proteicas em subfamílias isofuncionais é um problema complexo que é dificultado ainda mais pela ocorrência de forte desbalanceamento entre as subfamílias, como para as Serino Proteases, e pela existência de proteínas que desempenham múltiplas funções, como observado entre as proteínas da família DUF849. Os resultados aqui apresentados mostram avanço em relação ao ASMC, uma vez que os agrupamentos obtidos estão mais de acordo com as subfamílias existentes nas famílias proteicas estudadas. No entanto, é preciso investigar mais a fundo o motivo de ambas as

Tabela 6.21: Ocorrências dos tipos de dados entre as equações que levaram aos melhores agrupamentos no sistema de GP.

Tipo de Dados	**Ocorrências**	**Peso Médio**
$ASid$	26	1,63
$strAliId$	14	0,67
$ASscr$	13	0,77
$seqAliG$	12	0,60
go	8	0,47
$difAliphRes$	7	0,40
$seqAliL$	6	0,40
$interpro$	6	0,30
$cooccurence$	6	0,23
$strAliScr$	5	0,47
$strAliSize$	5	0,27
$csmDist$	4	0,37
$neighborhood$	4	0,27
$difChargedRes$	4	0,23
$difIsoPoint$	4	0,20
$difAcidicRes$	4	0,17
$aaCompDist$	4	0,17
$difMolWeight$	4	0,17
$difBasicRes$	4	0,13
$coexpression$	3	0,30
$difInstab$	3	0,27
$difPolarRes$	3	0,27
$difAromRes$	3	0,17
$difGRAVY$	3	0,13

técnicas subdividirem uma mesma subfamília em grupos menores. Sabe-se que uma proteína como, por exemplo, a Tripsina, apresenta diferenças dependendo da espécie em que ocorre. Uma vez que as proteínas aqui estudadas são oriundas de diferentes espécies, é possível que seja esse o motivo pelo qual uma mesma subfamília é subdividida em grupos menores mesmo que já seja muito homogênea.

Capítulo 7
Conclusões

Apesar dos melhores esforços de pesquisa, uma quantidade substancial e cada vez maior de proteínas previstas ainda apresentam função desconhecida (Babbitt, 2003). O aumento sem precedentes do número de novas sequências proteicas sendo produzidas por projetos de genômica e proteômica, além das muitas estruturas proteicas de função desconhecida sendo resolvidas pela genômica estrutural, enfatiza diretamente a necessidade de métodos computacionais para determinar, rápida e confiavelmente, as funções moleculares e celulares dessas proteínas, uma vez que a investigação experimental é difícil e tem altos custos financeiro e temporal (Zhang e Kim, 2003; Lee et al., 2007). No entanto, atualmente não existem abordagens de larga escala capazes de revelar a função de todos os genes hipotéticos nos genomas já sequenciados. Esse objetivo só é alcançável por meio dos esforços de biólogos experimentais, computacionais e estruturais (Galperin e Koonin, 2010). O presente trabalho é um esforço computacional visando a dar um passo em direção a esse objetivo.

A abordagem mais comum para a anotação funcional de proteínas é a transferência de anotações por meio de homologia, que apresenta várias limitações, visto que homologia não implica, necessariamente, em isofuncionalidade (Smith, 2012). De fato, esses métodos são considerados uma das principais fontes de erros de anotação em consequência da aplicação excessivamente liberal de homologia (Lee et al., 2007). Além disso, a literatura mostra que a utilização de apenas um tipo de dados como, por exemplo, similaridade de sequências, é insuficiente para transferir anotações com precisão, devido à plasticidade das funções proteicas, à imensa quantidade de fatores envolvidos na determinação de uma função, e à consequente complexidade do problema de anotação automática. A combinação de diversos tipos de informação é, portanto, crucial (Furnham et al., 2012).

Dada uma família composta por proteínas com múltiplos enovelamentos ou funções, a determinação de possíveis subfamílias pode levar a informações importantes sobre a função e estrutura de uma proteína de função desconhecida associada à família, assim como sobre a diversificação funcional adquirida pela família ao longo da evolução (Devos e Valencia, 2000). Assim, a família é dividida em subtipos que compartilham funções específicas mas que não são comuns à família como um todo (Capra e Singh, 2008). Acredita-se que a determinação dessas subfamílias seja um primeiro passo para reduzir a complexidade do problema de anotação funcional de proteínas. Por isso, o propósito desta tese foi a detecção de subfamílias isofuncionais em uma família de proteínas de função desconhecida, além da identificação dos resíduos responsáveis pela diferenciação entre elas. Para tanto, são empregados algoritmos de agrupamento, uma das principais tarefas da mineração de dados, que

permitem agrupar um conjunto de objetos de modo que aqueles em um mesmo grupo sejam mais similares entre si do que a objetos de outros grupos. Entre os muitos algoritmos de agrupamento apresentados na literatura, optou-se por empregar o agrupamento espectral, uma vez que ele é capaz de resolver problemas bastante complexos, como o caso em que, quando plotados, os objetos pertencentes a cada grupo são posicionados em espirais entrelaçadas, não separáveis por algo simples como uma reta ou curva. Isso é necessário para o cenário de aplicação desta tese, uma vez que trata-se de famílias de proteínas homólogas, em que uma divisão em subfamílias raramente é fácil de detectar.

Considerando que as muitas dificuldades enfrentadas para a realização de anotação automática de função podem ser estendidas ao problema de determinar subfamílias, para contornar as principais falhas comuns a métodos de anotação baseados em homologia, é adotada nesta tese uma abordagem que integra diversos tipos de dados que são possíveis indicadores de similaridade entre proteínas. Assim, neste trabalho, a similaridade entre pares de proteínas em relação a vários tipos de dados é estudada e interpretada como evidência, ainda que fraca, de similaridade funcional. Uma vez que é infactível realizar testes para todas as formas possíveis de integrar os tipos de dados aqui estudados, optou-se por realizar a integração de dados empregando programação genética (GP), uma técnica de aprendizagem de máquina. A partir de uma combinação inicial aleatória dos dados, o sistema de GP aprende, ao longo das gerações, quais combinações de dados levam aos melhores resultados. Isso a torna excelente para integrar dados biológicos, que são comumente imprecisos e, muitas vezes, encontram-se disponíveis apenas para um subconjunto das proteínas de interesse. A programação genética mostrou-se muito eficiente para integrar os diferentes tipos de dados biológicos estudados, encontrando combinações que obtiveram agrupamentos compatíveis com as subfamílias conhecidas.

O principal objetivo deste trabalho foi analisar o modo como a integração de informações provenientes de diferentes domínios de conhecimento é capaz de direcionar um processo não-supervisionado de agrupamento a detectar, em uma família de proteínas de função desconhecida, subfamílias possivelmente isofuncionais. Para tanto, foram realizados quatro estudos de caso aplicando a metodologia aqui desenvolvida às famílias proteicas de Nucleotidil Ciclases, Serino Proteases, Proteínas Cinases e à família de proteínas com função desconhecida DUF849, e comparando seus resultados àqueles obtidos por uma técnica similar, o ASMC (Melo-Minardi et al., 2010).

Foram comparadas duas formas para construir, a partir da combinação de dados produzida pelo sistema de GP, o grafo de similaridades utilizado pelo algoritmo de agrupamento espectral. Concluiu-se que são melhores os resultados obtidos ao empregar o grafo que usa apenas os valores positivos da matriz de similaridades. Além disso, diferentes configurações de valores das taxas de cruzamento, mutação e reprodução do sistema de GP foram comparadas, a fim de determinar a configuração de parâmetros que obtém os melhores agrupamentos. Concluiu-se que a configuração que mais ocorreu entre os melhores resultados foi utilizando as taxas de 80% de cruzamento, 5% de reprodução e 15% de mutação. Então, os melhores agrupamentos obtidos pelo sistema de GP, empregando o grafo de

similaridades construído apenas com valores positivos da matriz de similaridades e essa configuração de parâmetros, foram analisados e comparados àqueles produzidos pelo ASMC (Melo-Minardi et al., 2010) para as famílias Nucleotidil Ciclases, Serino Proteases e Proteínas Cinases, e com aqueles gerados por Bastard et al. (2014) para a família DUF849. Além disso, foram analisados os resíduos mais importantes para discriminar entre os diferentes grupos produzidos pelo sistema de GP.

Consistentemente, os melhores resultados obtidos pelo sistema de GP envolvem a combinação de múltiplos tipos de dados, o que confirma a premissa deste trabalho de que similaridades entre proteínas em relação a diferentes domínios do conhecimento podem ser utilizadas como evidências de similaridade funcional. Uma vez que a métrica de qualidade, a informação mútua (MI), é calculada com base na composição do sítio ativo, era esperado que os dados de similaridade dele derivados teriam participação nas combinações que produziram os melhores resultados. Outros tipos de dados que destacaram-se foram as identidades dos alinhamentos estruturais, as pontuações dos alinhamentos globais de sequências, as similaridades de termos do GO, as pontuações de coocorrência dos genes codificadores em diferentes genomas, as similaridades de domínios do InterPro e as pontuações dos alinhamentos locais de sequências. Para cada família proteica e cada quantidade de grupos considerada, diferentes combinações de dados obtiveram os melhores agrupamentos.

A análise dos perfis dos grupos de proteínas obtidos pelo sistema de GP mostrou que, para um dado número de grupos, a métrica de qualidade dos agrupamentos empregada (MI) reflete bem o objetivo de encontrar grupos que contenham, em determinadas posições do sítio ativo, resíduos que são exclusivos ou quase exclusivos às proteínas que o compõem. Essas são posições determinantes de especificidade, que diferenciam as subfamílias. Os agrupamentos obtidos apresentam uma grande correspondência com as subfamílias conhecidas das famílias proteicas estudadas. Infelizmente, o modo como a MI é calculada inviabiliza seu uso para encontrar o número ideal de grupos em uma família proteica, uma vez que seu valor diminui à medida que o número de grupos aumenta. No entanto, esse mesmo modo de cálculo, que envolve valores parciais de MI para cada resíduo, em cada posição e em cada grupo, permite avaliar numericamente quais resíduos, em quais posições, mais diferenciam um grupo dos demais. De fato, para as famílias cujas posições determinantes de especificidade são conhecidas, esses resíduos estavam entre os considerados mais importantes para diferenciar grupos segundo essa métrica.

Na comparação da metodologia desenvolvida neste trabalho com a técnica similar da literatura, o ASMC (Melo-Minardi et al., 2010), observou-se que o sistema de GP obteve agrupamentos em maior acordo com as subfamílias do que aqueles produzidos pelo ASMC para duas das famílias às quais a metodologia foi aplicada. Para os outros dois estudos de caso, os resultados foram equivalentes. O estudo de caso com a família de Nucleotidil Ciclases mostrou que os grupos obtidos pelo sistema de GP apresentam maior concordância com a divisão da família em suas duas subfamílias do que os grupos produzidos pelo ASMC, uma vez que ele foi capaz de separar essa família com

sucesso. Além disso, quando há uma quantidade maior de grupos do que o número de subfamílias existente, o sistema de GP tende a produzir agrupamentos com diferenças maiores entre os grupos do que aqueles produzidos pelo ASMC, que fragmentou uma das subfamílias em subgrupos, mesmo não havendo grande variabilidade.

Outro caso de sucesso do sistema de GP foi para a família de Proteínas Cinases, para a qual ele produziu grupos cuja correspondência com as subfamílias existentes foi quase total, apesar de alguns equívocos no agrupamento em menos de 2% das proteínas da família. Já o ASMC acabou por priorizar a produção de um grupo de proteínas contendo muitas lacunas e, mesmo aumentando a quantidade de grupos ao considerar o segundo nível do agrupamento hierárquico, as duas subfamílias continuaram misturadas.

Devido ao imenso desbalanceamento, a família de Serino Proteases mostrou-se a mais difícil de separar em subfamílias tanto para o sistema de GP quanto para o ASMC. Como as Tripsinas, que representam 94,7% da família, apresentam grande variabilidade, ambas as técnicas encontraram subgrupos de Tripsinas mais facilmente do que as outras subfamílias. Ainda assim, os subgrupos obtidos pelo sistema de GP mostraram-se relevantes, tendo sido encontrados grupos de Protrombinas, algumas das quais foram manualmente curadas, por exemplo, e um grupo de Calicreínas maior do que o observado por Melo-Minardi et al. (2010), incluindo várias Calicreínas com anotações revisadas.

O último estudo de caso envolveu a família de proteínas de função desconhecida DUF849 estudada por Bastard et al. (2014), que manualmente manipularam a árvore produzida pelo ASMC, gerando sete grupos. Tal manipulação permite considerar, para cada ramo, níveis diferentes do agrupamento hierárquico, alterando a saída do algoritmo para construir grupos conforme convier. Essa família mostrou-se difícil de agrupar em subfamílias isofuncionais devido à promiscuidade que apresenta. Ainda assim, o agrupamento produzido pela metodologia totalmente automática aqui desenvolvida obteve uma grande correspondência com os grupos produzidos manualmente pelos autores.

Os resultados aqui obtidos representam avanço em relação àqueles da literatura porque observou-se que o sistema de GP, que é totalmente automatizado, obteve agrupamentos melhores que aqueles produzidos pelo ASMC para duas das famílias às quais a metodologia foi aplicada, e resultados equivalentes para os outros dois estudos de caso, incluindo um agrupamento definido com intervenção manual. Em geral, os grupos gerados pelo sistema de GP foram mais relevantes que aqueles produzidos pelo ASMC, por apresentarem maiores diferenças entre si, enquanto o ASMC tende a subdividir grupos que já são bastante homogêneos em relação aos demais. Adicionalmente, o ASMC não mostrou-se muito estável em relação aos grupos gerados, pois a atualização dos conjuntos de proteínas estudados o levou a produzir, com as mesmas configurações de parâmetros, agrupamentos bem diferentes daqueles obtidos para o conjunto original, pouco maior.

Adicionalmente, o algoritmo de agrupamento hierárquico empregado pelo ASMC não permite que eventuais erros durante o processo sejam reparados, uma vez que, quando um nó da árvore que

representa o agrupamento é subdividido, não é possível que uma proteína mude de um ramo a outro. Assim, caso uma subdivisão seja realizada equivocadamente durante o processo, o erro será propagado aos demais níveis da hierarquia e jamais poderá ser consertado. Já no caso de um agrupamento de particionamento como empregado na metodologia aqui proposta, em que define-se a priori o número de grupos buscados, proteínas podem migrar para um grupo que torne-se mais adequado conforme o número de grupos aumenta, o que seria equivalente a uma mudança de ramo na árvore para reparar um erro.

Os resultados deste trabalho comprovam que a metodologia aqui proposta apresentou-se útil e capaz de detectar subfamílias possivelmente isofuncionais em famílias de proteínas, mesmo em uma família de função desconhecida. Esse tipo de metodologia, que integra informações provenientes de fontes diversas, é de grande interesse para um cenário de aplicação como o desta tese, visto que a função molecular de uma proteína é determinada por diversos fatores, e a complementaridade das diferentes fontes de dados permite que os algoritmos trabalhem com o máximo de informação possível. Além disso, a capacidade de lidar com dados incertos e incompletos é fundamental, uma vez que dados biológicos em muitos casos são inerentemente imprecisos devido à natureza dinâmica dos fenômenos investigados, assim como alguns tipos de dados são relativamente escassos.

Perspectivas

A fim de comparar a metodologia desenvolvida nesta tese à técnica similar existente na literatura, o ASMC (Melo-Minardi et al., 2010), foram feitos estudos de caso com famílias proteicas às quais esse método foi previamente aplicado. Como anteriormente discutido, os resultados apresentados mostraram que a técnica aqui proposta levou a agrupamentos melhores que aqueles obtidos pelo ASMC. No entanto, é necessário testar o desempenho da técnica para outras famílias proteicas que contenham um número maior de subfamílias de diferentes especificidades. Além disso, é necessário investigar mais profundamente a semântica das combinações de dados produzidas pelo sistema de GP. Uma rede de variáveis dependentes ou sinônimas pode ajudar a melhor compreender a redundância existente entre os tipos de dados estudados como evidência de similaridade funcional. A eliminação redundâncias do sistema de GP pode facilitar a análise semântica das combinações de dados obtidas, assim como melhorar a qualidade dos agrupamentos gerados. Isso requer a realização de experimentos com diferentes subconjuntos dos tipos de dados estudados.

Considerando a dificuldade encontrada para dividir a família de Serino Proteases devido ao grande desbalanceamento existente entre suas subfamílias, seria interessante aplicar métodos de amostragem a famílias como essa, a fim de avaliar o desempenho da metodologia aqui proposta em bases de dados melhor balanceadas. Além disso, considerando a complexidade adicional existente para famílias de proteínas promíscuas, como a DUF849 mostrou-se ser, é preciso investigar a possibilidade de adaptação da técnica aqui proposta para utilizar algoritmos de agrupamento *fuzzy*.

Diferentemente do algoritmo de particionamento aqui empregado, em que cada proteína está em um único grupo, um algoritmo de agrupamento *fuzzy* retorna, para cada proteína, um grau de pertinência da mesma a cada grupo existente. Dessa forma, uma proteína que desempenha múltiplas funções poderia estar, ao mesmo tempo, em grupos diferentes.

Por fim, esforços adicionais são necessários na busca de uma métrica de qualidade de agrupamento apropriada para o cenário de aplicação deste trabalho que permita comparar agrupamentos contendo diferentes números de grupos, de modo a determinar o número ideal de grupos em uma família de proteínas.

Referências Bibliográficas

Alberts, B.; Bray, D.; Hopkin, K.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K. & Walter, P. (2010). *Essential cell biology*. Garland Science, 3 edição.

Arakaki, A. K.; Huang, Y. & Skolnick, J. (2009). EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinformatics*, 10:107.

Arakaki, A. K.; Tian, W. & Skolnick, J. (2006). High precision multi-genome scale reannotation of enzyme function by EFICAz. *BMC Genomics*, 7:315.

Ares, M. E.; Parapar, J. & Barreiro, A. (2012). An experimental study of constrained clustering effectiveness in presence of erroneous constraints. *Information Processing & Management*, 48(3):537--551.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M. & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25(25-9).

Babbitt, P. C. (2003). Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol.*, 7(2):230--237.

Bach, F. R. & Jordan, M. I. (2003). Learning spectral clustering. In *Advances in Neural Information Processing Systems 16*. MIT Press.

Bartlett, G. J.; Porter, C. T.; Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, 324:105--21.

Bastard, K.; Smith, A. A. T.; Vergne-Vaxelaire, C.; Perret, A.; Zaparucha, A.; Melo-Minardi, R. D.; Mariage, A.; Boutard, M.; Debard, A.; Lechaplais, C.; Pelle, C.; Pellouin, V.; Perchat, N.; Petit, J.-L.; Kreimeyer, A.; Medigue, C.; Weissenbach, J.; Artiguenave, F.; Berardinis, V. D.; Vallenet, D. & Salanoubat, M. (2014). Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.*, 10:42–49.

Basu, S.; Davidson, I. & Wagstaff, K. L. (2008). *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC Press.

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

Betts, M. J. & Russell, R. B. (2003). *Bioinformatics for Geneticists*, chapter Amino acid propensities and consequences of substitutions. John Wiley & Sons, Chichester, UK.

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).

Boareto, M.; Yamagishi, M. E. B.; Caticha, N. & Leite, V. B. P. (2012). Relationship between global structural parameters and Enzyme Commission hierarchy: Implications for function prediction. *Comput. Biol. Chem.*, 40:15--19.

Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol*, 283(4):707--25.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pp. 31--40.

Bray, T.; Doig, A. J. & Warwicker, J. (2009). Sequence and structural features of enzymes and their active sites by EC class. *J. Mol. Biol*, 386(5):1423--1436.

Brown, S. D.; Gerlt, J. A.; Seffernick, J. L. & Babbitt, P. C. (2006). A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.*, 7(1):R8.

Capra, J. a. J. A. & Singh, M. (2008). Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, 24(13):1473--1480.

Casari, G.; Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nat. Struct. Biol*, 2(2):171--178.

Chitale, M.; Hawkins, T.; Park, C. & Kihara, D. (2009). ESG: extended similarity group method for automated protein function prediction. *Bioinformatics*, 25(14):1739–1745.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357:543--544.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22--9.

Clark, W. & Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. *Proteins*, 79:2086—96.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

Crooks, G. E.; Hon, H.; Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.*, 14:1188--90.

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

Deng, M.; Chen, T. & Sun, F. (2004). An integrated probabilistic model for functional prediction of proteins. *J. Comput. Biol.*, 11(2-3):463–475.

Deng, M.; Zhang, K.; Mehta, S.; Chen, T. & Sun, F. (2003). Prediction of protein function using protein-protein interaction data. *J Comput Biol.*, 10:947--60.

Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins*, 41(1):98--107.

Dobson, P. D. & Doig, A. J. (2005). Predicting enzyme class from protein structure without alignments. *J. Mol. Biol*, 345(1):187--199.

Duan, Z.-H.; Hughes, B.; Reichel, L.; Perez, D. M. & Shi, T. (2006). The relationship between protein sequences and their gene ontology functions. *BMC Bioinformatics*, 7(Suppl 4):S11.

Enault, F.; Suhre, K. & Claverie, J. (2005). Phydbac "gene function predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics*, 6(247).

Engelhardt, B.; Jordan, M.; Muratore, K. & Brenner, S. (2005). Protein molecular function prediction by bayesian phylogenomics. *PLoS Comput. Biol.*, 1(e45).

Erdin, S.; Lisewski, A. M. & Lichtarge, O. (2011). Protein function prediction: towards integration of similarity metrics. *Curr. Opin. Struct. Biol.*, 21(2):180–188.

Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M.; Eramian, D.; yi Shen, M.; Pieper, U. & Sali, A. (2006). Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, Suppl. 15:5.6.1--5.6.30.

Finn, R.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.; Eddy, S.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E.; Tate, J. & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res.*, 42(Database issue):D222–30.

Finn, R.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.; Gavin, O.; Gunesekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E.; Eddy, S. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res.*, 38(Database issue):211--222.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Mach. Learn*, 2:139--172.

Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends in Genetics*, 16(5):227--31.

Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C. & Jensen, L. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*

Friedberg, I. (2006). Automated protein function prediction – the genomic challenge. *Brief. Bioinform.*, 7(3):225--242.

Furnham, N.; Sillitoe, I.; Holliday, G. L.; Cuff, A. L.; Rahman, S. A.; Laskowski, R. A.; Orengo, C. A. & Thornton, J. M. (2012). FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.*, 40(D1):D776--D782.

Galperin, M. Y. & Koonin, E. V. (2010). From complete genome sequence to 'complete' understanding? *Trends Biotechnol.*, 28(8):398–406.

Gao, M. & Skolnick, J. (2013). A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.*, 9(10).

Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D. & Bairoch, A. (2005). Protein identification and analysis tools in the ExPASy server. In Walker, J. M., editor, *The Proteomics Protocols Handbook*, chapter 52, pp. 571--607. Humana Press.

Gaudet, P.; Livstone, M.; Lewis, S. & Thomas, P. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.*, 12:449—62.

Gherardini, P. F. & Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. *Brief. Funct. Genomic Proteomic*, 7(4):291–302.

Guilloux, V. L.; Schmidtke, P. & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168.

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, 2nd edição.

Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol*, 303:61--76.

Hawkins, T.; Luban, S. & Kihara, D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by pfp. *Protein Sci.*, 15:1550—6.

Hedstrom, L. (2002). Serine protease mechanism and specificity. *Chem Rev.*, 102:4501--23.

Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T. K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; de Castro, E.; Coggill, P.; Corbett, M.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Fraser, M.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; McMenamin, C.; Mi, H.; Mutowo-Muellenet, P.; Mulder, N.; Natale, D.; Orengo, C.; Pesseat, S.; Punta, M.; Quinn, A. F.; Rivoire, C.; Sangrador-Vegas, A.; Selengut, J. D.; Sigrist, C. J. A.; Scheremetjew, M.; Tate, J.; Thimmajanarthanan, M.; Thomas, P. D.; Wu, C. H.; Yeats, C. & Yong, S.-Y. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, 40:D306--D312.

Huttenhower, C.; Hibbs, M.; Myers, C. & Troyanskaya, O. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22:2890—7.

Huynen, M. (2000). Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.*, 10(8):1204--1210.

Jain, R. K. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley.

Jensen, L. J.; Gupta, R.; Blom, N.; Devos, D.; Tamames, J.; Kesmir, C.; Nielsen, H.; Stærfeldt, H. H.; Rapacki, K.; Workman, C.; Andersen, C. A. F.; Knudsen, S.; Krogh, A.; Valencia, A. & Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, 319:1257—65.

Kawabata, T. (2010). Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1195--211.

Kinoshita, K. & Nakamura, H. (2003). Protein informatics towards function identification. *Curr. Opin. Struct. Biol.*, 13:396–400.

Kolesov, G.; Mewes, H.-W. & Frishman, D. (2002). Snapper: gene order predicts gene function. *Bioinformatics*, 18(7):1017--9.

Kumar, C. & Choudhary, A. (2012). A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP J Bioinform Syst Biol*, 2012(1).

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5):323--30.

Laskowski, R. A.; Watson, J. D. & Thornton, J. M. (2005a). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33(suppl 2):W89--W93.

Laskowski, R. a.; Watson, J. D. & Thornton, J. M. (2005b). Protein function prediction using local 3d templates. *J. Mol. Biol*, 351(3):614--26.

Lee, D.; Redfern, O. & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, 8:995–10005.

Letovsky, S. & Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(Suppl 1):i197--204.

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

Luke, S. & Spector, L. (1998). A revised comparison of crossover and mutation in genetic programming. In Koza, J. R., editor, *Proceedings of the Third Annual conference in Genetic Programming*, pp. 208–213.

Manning, C. D.; Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marcotte, E.; Pellegrini, M.; Ng, H.; Rice, D.; Yeates, T. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751—3.

Martin, D. M. a.; Berriman, M. & Barton, G. J. (2004). Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178.

Melo-Minardi, R. C.; Bastard, K. & Artiguenave, F. (2010). Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics*, 26(24):3075--3082.

Mesa, A. d. S.; Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol*, 326:1289--1302.

Mitchell, A.; Chang, H.-Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; Sangrador-Vegas, A.; Scheremetjew, M.; Rato, C.; Yong, S.-Y.; Bateman, A.; Punta, M.; Attwood, T. K.; Sigrist, C. J.; Redaschi, N.; Rivoire, C.; Xenarios, I.; Kahn, D.; Guyot, D.; Bork, P.; Letunic, I.; Gough, J.; Oates, M.; Haft, D.; Huang, H.; Natale, D. A.; Wu, C. H.; Orengo, C.; Sillitoe, I.; Mi, H.; Thomas, P. D. & Finn, R. D. (2015). The interpro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, 43(D1):D213--21.

Moult, J. (2005). A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struc. Biol*, 15:285--289.

Munroe, D. R. (2004). Genetic programming: the ratio of crossover to mutation as a function of time. *Res. Lett. Inf. Math. Sci.*, 6:83–96.

Murray, R. K.; Granner, D. K.; Mayes, P. A. & Rodwell, V. W. (2003). *Harper's Illustrated Biochemistry*. Lange Medical Books/McGraw-Hill, 26 edição.

Nabieva, E.; Jim, K.; Agarwal, A.; Chazelle, B. & Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl 1):i302--10.

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

Neitzel, J. J. (2010). Enzyme catalysis: the serine proteases. *Nature Education*, 3(9):21.

Nelson, D. & Cox, M. (2005). *Lehninger Principles of Biochemistry*. W. H. Freeman, New York, 4 edição.

Norin, M. & Sundström, M. (2002). Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol.*, 20(2):79–84.

Pages, H.; Aboyoung, P.; Gentleman, R. & DebRoy, S. (2012). *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.22.0.

Pal, D. & Eisenberg, D. (2005). Inference of protein function from protein structure. *Structure*, 13(1):121--30.

Pazos, F.; Rausell, A. & Valencia, A. (2006). Phylogeny-independent detection of functional residues. *Bioinformatics*, 22(12):1440--8.

Pazos, F. & Sternberg, M. J. E. (2004). Automated prediction of protein function and detection of functional sites from structure. *P. Natl. Acad. Sci. USA*, 101(41):14754--9.

Pellegrini, M.; Marcotte, E.; Thompson, M.; Eisenberg, D. & Yeates, T. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96:4285—8.

Petretti, C. & Prigent, C. (2005). The Protein Kinase Resource: everything you always wanted to know about protein kkinase but were afraid to ask. *Biol. Cell*, 97:113--8.

Pires, D. E. V.; de Melo-Minardi, R. C.; Santos, M. A.; Silveira, C. H.; Santoro, M. M. & Meira Jr, W. (2011). Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(S4):S12.

Poli, R.; Langdon, W. B.; McPhee, N. F.; Systems, E.; Sciences, M. & Koza, J. R. (2008). *A Field Guide to Genetic Programming*. Freely available at http:/www.gp-field-guide.org.uk.

Pravda, L.; Berka, K.; Vařeková, R. S.; Sehnal, D.; Banáš, P.; Laskowski, R. A.; Koča, J. & Otyepka, M. (2014). Anatomy of enzyme channels. *BMC Bioinformatics*, 15:379.

Punta, M.; Coggill, P.; Eberhardt, R.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E.; Eddy, S.; Bateman, A. & Finn, R. (2012). The pgam protein families database. *Nucleic Acids Res.*, 40(Database issue):D290–301.

Punta, M. & Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput. Biol*, 4(10):e1000160.

Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; Pandey, G.; Yunes, J. M.; Talwalkar, A. S.; Repo, S.; Souza, M. L.; Piovesan, D.; Casadio, R.; Wang, Z.; Cheng, J.; Fang, H.; Gough, J.; Koskinen, P.; Törönen, P.; Nokso-Koivisto, J.; Holm, L.; Cozzetto, D.; Buchan, D. W. A.; Bryson, K.; Jones, D. T.; Limaye, B.; Inamdar, H.; Datta, A.; Manjari, S. K.; Joshi, R.; Chitale, M.; Kihara, D.; Lisewski, A. M.; Erdin, S.; Venner, E.; Lichtarge, O.; Rentzsch, R.; Yang, H.; Romero, A. E.; Bhat, P.; Paccanaro, A.; Hamp, T.; Kaßner, R.; Seemayer, S.; Vicedo, E.; Schaefer, C.; Achten, D.; Auer, F.; Boehm, A.; Braun, T.; Hecht, M.; Heron, M.; Hönigschmid, P.; Hopf, T. A.; Kaufmann, S.; Kiening, M.; Krompass, D.; Landerer, C.; Mahlich, Y.; Roos, M.; Björne, J.; Salakoski, T.; Wong, A.; Shatkay, H.; Gatzmann, F.; Sommer, I.; Wass, M. N.; Sternberg, M. J. E.; Škunca, N.; Supek, F.; Bošnjak, M.; Panov, P.; Džeroski, S.; Šmuc, T.; Kourmpetis, Y. A. I.; van Dijk, A. D. J.; ter Braak, C. J. F.; Zhou, Y.; Gong, Q.; Dong, X.; Tian, W.; Falda, M.; Fontana, P.; Lavezzo, E.; Di Camillo, B.; Toppo, S.; Lan, L.; Djuric, N.; Guo, Y.; Vucetic, S.; Bairoch, A.; Linial, M.; Babbitt, P. C.; Brenner, S. E.; Orengo, C.; Rost, B.; Mooney, S. D. & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nat. Methods*, 10(3):221--227.

Redfern, O. C.; Dessailly, B. H.; Dallman, T. J.; Sillitoe, I. & Orengo, C. A. (2009). Flora: a novel method to predict protein function from structure in diverse superfamilies. *PLoS Comput. Biol.*, 5(8).

Rentzsch, R. & Orengo, C. a. (2009). Protein function prediction–the power of multiplicity. *Trends Biotechnol.*, 27(4):210--9.

Rice, P.; Longden, I. & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6):276--277.

Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O. & Ofran, Y. (2003). Automatic prediction of protein function. *Cell. Mol. Life Sci*, 60(12):2637--50.

Roy, A.; Yang, J. & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucl. Acids Res.*, 40(W1):W471–W477.

Sadowski, M. I. & Jones, D. T. (2009). The sequence-structure relationship and protein function prediction. *Curr. Opin. Struc. Biol*, 19(3):357--362.

Salem, S. & Zaki, M. J. (2009). Iterative non-sequential protein structural alignment. *J. Bioinf. Comput. Biol*, 7(3):183--194.

Schmitt, S.; Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, 323:387–406.

Shah, I. & Hunter, L. (1997). Predicting enzyme function from sequence: a systematic appraisal. *Int. Conf. Intell. Syst. Mol. Biol*, 5:276--283.

Sharan, R.; Ulitsky, I. & Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.*, 3(88).

Shatsky, M.; Nussinov, R. & Wolfson, H. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins: Struct. Funct. Bioinf.*, 56(1):143–156.

Sigrist, C.; Cerutti, L.; de Castro, E.; Langendijk-Genevaux, P.; Bulliard, V.; Bairoch, A. & Hulo, N. (2010). Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, 38(Database issue):161--166.

Smith, A. A. T. (2012). *Automatically exploiting genomic and metabolic contexts to aid the functional annotation of prokaryote genomes*. PhD thesis, Université d'Évry Val d'Essone.

Smith, C. M.; Shindyalov, I. N.; Veretnik, S.; Gribskov, M.; Taylor, S. S.; Eyck, L. F. T. & Bourne, P. E. (1997). The protein kinase resource. *Trends Biochem. Sci.*, 22(11):444--6.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197.

Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; & von Mering, C. (2015). String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(D1):D447–52.

Tetko, I.; Rodchenkov, I.; Walter, M.; Rattei, T. & Mewes, H. (2008). Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information. *Bioinformatics*, 24(5):621–628.

The Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43(D1):D1049--56.

Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N. & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nat. Struct. Biol.*, 7:991–994.

Tian, W.; Arakaki, A. K. & Skolnick, J. (2004). Eficaz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, 32(21):6226--6239.

Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, 333(4):863--882.

Tramontano, A.; Leplae, R. & Morea, V. (2001). Analysis and assessment of comparative modeling predictions in casp4. *Proteins*, Suppl. 5(Jan.):22--38.

Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in casp5. *Proteins*, 53:352--368.

Tucker, C. L.; Hurley, J. H.; Miller, T. R. & Hurley, J. B. (1998). Two amino acid ssubstitution convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proc. Natl. Acad. Sci. USA*, 95(11):5993--7.

Vazquez, A.; Flammini, A.; Maritan, A. & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol.*, 21:697—700.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395--416.

von Mering, C.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P. & Snel, B. (2003). String: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, 31(1):258--261.

von Mering, C.; Jensen, L. J.; Snel, B.; Hooper, S. D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M. A. & Bork, P. (2005). String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, 33(S1):D433–7.

Wass, M. N. & Sternberg, M. J. E. (2008). Confunc-functional annotation in the twilight zone. *Bioinformatics*, 24(6):798--806.

Watson, J. D.; Laskowski, R. a. & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Curr. Opin. Struc. Biol*, 15(3):275--84.

White, R. H. (2006). The difficult road from sequence to function. *J. Bacteriol.*, 188(10):3431--3432.

Xin, F. & Radivojac, P. (2011). Computational methods for identification of functional residues in protein structures. *Curr. Protein Pept. Sci.*, 12:456—69.

Yu, G.-X.; Park, B.-H.; Chandramohan, P.; Munavalli, R.; Geist, A. & Samatova, N. F. (2005). In silico discovery of enzyme-substrate specificity-determining residue clusters. *J. Mol. Biol*, 352:1105--1117.

Zaha, A.; Ferreira, H. B.; Passaglia, L. M. P.; de Vasconcelos, A. T. R.; Schrank, A.; de Almeida, D. F.; Rossetti, M. L. R.; Loreto, E.; Vainstein, M. H.; Schrank, I. S.; da Silva, S. C.; de Castro, L. A. & Gaiesky, V. L. S. V. (2012). *Biologia Molecular Básica*. Artmed, Porto Alegre, RS, Brazil, 4 edição.

Zaki, M. J. & Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zhang, C. & Kim, S.-H. (2003). Overvew of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.*, 7:28–32.

Zhang, Y. & Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res.*, 33(7):2302--2309.

Zhao, S.; Kumar, R.; Sakai, A.; Vetting, M. W.; Wood, B. M.; Brown, S.; Bonanno, J. B.; Hillerich, B. S.; Seidel, R. D.; Babbitt, P. C.; Almo, S. C.; Sweedler, J. V.; Gerlt, J. A.; Cronan, J. E. & Jacobson, M. P. (2013). Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*, 502:698–802.

Zongker, D. & Punch, B. (1996). lil-gp 1.01 user's manual. Technical report, Michigan State University.

Apêndice A
Informação Mútua dos Agrupamentos Gerados pelo Sistema de GP

Neste trabalho, foram testadas muitas métricas de qualidade para os agrupamentos, visando a encontrar uma cujos altos valores correspondessem a agrupamentos considerados interessantes para o cenário de aplicação desta tese. Essas métricas eram baseadas em várias comumente aplicadas na avaliação de agrupamentos como, por exemplo, coeficiente de silhueta, BetaCV e *Normalized Cut*. Também foram estudadas métricas baseadas na entropia dos grupos e na verossimilhança (*log-likelihood*). No entanto, os agrupamentos que apresentaram os maiores valores para essas métricas não foram considerados bons para todas as famílias estudadas. Por isso, e por restrições de espaço, esses resultados não são apresentados nesta tese.

Foi um longo processo até chegar ao cálculo de informação mútua empregado nesta tese. Quando empregada na mineração de dados, a informação mútua é utilizada para comparar dois agrupamentos diferentes do mesmo conjunto de objetos, o que não é interessante para o cenário de aplicação desta tese. A aplicação feita neste trabalho é a mesma utilizada na área de recuperação de informação, em que a informação mútua é utilizada para avaliar a relevância de um dado termo para definir um documento de texto.

Neste apêndice são apresentados os valores de informação mútua (MI) do melhor indivíduo encontrado em cada execução do sistema de programação genética (GP), calculados conforme apresentado no Capítulo 5. Para as famílias Nucleotidil Ciclases, Proteínas Cinases e Serino Proteases, os resultados correspondem aos obtidos usando as famílias atualizadas em relação às originais empregadas por Melo-Minardi et al. (2010), ou seja, foram removidas as proteínas excluídas do UniProt desde então. As tabelas contêm as seguintes colunas:

1. **Parâmetros:** corresponde às taxas dos operadores de cruzamento, reprodução e mutação do sistema de GP, no formato *cc_rr_mm*. O valor 70_10_20, por exemplo, significa que foram usadas as taxas de 70% de cruzamento, 10% de reprodução e 20% de mutação.

2. **Repetição:** valor da semente utilizada para o gerador de números aleatórios e que marca também o número da repetição. A fim de que o experimento possa ser repetido, é preciso fixar o valor dessa semente.

3. **Grafo com Positivos:** indica resultados obtidos construindo o grafo de similaridades utilizado pelo agrupamento espectral apenas com valores positivos da matriz de similaridades gerada pelo sistema de GP.

4. **Grafo com Todos:** indica resultados obtidos construindo o grafo de similaridades utilizado pelo agrupamento espectral com todos os valores da matriz de similaridades calculada pelo sistema de GP, redimensionando-os para o intervalo [0, 1].

5. **Indivíduo:** identificador do indivíduo que levou ao melhor agrupamento.

6. **MI:** valor de informação mútua do agrupamento em questão.

7. **Diferença:** corresponde à diferença entre os valores de MI obtidos empregando cada método de construção do grafo de similaridades. Quanto menor a diferença, menor o efeito de escolher uma ou outra forma de construir o grafo de similaridades a partir da matriz de similaridades calculada pelo sistema de GP.

A.1 Estudo de Caso I: Nucleotidil Ciclases

Como mencionado no Capítulo 5, essa família de proteínas contém duas subfamílias: Adenilato Ciclases e Guanilato Ciclases. Com os mesmos parâmetros usados por Melo-Minardi et al. (2010) para o conjunto de proteínas original, o ASMC produz um agrupamento hierárquico cujo primeiro nível divide essa família em três grupos, e cujo segundo nível a divide em seis. Por essas razões, o sistema de GP foi executado com dois a seis grupos, e os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados, respectivamente, nas Tabelas A.1 a A.5.

Tabela A.1: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Nucleotidil Ciclases em dois grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	134	25,7782	122	25,7782	0,0000
70_10_20	2	92	25,7541	94	25,7541	0,0000
70_10_20	3	154	25,7782	138	25,7541	0,0241
70_10_20	4	23	25,7782	183	25,7782	0,0000
70_10_20	5	69	25,7782	24	25,7541	0,0241
70_20_10	1	81	25,7782	182	25,7782	0,0000
70_20_10	2	98	25,7782	98	25,7782	0,0000
70_20_10	3	97	25,7541	176	25,7782	-0,0242
70_20_10	4	23	25,7782	182	25,7782	0,0000
70_20_10	5	269	25,7782	24	25,7541	0,0241
80_05_15	1	112	25,7782	144	25,7782	0,0000

Tabela A.1: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_05_15	2	192	25,7782	191	25,7782	0,0000
80_05_15	3	78	25,7782	81	25,7541	0,0241
80_05_15	4	23	25,7782	65	25,7782	0,0000
80_05_15	5	132	25,7541	24	25,7541	0,0000
80_15_05	1	84	25,7782	167	25,7782	0,0000
80_15_05	2	180	25,7782	99	25,7782	0,0000
80_15_05	3	78	25,7782	203	25,7782	0,0000
80_15_05	4	23	25,7782	90	25,7782	0,0000
80_15_05	5	236	25,7782	270	25,7782	0,0000
80_20_00	1	63	25,7782	135	25,7782	0,0000
80_20_00	2	113	25,7541	268	25,7782	-0,0241
80_20_00	3	78	25,7782	81	25,7541	0,0241
80_20_00	4	23	25,7782	139	25,7782	0,0000
80_20_00	5	116	25,7541	24	25,7541	0,0000
85_05_10	1	73	25,7782	94	25,7782	0,0000
85_05_10	2	61	25,7782	124	25,7541	0,0241
85_05_10	3	278	25,7782	81	25,7541	0,0241
85_05_10	4	23	25,7782	90	25,7782	0,0000
85_05_10	5	293	25,7782	24	25,7541	0,0241
85_10_05	1	84	25,7782	109	25,7782	0,0000
85_10_05	2	134	25,7782	233	25,7782	0,0000
85_10_05	3	269	25,7782	274	25,7782	0,0000
85_10_05	4	23	25,7782	90	25,7782	0,0000
85_10_05	5	137	25,7541	170	25,7782	-0,0241
90_05_05	1	139	25,7782	109	25,7782	0,0000
90_05_05	2	134	25,7782	141	25,7541	0,0241
90_05_05	3	178	25,7782	167	25,7782	0,0000
90_05_05	4	23	25,7782	90	25,7782	0,0000
90_05_05	5	130	25,7541	24	25,7541	0,0000
90_10_00	1	86	25,7782	142	25,7782	0,0000
90_10_00	2	94	25,7541	110	25,7782	-0,0241
90_10_00	3	67	25,7541	192	25,7782	-0,0242
90_10_00	4	23	25,7782	104	25,7782	0,0000
90_10_00	5	263	25,7782	24	25,7541	0,0241

Tabela A.2: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Nucleotidil Ciclases em três grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	210	22,4108	10	22,2276	0,1832
70_10_20	2	215	22,2177	29	22,0781	0,1396
70_10_20	3	51	22,2161	40	22,2276	-0,0115
70_10_20	4	166	22,2161	8	22,2276	-0,0115
70_10_20	5	95	22,0829	95	22,2276	-0,1447
70_20_10	1	119	22,4232	10	22,2276	0,1956
70_20_10	2	211	22,2177	29	22,0781	0,1396
70_20_10	3	190	22,3641	40	22,2276	0,1365
70_20_10	4	94	22,2276	8	22,2276	0,0000
70_20_10	5	95	22,0829	102	22,2276	-0,1447
80_05_15	1	63	22,3483	10	22,2276	0,1207
80_05_15	2	158	22,2177	157	22,2276	-0,0100
80_05_15	3	116	22,2276	40	22,2276	0,0000
80_05_15	4	154	22,2276	8	22,2276	0,0000
80_05_15	5	117	22,2177	65	22,2276	-0,0100
80_15_05	1	109	22,2161	10	22,2276	-0,0115
80_15_05	2	87	22,2177	83	22,2276	-0,0100
80_15_05	3	51	22,2161	40	22,2276	-0,0115
80_15_05	4	247	22,1253	8	22,2276	-0,1023
80_15_05	5	270	22,3641	135	22,2276	0,1365
80_20_00	1	109	22,2161	10	22,2276	-0,0115
80_20_00	2	87	22,2177	83	22,2276	-0,0100
80_20_00	3	51	22,2161	40	22,2276	-0,0115
80_20_00	4	227	22,2161	8	22,2276	-0,0115
80_20_00	5	256	22,3641	136	22,2276	0,1365
85_05_10	1	193	22,2161	10	22,2276	-0,0115
85_05_10	2	127	22,2177	236	22,2276	-0,0100
85_05_10	3	124	22,2276	40	22,2276	0,0000
85_05_10	4	221	22,2276	8	22,2276	0,0000
85_05_10	5	246	22,0829	214	22,2276	-0,1447
85_10_05	1	224	22,1420	10	22,2276	-0,0856
85_10_05	2	248	22,1253	94	22,2276	-0,1023
85_10_05	3	124	22,2276	40	22,2276	0,0000
85_10_05	4	110	22,0951	8	22,2276	-0,1326
85_10_05	5	273	22,3641	126	22,2276	0,1365
90_05_05	1	190	22,4232	10	22,2276	0,1956
90_05_05	2	74	22,2161	29	22,0781	0,1380
90_05_05	3	298	22,2276	40	22,2276	0,0000

Tabela A.2: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_05_05	4	71	22,2276	8	22,2276	0,0000
90_05_05	5	125	22,2177	104	22,2276	-0,0100
90_10_00	1	280	22,1420	10	22,2276	-0,0856
90_10_00	2	87	22,2177	83	22,2276	-0,0100
90_10_00	3	51	22,2161	40	22,2276	-0,0115
90_10_00	4	242	22,3641	8	22,2276	0,1365
90_10_00	5	125	22,2177	15	22,0951	0,1226

Tabela A.3: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Nucleotidil Ciclases em quatro grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	17	21,0468	229	20,8817	0,1651
70_10_20	2	115	21,1197	258	21,0777	0,0420
70_10_20	3	32	21,0468	169	19,5328	1,5141
70_10_20	4	17	21,0468	279	17,9621	3,0848
70_10_20	5	15	21,0468	279	18,1240	2,9228
70_20_10	1	17	21,0468	10	16,3157	4,7311
70_20_10	2	182	21,1197	270	20,9332	0,1865
70_20_10	3	32	21,0468	32	18,6456	2,4013
70_20_10	4	17	21,0468	17	17,4162	3,6306
70_20_10	5	15	21,0468	238	18,1240	2,9228
80_05_15	1	17	21,0468	283	16,5546	4,4922
80_05_15	2	187	21,1197	206	21,0813	0,0384
80_05_15	3	32	21,0468	32	18,6456	2,4013
80_05_15	4	17	21,0468	225	20,8872	0,1596
80_05_15	5	15	21,0468	170	21,0800	-0,0332
80_15_05	1	17	21,0468	141	16,5546	4,4922
80_15_05	2	152	21,1197	268	20,8872	0,2324
80_15_05	3	32	21,0468	32	18,6456	2,4013
80_15_05	4	17	21,0468	17	17,4162	3,6306
80_15_05	5	15	21,0468	263	18,0745	2,9723
80_20_00	1	17	21,0468	141	16,5546	4,4922
80_20_00	2	29	21,0468	230	21,0465	0,0003
80_20_00	3	32	21,0468	32	18,6456	2,4013
80_20_00	4	17	21,0468	17	17,4162	3,6306
80_20_00	5	15	21,0468	188	17,9965	3,0503
85_05_10	1	195	21,0787	10	16,3157	4,7631

Tabela A.3: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	2	161	21,0468	222	20,9332	0,1137
85_05_10	3	32	21,0468	32	18,6456	2,4013
85_05_10	4	17	21,0468	17	17,4162	3,6306
85_05_10	5	15	21,0468	119	17,9965	3,0503
85_10_05	1	153	21,1540	140	16,5546	4,5993
85_10_05	2	168	21,1197	238	21,0813	0,0384
85_10_05	3	32	21,0468	32	18,6456	2,4013
85_10_05	4	17	21,0468	252	21,0465	0,0003
85_10_05	5	15	21,0468	279	18,0303	3,0165
90_05_05	1	158	21,0777	277	19,6490	1,4287
90_05_05	2	156	21,1540	289	20,8872	0,2667
90_05_05	3	32	21,0468	32	18,6456	2,4013
90_05_05	4	17	21,0468	17	17,4162	3,6306
90_05_05	5	15	21,0468	130	17,9965	3,0503
90_10_00	1	17	21,0468	79	16,5546	4,4922
90_10_00	2	29	21,0468	227	21,0465	0,0003
90_10_00	3	32	21,0468	32	18,6456	2,4013
90_10_00	4	17	21,0468	279	20,7750	0,2719
90_10_00	5	15	21,0468	237	17,8577	3,1891

Tabela A.4: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Nucleotidil Ciclases em cinco grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	64	18,1782	258	18,1296	0,0486
70_10_20	2	86	17,8929	29	17,8229	0,0700
70_10_20	3	166	18,1782	166	18,1782	0,0000
70_10_20	4	151	18,1782	61	18,1182	0,0600
70_10_20	5	222	16,8849	283	16,7114	0,1735
70_20_10	1	17	17,7791	17	17,8229	-0,0438
70_20_10	2	86	17,8929	105	17,8750	0,0179
70_20_10	3	227	18,1782	161	18,1182	0,0600
70_20_10	4	254	18,1658	17	17,8229	0,3429
70_20_10	5	259	17,8231	207	16,5248	1,2984
80_05_15	1	64	18,1782	271	18,1424	0,0358
80_05_15	2	86	17,8929	138	17,8750	0,0179
80_05_15	3	180	18,1782	40	17,3801	0,7980
80_05_15	4	71	18,1782	61	18,1182	0,0600

Tabela A.4: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_05_15	5	292	17,5956	149	17,3801	0,2154
80_15_05	1	267	18,0779	17	17,8229	0,2549
80_15_05	2	90	18,1321	29	17,8229	0,3091
80_15_05	3	191	18,1782	240	17,9638	0,2144
80_15_05	4	102	18,1020	17	17,8229	0,2791
80_15_05	5	284	17,7377	263	17,3801	0,3576
80_20_00	1	111	18,1782	186	18,1182	0,0600
80_20_00	2	90	18,1321	29	17,8229	0,3091
80_20_00	3	104	18,1782	40	17,3801	0,7980
80_20_00	4	140	18,1321	17	17,8229	0,3091
80_20_00	5	260	17,8975	271	17,5859	0,3116
85_05_10	1	285	18,0844	256	18,1424	-0,0580
85_05_10	2	86	17,8929	153	18,0702	-0,1772
85_05_10	3	71	18,1782	40	17,3801	0,7980
85_05_10	4	236	18,1782	17	17,8229	0,3553
85_05_10	5	278	17,1228	249	16,3684	0,7544
85_10_05	1	259	18,1020	211	18,1424	-0,0404
85_10_05	2	195	18,1424	29	17,8229	0,3195
85_10_05	3	71	18,1782	243	18,1180	0,0601
85_10_05	4	144	18,1180	184	18,0840	0,0340
85_10_05	5	192	17,9283	273	17,5859	0,3424
90_05_05	1	235	17,8255	233	18,1296	-0,3041
90_05_05	2	281	18,1451	29	17,8229	0,3222
90_05_05	3	71	18,1782	107	18,1182	0,0600
90_05_05	4	262	18,1296	17	17,8229	0,3066
90_05_05	5	192	17,9283	224	17,7402	0,1881
90_10_00	1	219	18,1782	17	17,8229	0,3553
90_10_00	2	90	18,1321	29	17,8229	0,3091
90_10_00	3	71	18,1782	107	18,1182	0,0600
90_10_00	4	255	17,6885	17	17,8229	-0,1345
90_10_00	5	265	17,8244	232	17,3801	0,4443

Tabela A.5: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Nucleotidil Ciclases em seis grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	211	16,1195	264	16,1080	0,0114
70_10_20	2	205	16,0963	243	16,1124	-0,0161

Tabela A.5: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	3	267	16,1316	126	16,0963	0,0353
70_10_20	4	169	16,1204	236	15,5422	0,5782
70_10_20	5	242	15,9896	220	16,0963	-0,1067
70_20_10	1	230	16,1316	183	16,1080	0,0236
70_20_10	2	191	16,1138	218	16,0963	0,0175
70_20_10	3	203	16,1341	266	16,1273	0,0067
70_20_10	4	267	16,1014	192	16,1204	-0,0191
70_20_10	5	178	15,2173	220	16,0963	-0,8790
80_05_15	1	286	16,1273	108	16,1080	0,0193
80_05_15	2	263	16,1138	228	16,0963	0,0175
80_05_15	3	293	16,1341	98	16,0963	0,0377
80_05_15	4	212	16,1273	151	16,1028	0,0245
80_05_15	5	275	16,1273	264	16,1205	0,0068
80_15_05	1	255	16,0963	216	16,1180	-0,0217
80_15_05	2	245	16,1148	213	16,1243	-0,0095
80_15_05	3	238	16,1341	232	16,1014	0,0327
80_15_05	4	179	15,7321	232	16,1243	-0,3922
80_15_05	5	249	16,0276	252	16,0584	-0,0308
80_20_00	1	255	16,1541	102	16,1080	0,0461
80_20_00	2	251	16,1273	277	16,1243	0,0031
80_20_00	3	264	16,1316	267	16,1014	0,0303
80_20_00	4	217	16,1273	227	16,1204	0,0068
80_20_00	5	192	15,9350	221	16,0963	-0,1614
85_05_10	1	256	16,1341	215	16,1036	0,0305
85_05_10	2	260	16,0963	219	16,1138	-0,0175
85_05_10	3	245	16,1316	285	16,1014	0,0303
85_05_10	4	141	16,1273	122	16,1028	0,0245
85_05_10	5	232	16,1036	262	16,1316	-0,0281
85_10_05	1	162	16,1316	133	16,1080	0,0236
85_10_05	2	289	16,1273	148	16,1138	0,0135
85_10_05	3	277	16,1316	203	16,1014	0,0303
85_10_05	4	141	16,1273	158	16,1204	0,0068
85_10_05	5	256	15,9810	235	16,1080	-0,1270
90_05_05	1	226	16,1273	192	16,1036	0,0238
90_05_05	2	265	16,1138	239	16,0963	0,0175
90_05_05	3	286	16,1341	191	16,0963	0,0377
90_05_05	4	180	16,0802	168	16,1028	-0,0226
90_05_05	5	251	16,1273	274	16,1273	0,0000

Tabela A.5: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	1	248	16,1273	215	16,1080	0,0193
90_10_00	2	141	16,0963	265	16,1138	-0,0175
90_10_00	3	263	16,1316	190	16,0963	0,0353
90_10_00	4	177	15,8968	214	15,8600	0,0369
90_10_00	5	288	15,9721	263	16,0963	-0,1242

A.2 Estudo de Caso II: Serino Proteases

Como mencionado no Capítulo 5, essa família de proteínas contém três subfamílias: Tripsinas, Quimotripsinas e Elastases, além de uma subfamília das Tripsinas encontradas por Melo-Minardi et al. (2010) que corresponde a Calicreínas. Após eliminar do conjunto de proteínas usado por esses autores aquelas que foram excluídas do UniProt desde então, o ASMC, quando utilizado com os mesmos parâmetros do algoritmo Cobweb empregados por eles (-A 1,0 e -C 0,25), não divide essa família. Então, o principal parâmetro (-C) foi reduzido de 0,05 em 0,05 até encontrar um valor que dividisse a família. Esse valor foi de 0,15, que produziu um agrupamento hierárquico cujo primeiro nível divide a família em quatro grupos, e cujo segundo nível a divide em onze grupos. Por essas razões, e considerando a variabilidade da família em relação à composição dos sítios ativos putativos, o sistema de GP foi executado com várias quantidades de grupos: de três a treze grupos. Os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados nas Tabelas A.6 a A.16.

Tabela A.6: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em três grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	17	16,3207	77	13,2626	3,0581
70_10_20	2	29	16,3207	147	13,3723	2,9483
70_10_20	3	32	16,3207	258	13,3414	2,9792
70_10_20	4	17	16,3207	172	13,3300	2,9907
70_10_20	5	15	16,3207	212	13,0908	3,2298
70_20_10	1	17	16,3207	134	13,2626	3,0581
70_20_10	2	29	16,3207	186	13,4283	2,8924
70_20_10	3	32	16,3207	238	13,5952	2,7254
70_20_10	4	17	16,3207	17	11,5624	4,7583
70_20_10	5	15	16,3207	237	13,1917	3,1290
80_05_15	1	17	16,3207	196	13,3303	2,9903
80_05_15	2	29	16,3207	146	13,5109	2,8098
80_05_15	3	32	16,3207	286	13,3424	2,9783
80_05_15	4	17	16,3207	17	11,5624	4,7583
80_05_15	5	15	16,3207	277	13,2125	3,1081
80_15_05	1	17	16,3207	138	13,2626	3,0581
80_15_05	2	29	16,3207	114	13,3723	2,9483
80_15_05	3	32	16,3207	256	13,5952	2,7254
80_15_05	4	17	16,3207	225	13,4613	2,8593
80_15_05	5	15	16,3207	262	12,8815	3,4392
80_20_00	1	17	16,3207	213	13,2699	3,0508
80_20_00	2	29	16,3207	242	13,5971	2,7235
80_20_00	3	32	16,3207	211	13,2753	3,0454

Tabela A.6: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_20_00	4	17	16,3207	17	11,5624	4,7583
80_20_00	5	15	16,3207	244	12,9032	3,4175
85_05_10	1	17	16,3207	267	13,3223	2,9984
85_05_10	2	29	16,3207	236	13,5952	2,7254
85_05_10	3	217	16,5232	202	13,5952	2,9280
85_05_10	4	17	16,3207	231	13,2753	3,0454
85_05_10	5	15	16,3207	203	13,2414	3,0792
85_10_05	1	17	16,3207	267	13,4121	2,9086
85_10_05	2	29	16,3207	61	12,4704	3,8502
85_10_05	3	32	16,3207	279	13,5952	2,7254
85_10_05	4	17	16,3207	182	13,2753	3,0454
85_10_05	5	15	16,3207	244	13,1207	3,2000
90_05_05	1	196	16,3207	247	13,3164	3,0042
90_05_05	2	29	16,3207	61	12,4704	3,8502
90_05_05	3	32	16,3207	286	13,3747	2,9459
90_05_05	4	17	16,3207	170	13,2753	3,0454
90_05_05	5	15	16,3207	277	13,4672	2,8535
90_10_00	1	17	16,3207	116	13,2626	3,0581
90_10_00	2	29	16,3207	133	12,4831	3,8376
90_10_00	3	32	16,3207	283	13,3414	2,9792
90_10_00	4	17	16,3207	17	11,5624	4,7583
90_10_00	5	15	16,3207	231	13,3723	2,9483

Tabela A.7: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em quatro grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	222	17,8037	17	16,0867	1,7170
70_10_20	2	241	17,8533	29	16,0867	1,7666
70_10_20	3	184	17,5642	199	17,5364	0,0278
70_10_20	4	17	16,1090	17	16,0867	0,0223
70_10_20	5	102	13,8851	15	16,0691	-2,1840
70_20_10	1	106	17,7094	17	16,0867	1,6227
70_20_10	2	141	17,6544	187	16,2248	1,4295
70_20_10	3	118	17,5642	121	17,5364	0,0278
70_20_10	4	181	17,7156	17	16,0867	1,6289
70_20_10	5	269	13,8507	15	16,0691	-2,2184
80_05_15	1	111	17,7094	17	16,0867	1,6227

Tabela A.7: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_05_15	2	148	17,6544	237	16,2248	1,4295
80_05_15	3	32	16,1063	32	16,0691	0,0372
80_05_15	4	120	17,5701	237	17,5446	0,0255
80_05_15	5	291	13,8851	15	16,0691	-2,1840
80_15_05	1	211	17,5701	17	16,0867	1,4834
80_15_05	2	135	17,7184	196	17,6072	0,1112
80_15_05	3	234	17,7660	32	16,0691	1,6969
80_15_05	4	17	16,1090	17	16,0867	0,0223
80_15_05	5	167	13,4801	15	16,0691	-2,5889
80_20_00	1	225	17,5701	17	16,0867	1,4834
80_20_00	2	200	17,8351	277	17,6072	0,2279
80_20_00	3	202	17,7660	32	16,0691	1,6969
80_20_00	4	219	17,7156	17	16,0867	1,6289
80_20_00	5	278	13,7223	15	16,0691	-2,3468
85_05_10	1	121	17,7094	17	16,0867	1,6227
85_05_10	2	242	17,5701	29	16,0867	1,4834
85_05_10	3	257	17,8351	32	16,0691	1,7660
85_05_10	4	188	17,7156	17	16,0867	1,6289
85_05_10	5	285	13,6395	15	16,0691	-2,4295
85_10_05	1	148	17,2979	17	16,0867	1,2111
85_10_05	2	242	17,8533	277	16,2248	1,6285
85_10_05	3	286	17,8351	32	16,0691	1,7660
85_10_05	4	17	16,1090	17	16,0867	0,0223
85_10_05	5	276	13,7072	15	16,0691	-2,3619
90_05_05	1	289	17,6843	17	16,0867	1,5976
90_05_05	2	110	17,6544	295	16,2248	1,4295
90_05_05	3	245	17,7660	32	16,0691	1,6969
90_05_05	4	17	16,1090	130	17,6072	-1,4982
90_05_05	5	229	13,8851	15	16,0691	-2,1840
90_10_00	1	146	17,6065	224	16,2214	1,3851
90_10_00	2	110	17,6544	192	16,2248	1,4295
90_10_00	3	272	17,8343	32	16,0691	1,7652
90_10_00	4	238	17,7094	17	16,0867	1,6227
90_10_00	5	209	13,6937	15	16,0691	-2,3754

Tabela A.8: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em cinco grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	206	15,8094	201	16,1940	-0,3846
70_10_20	2	185	16,6328	29	14,2942	2,3387
70_10_20	3	164	16,0559	117	14,1508	1,9050
70_10_20	4	17	14,8254	17	14,8889	-0,0636
70_10_20	5	246	16,2707	236	16,2814	-0,0107
70_20_10	1	143	15,8094	17	14,2942	1,5152
70_20_10	2	257	16,8174	246	16,5545	0,2629
70_20_10	3	220	16,5830	166	16,1449	0,4381
70_20_10	4	109	14,8761	221	16,3019	-1,4258
70_20_10	5	204	14,7443	273	15,9973	-1,2530
80_05_15	1	269	16,2930	17	14,2942	1,9988
80_05_15	2	234	16,6945	29	14,2942	2,4004
80_05_15	3	32	14,1930	113	16,1825	-1,9895
80_05_15	4	103	15,4465	17	14,8889	0,5576
80_05_15	5	191	16,3114	285	16,2418	0,0696
80_15_05	1	254	16,7539	17	14,2942	2,4597
80_15_05	2	124	16,6623	90	16,6684	-0,0061
80_15_05	3	168	16,6209	189	16,1449	0,4760
80_15_05	4	17	14,8254	17	14,8889	-0,0636
80_15_05	5	198	15,3593	173	16,0557	-0,6963
80_20_00	1	213	15,8094	195	14,6994	1,1100
80_20_00	2	90	16,6320	90	16,6684	-0,0363
80_20_00	3	214	16,5489	132	14,1508	2,3980
80_20_00	4	17	14,8254	17	14,8889	-0,0636
80_20_00	5	225	15,3593	15	14,1212	1,2381
85_05_10	1	116	16,7045	152	16,1683	0,5362
85_05_10	2	289	16,7102	29	14,2942	2,4160
85_05_10	3	253	16,7125	177	16,1449	0,5676
85_05_10	4	131	16,5355	17	14,8889	1,6466
85_05_10	5	258	16,5881	234	16,0823	0,5059
85_10_05	1	116	16,7045	192	16,1683	0,5362
85_10_05	2	282	16,7102	195	16,5545	0,1556
85_10_05	3	277	16,7778	278	16,7554	0,0224
85_10_05	4	139	15,4465	17	14,8889	0,5576
85_10_05	5	168	16,5881	160	16,0647	0,5234
90_05_05	1	294	16,7552	206	16,1940	0,5612
90_05_05	2	157	14,3240	29	14,2942	0,0299
90_05_05	3	250	16,6115	274	16,5823	0,0292

Tabela A.8: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_05_05	4	217	16,6339	17	14,8889	1,7450
90_05_05	5	221	16,5881	260	16,0909	0,4972
90_10_00	1	232	16,6530	17	14,2942	2,3588
90_10_00	2	153	16,7490	90	16,6684	0,0806
90_10_00	3	245	16,7295	168	16,1358	0,5938
90_10_00	4	217	16,5728	17	14,8889	1,6839
90_10_00	5	175	16,5881	270	15,8757	0,7124

Tabela A.9: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em seis grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	259	15,3117	280	14,9576	0,3540
70_10_20	2	274	15,4987	174	15,2638	0,2348
70_10_20	3	32	13,6827	32	13,7010	-0,0183
70_10_20	4	17	14,3891	17	14,3685	0,0206
70_10_20	5	205	14,9471	249	14,9160	0,0311
70_20_10	1	154	14,6810	240	14,8115	-0,1305
70_20_10	2	206	15,4987	163	14,7848	0,7138
70_20_10	3	32	13,6827	32	13,7010	-0,0183
70_20_10	4	17	14,3891	17	14,3685	0,0206
70_20_10	5	171	14,8073	15	14,4342	0,3731
80_05_15	1	281	14,8688	287	14,7415	0,1273
80_05_15	2	156	15,4987	283	15,2638	0,2348
80_05_15	3	32	13,6827	199	13,7423	-0,0596
80_05_15	4	234	15,3091	287	14,3685	0,9406
80_05_15	5	215	15,0088	172	14,9767	0,0321
80_15_05	1	232	15,4059	247	14,8993	0,5066
80_15_05	2	178	15,4987	29	14,3715	1,1271
80_15_05	3	32	13,6827	32	13,7010	-0,0183
80_15_05	4	17	14,3891	17	14,3685	0,0206
80_15_05	5	209	14,9471	227	14,8663	0,0807
80_20_00	1	269	15,4731	150	14,6870	0,7860
80_20_00	2	204	15,1539	226	15,4986	-0,3447
80_20_00	3	32	13,6827	32	13,7010	-0,0183
80_20_00	4	17	14,3891	17	14,3685	0,0206
80_20_00	5	249	15,3307	148	15,0400	0,2907
85_05_10	1	173	14,9940	283	14,6880	0,3060

Tabela A.9: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	2	111	15,1242	253	14,6052	0,5190
85_05_10	3	32	13,6827	32	13,7010	-0,0183
85_05_10	4	288	15,4915	17	14,3685	1,1230
85_05_10	5	293	14,9237	15	14,4342	0,4895
85_10_05	1	170	14,9940	203	14,7923	0,2017
85_10_05	2	29	14,4936	29	14,3715	0,1221
85_10_05	3	32	13,6827	32	13,7010	-0,0183
85_10_05	4	253	15,3458	17	14,3685	0,9773
85_10_05	5	180	15,3307	259	14,9796	0,3511
90_05_05	1	257	15,1930	232	14,7923	0,4007
90_05_05	2	174	15,4987	151	15,4986	0,0000
90_05_05	3	32	13,6827	32	13,7010	-0,0183
90_05_05	4	275	15,4588	17	14,3685	1,0904
90_05_05	5	268	14,9877	261	14,8174	0,1703
90_10_00	1	203	15,0613	274	14,4233	0,6380
90_10_00	2	177	15,4987	29	14,3715	1,1271
90_10_00	3	32	13,6827	32	13,7010	-0,0183
90_10_00	4	17	14,3891	17	14,3685	0,0206
90_10_00	5	249	15,0875	15	14,4342	0,6533

Tabela A.10: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em sete grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	17	14,2924	275	15,0489	-0,7564
70_10_20	2	198	13,3399	189	13,2249	0,1150
70_10_20	3	99	13,2942	230	13,4063	-0,1120
70_10_20	4	17	13,2174	17	14,2533	-1,0358
70_10_20	5	15	14,3184	15	13,7290	0,5893
70_20_10	1	17	14,2924	200	14,9775	-0,6851
70_20_10	2	86	13,2572	29	13,0430	0,2142
70_20_10	3	40	13,1692	40	13,2468	-0,0776
70_20_10	4	207	15,0206	17	14,2533	0,7673
70_20_10	5	198	14,4298	188	14,3963	0,0336
80_05_15	1	17	14,2924	17	13,6990	0,5934
80_05_15	2	86	13,2572	29	13,0430	0,2142
80_05_15	3	289	13,2613	40	13,2468	0,0145
80_05_15	4	245	15,0206	241	15,1702	-0,1496

Tabela A.10: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_05_15	5	15	14,3184	15	13,7290	0,5893
80_15_05	1	17	14,2924	17	13,6990	0,5934
80_15_05	2	109	13,2572	29	13,0430	0,2142
80_15_05	3	225	14,5185	40	13,2468	1,2717
80_15_05	4	17	13,2174	17	14,2533	-1,0358
80_15_05	5	15	14,3184	15	13,7290	0,5893
80_20_00	1	17	14,2924	17	13,6990	0,5934
80_20_00	2	109	13,2572	29	13,0430	0,2142
80_20_00	3	142	13,9076	262	13,5902	0,3174
80_20_00	4	17	13,2174	17	14,2533	-1,0358
80_20_00	5	15	14,3184	15	13,7290	0,5893
85_05_10	1	17	14,2924	17	13,6990	0,5934
85_05_10	2	86	13,2572	29	13,0430	0,2142
85_05_10	3	260	13,2613	40	13,2468	0,0145
85_05_10	4	114	13,9513	17	14,2533	-0,3019
85_05_10	5	117	14,3644	15	13,7290	0,6354
85_10_05	1	17	14,2924	17	13,6990	0,5934
85_10_05	2	244	13,3399	29	13,0430	0,2969
85_10_05	3	225	13,3861	40	13,2468	0,1392
85_10_05	4	227	13,2404	17	14,2533	-1,0129
85_10_05	5	263	14,4298	15	13,7290	0,7008
90_05_05	1	17	14,2924	17	13,6990	0,5934
90_05_05	2	243	13,6125	29	13,0430	0,5695
90_05_05	3	270	13,2613	40	13,2468	0,0145
90_05_05	4	17	13,2174	17	14,2533	-1,0358
90_05_05	5	15	14,3184	232	14,3963	-0,0779
90_10_00	1	17	14,2924	17	13,6990	0,5934
90_10_00	2	200	13,6125	29	13,0430	0,5695
90_10_00	3	142	13,9076	40	13,2468	0,6608
90_10_00	4	17	13,2174	17	14,2533	-1,0358
90_10_00	5	255	14,4298	15	13,7290	0,7008

Tabela A.11: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em oito grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	249	13,2973	17	12,9550	0,3422
70_10_20	2	232	11,6709	257	11,9089	-0,2380

Tabela A.11: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	3	32	12,9074	32	12,9229	-0,0154
70_10_20	4	83	13,5362	17	12,9118	0,6244
70_10_20	5	241	13,0220	15	12,9118	0,1103
70_20_10	1	222	13,2217	17	12,9550	0,2667
70_20_10	2	269	11,6795	263	11,9089	-0,2293
70_20_10	3	32	12,9074	32	12,9229	-0,0154
70_20_10	4	17	13,5207	17	12,9118	0,6089
70_20_10	5	271	14,1617	275	13,4811	0,6807
80_05_15	1	118	13,1243	252	13,5633	-0,4390
80_05_15	2	269	11,6795	18	11,4269	0,2527
80_05_15	3	32	12,9074	32	12,9229	-0,0154
80_05_15	4	17	13,5207	17	12,9118	0,6089
80_05_15	5	224	14,1617	136	13,0115	1,1502
80_15_05	1	225	13,2602	17	12,9550	0,3052
80_15_05	2	41	11,5546	191	12,0637	-0,5091
80_15_05	3	32	12,9074	32	12,9229	-0,0154
80_15_05	4	17	13,5207	17	12,9118	0,6089
80_15_05	5	15	12,9367	121	13,0115	-0,0748
80_20_00	1	83	13,2217	17	12,9550	0,2667
80_20_00	2	41	11,5546	202	12,6497	-1,0951
80_20_00	3	32	12,9074	32	12,9229	-0,0154
80_20_00	4	17	13,5207	17	12,9118	0,6089
80_20_00	5	15	12,9367	113	13,0115	-0,0748
85_05_10	1	239	13,2602	17	12,9550	0,3052
85_05_10	2	262	11,6709	286	11,9089	-0,2380
85_05_10	3	32	12,9074	32	12,9229	-0,0154
85_05_10	4	17	13,5207	17	12,9118	0,6089
85_05_10	5	15	12,9367	264	13,2049	-0,2682
85_10_05	1	239	13,2602	17	12,9550	0,3052
85_10_05	2	41	11,5546	101	12,7004	-1,1458
85_10_05	3	32	12,9074	32	12,9229	-0,0154
85_10_05	4	17	13,5207	17	12,9118	0,6089
85_10_05	5	234	13,0220	221	13,2574	-0,2354
90_05_05	1	210	13,2602	17	12,9550	0,3052
90_05_05	2	41	11,5546	292	12,8969	-1,3423
90_05_05	3	32	12,9074	32	12,9229	-0,0154
90_05_05	4	17	13,5207	17	12,9118	0,6089
90_05_05	5	15	12,9367	15	12,9118	0,0249

Tabela A.11: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	1	190	13,1554	17	12,9550	0,2003
90_10_00	2	41	11,5546	101	12,7004	-1,1458
90_10_00	3	32	12,9074	32	12,9229	-0,0154
90_10_00	4	17	13,5207	17	12,9118	0,6089
90_10_00	5	15	12,9367	125	13,0115	-0,0748

Tabela A.12: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em nove grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	287	13,4029	198	12,0453	1,3576
70_10_20	2	163	12,4110	256	11,6266	0,7844
70_10_20	3	168	12,4759	244	12,4953	-0,0194
70_10_20	4	202	13,2554	17	11,6280	1,6275
70_10_20	5	276	11,6656	15	11,4556	0,2100
70_20_10	1	253	12,1832	151	11,9807	0,2024
70_20_10	2	133	12,4060	179	11,6353	0,7708
70_20_10	3	32	11,7299	32	11,6876	0,0423
70_20_10	4	267	12,4813	94	11,7493	0,7320
70_20_10	5	204	13,3461	251	12,0986	1,2475
80_05_15	1	173	13,4029	17	11,5811	1,8219
80_05_15	2	241	12,4110	281	11,7486	0,6624
80_05_15	3	273	12,7268	145	12,4953	0,2315
80_05_15	4	244	12,4813	17	11,6280	0,8533
80_05_15	5	195	12,6626	15	11,4556	1,2070
80_15_05	1	184	12,1688	208	12,0453	0,1235
80_15_05	2	238	11,8544	223	12,0156	-0,1613
80_15_05	3	218	12,4702	246	11,9374	0,5328
80_15_05	4	8	11,8620	141	12,1854	-0,3234
80_15_05	5	251	12,6775	15	11,4556	1,2219
80_20_00	1	155	13,4029	256	12,0453	1,3576
80_20_00	2	158	12,4060	229	12,0059	0,4001
80_20_00	3	277	12,4702	168	12,3493	0,1209
80_20_00	4	215	12,4813	17	11,6280	0,8533
80_20_00	5	256	12,9645	280	12,2466	0,7179
85_05_10	1	210	12,1637	228	11,9807	0,1829
85_05_10	2	148	12,4060	189	11,6266	0,7794
85_05_10	3	32	11,7299	244	11,9677	-0,2378

Tabela A.12: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	4	8	11,8620	17	11,6280	0,2340
85_05_10	5	262	11,5216	15	11,4556	0,0659
85_10_05	1	275	12,1957	158	13,4305	-1,2348
85_10_05	2	29	10,5892	250	11,6266	-1,0374
85_10_05	3	32	11,7299	32	11,6876	0,0423
85_10_05	4	8	11,8620	232	12,1854	-0,3234
85_10_05	5	265	11,5596	15	11,4556	0,1040
90_05_05	1	212	13,4029	234	12,4506	0,9523
90_05_05	2	29	10,5892	29	10,5860	0,0033
90_05_05	3	231	12,0018	32	11,6876	0,3141
90_05_05	4	214	12,4813	253	11,7510	0,7303
90_05_05	5	267	12,5557	141	11,5565	0,9993
90_10_00	1	186	13,4029	216	12,6739	0,7290
90_10_00	2	261	11,2805	214	11,6266	-0,3462
90_10_00	3	250	12,4655	287	11,9374	0,5280
90_10_00	4	8	11,8620	203	12,1854	-0,3234
90_10_00	5	278	12,5501	141	11,5565	0,9936

Tabela A.13: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em dez grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	255	11,5365	272	11,9529	-0,4163
70_10_20	2	234	12,1167	139	12,1213	-0,0046
70_10_20	3	215	11,4169	122	11,6841	-0,2673
70_10_20	4	262	12,4478	234	12,4832	-0,0354
70_10_20	5	123	11,5354	260	12,3334	-0,7980
70_20_10	1	63	12,5680	162	12,1552	0,4127
70_20_10	2	142	10,9995	69	11,4633	-0,4638
70_20_10	3	240	12,5813	40	11,0219	1,5594
70_20_10	4	267	12,4988	175	12,6445	-0,1457
70_20_10	5	223	11,7297	152	12,6373	-0,9076
80_05_15	1	149	12,5358	292	12,3301	0,2058
80_05_15	2	263	12,0558	146	12,3165	-0,2607
80_05_15	3	270	12,6555	118	11,4043	1,2512
80_05_15	4	294	12,3060	171	10,9852	1,3208
80_05_15	5	246	12,5650	192	11,7597	0,8054
80_15_05	1	63	12,5680	231	12,5330	0,0350

Tabela A.13: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_15_05	2	241	12,5383	119	12,1213	0,4170
80_15_05	3	162	11,9792	267	11,3178	0,6614
80_15_05	4	215	12,5161	96	11,2780	1,2381
80_15_05	5	80	11,0484	223	11,4132	-0,3649
80_20_00	1	254	12,6580	232	12,1552	0,5027
80_20_00	2	135	10,9987	69	11,4633	-0,4646
80_20_00	3	32	10,8268	40	11,0219	-0,1950
80_20_00	4	185	11,3418	231	12,4832	-1,1414
80_20_00	5	122	11,5354	169	11,4132	0,1222
85_05_10	1	63	12,5680	260	12,6144	-0,0464
85_05_10	2	256	10,9995	100	10,0462	0,9533
85_05_10	3	108	11,7518	277	11,6583	0,0935
85_05_10	4	234	12,5747	127	12,4872	0,0875
85_05_10	5	211	12,1569	247	12,3344	-0,1775
85_10_05	1	63	12,5680	161	12,5330	0,0350
85_10_05	2	237	11,2510	153	11,5014	-0,2504
85_10_05	3	225	11,9792	287	12,3926	-0,4134
85_10_05	4	221	12,5161	96	11,2780	1,2381
85_10_05	5	80	11,0484	169	11,4570	-0,4086
90_05_05	1	216	12,5684	246	12,5330	0,0354
90_05_05	2	169	12,5790	137	12,1213	0,4577
90_05_05	3	293	11,7334	40	11,0219	0,7115
90_05_05	4	134	12,5691	255	12,5604	0,0087
90_05_05	5	80	11,0484	128	11,4570	-0,4086
90_10_00	1	168	12,6993	270	12,5330	0,1663
90_10_00	2	209	10,9987	200	12,3472	-1,3485
90_10_00	3	32	10,8268	141	11,2097	-0,3828
90_10_00	4	88	11,3418	242	12,6445	-1,3027
90_10_00	5	233	11,5431	261	12,2350	-0,6919

Tabela A.14: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em onze grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	163	11,8605	138	11,2046	0,6559
70_10_20	2	233	11,5305	29	10,6556	0,8749
70_10_20	3	210	10,9983	40	10,9025	0,0958
70_10_20	4	197	11,8265	17	11,2644	0,5621

Tabela A.14: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	5	135	11,8437	252	11,8432	0,0005
70_20_10	1	10	11,5110	140	11,2046	0,3064
70_20_10	2	265	11,5305	29	10,6556	0,8749
70_20_10	3	110	11,1309	40	10,9025	0,2284
70_20_10	4	252	11,8265	17	11,2644	0,5621
70_20_10	5	248	11,9697	192	11,7837	0,1859
80_05_15	1	243	12,0891	133	11,2046	0,8845
80_05_15	2	268	10,8520	29	10,6556	0,1963
80_05_15	3	125	11,1309	40	10,9025	0,2284
80_05_15	4	17	11,3230	17	11,2644	0,0586
80_05_15	5	148	11,9727	270	11,8234	0,1493
80_15_05	1	10	11,5110	103	11,2046	0,3064
80_15_05	2	87	11,0820	125	11,7242	-0,6422
80_15_05	3	40	10,2737	40	10,9025	-0,6288
80_15_05	4	17	11,3230	17	11,2644	0,0586
80_15_05	5	241	11,7153	140	11,4526	0,2627
80_20_00	1	10	11,5110	17	11,0393	0,4717
80_20_00	2	87	11,0820	249	11,7524	-0,6704
80_20_00	3	128	11,1309	40	10,9025	0,2284
80_20_00	4	17	11,3230	17	11,2644	0,0586
80_20_00	5	251	11,3166	231	11,5857	-0,2691
85_05_10	1	10	11,5110	138	11,2046	0,3064
85_05_10	2	271	11,8600	29	10,6556	1,2043
85_05_10	3	223	11,4356	40	10,9025	0,5331
85_05_10	4	17	11,3230	17	11,2644	0,0586
85_05_10	5	265	11,8761	280	11,7637	0,1123
85_10_05	1	10	11,5110	128	11,4406	0,0705
85_10_05	2	87	10,6629	29	10,6556	0,0073
85_10_05	3	110	11,1507	40	10,9025	0,2482
85_10_05	4	17	11,3230	17	11,2644	0,0586
85_10_05	5	279	11,6609	176	11,7421	-0,0812
90_05_05	1	10	11,5110	132	11,2046	0,3064
90_05_05	2	87	10,6629	29	10,6556	0,0073
90_05_05	3	110	11,1507	40	10,9025	0,2482
90_05_05	4	247	11,8361	17	11,2644	0,5718
90_05_05	5	206	11,8761	151	11,6679	0,2082
90_10_00	1	10	11,5110	17	11,0393	0,4717
90_10_00	2	87	10,6629	141	10,8721	-0,2092

Tabela A.14: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	3	40	10,2737	40	10,9025	-0,6288
90_10_00	4	17	11,3230	17	11,2644	0,0586
90_10_00	5	209	11,8555	206	11,7837	0,0718

Tabela A.15: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em doze grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	280	11,3348	202	11,1580	0,1768
70_10_20	2	171	10,5592	29	10,4847	0,0745
70_10_20	3	40	10,7440	180	10,9399	-0,1959
70_10_20	4	91	11,0090	91	10,9940	0,0150
70_10_20	5	261	11,6845	108	10,4934	1,1911
70_20_10	1	17	10,6734	147	11,0149	-0,3415
70_20_10	2	173	11,2264	212	11,3942	-0,1678
70_20_10	3	40	10,7440	198	10,8163	-0,0723
70_20_10	4	175	10,9940	175	10,9940	0,0000
70_20_10	5	155	11,0997	277	11,4691	-0,3693
80_05_15	1	292	11,4672	288	11,3254	0,1418
80_05_15	2	214	10,7339	271	11,0456	-0,3116
80_05_15	3	212	10,8992	238	10,9399	-0,0407
80_05_15	4	8	10,3784	149	10,9940	-0,6156
80_05_15	5	284	11,4285	195	10,4934	0,9351
80_15_05	1	17	10,6734	211	11,0152	-0,3418
80_15_05	2	256	11,2216	249	10,8674	0,3543
80_15_05	3	40	10,7440	152	10,9399	-0,1959
80_15_05	4	123	10,9940	123	10,9940	0,0000
80_15_05	5	264	11,3837	112	11,3306	0,0531
80_20_00	1	166	11,4060	196	11,2575	0,1485
80_20_00	2	248	11,1601	203	11,1167	0,0433
80_20_00	3	40	10,7440	181	10,9399	-0,1959
80_20_00	4	79	10,9940	79	10,9940	0,0000
80_20_00	5	224	11,3422	184	11,3306	0,0116
85_05_10	1	240	11,5038	284	11,0152	0,4885
85_05_10	2	89	10,5935	279	11,3942	-0,8006
85_05_10	3	270	11,1439	132	10,9399	0,2040
85_05_10	4	123	11,0090	119	10,9940	0,0150
85_05_10	5	274	11,3549	114	11,3306	0,0243

Tabela A.15: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_10_05	1	275	11,6060	247	11,4481	0,1580
85_10_05	2	155	11,2216	29	10,4847	0,7369
85_10_05	3	104	10,9987	132	10,9399	0,0588
85_10_05	4	8	10,3784	134	10,9940	-0,6156
85_10_05	5	282	11,3541	114	11,3306	0,0235
90_05_05	1	172	11,2728	280	11,0152	0,2575
90_05_05	2	244	11,2216	29	10,4847	0,7369
90_05_05	3	104	10,9987	132	10,9399	0,0588
90_05_05	4	187	11,0198	274	11,3491	-0,3293
90_05_05	5	281	11,5906	293	11,3433	0,2474
90_10_00	1	211	11,4060	283	11,0684	0,3376
90_10_00	2	151	11,1116	252	10,9791	0,1325
90_10_00	3	40	10,7440	159	10,9399	-0,1959
90_10_00	4	96	10,9940	96	10,9940	0,0000
90_10_00	5	252	11,4754	293	11,6968	-0,2214

Tabela A.16: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Serino Proteases em treze grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	17	11,1250	273	11,2042	-0,0792
70_10_20	2	29	10,0437	201	11,0483	-1,0046
70_10_20	3	228	10,5186	226	10,3375	0,1812
70_10_20	4	258	10,7371	17	10,1323	0,6047
70_10_20	5	244	10,8143	235	10,7913	0,0230
70_20_10	1	17	11,1250	100	11,0728	0,0522
70_20_10	2	29	10,0437	264	10,6610	-0,6173
70_20_10	3	189	10,4488	236	10,5806	-0,1318
70_20_10	4	17	10,1297	265	10,4490	-0,3193
70_20_10	5	138	10,8967	180	10,9670	-0,0703
80_05_15	1	234	11,1628	17	10,8239	0,3390
80_05_15	2	29	10,0437	195	10,3218	-0,2781
80_05_15	3	276	10,3947	266	10,3714	0,0233
80_05_15	4	121	10,3387	17	10,1323	0,2064
80_05_15	5	204	10,8967	273	11,1300	-0,2333
80_15_05	1	163	11,1628	17	10,8239	0,3390
80_15_05	2	250	10,2012	29	10,2063	-0,0051
80_15_05	3	275	10,8934	211	10,7012	0,1922

Tabela A.16: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_15_05	4	17	10,1297	17	10,1323	-0,0026
80_15_05	5	141	10,8143	277	11,1300	-0,3157
80_20_00	1	17	11,1250	17	10,8239	0,3011
80_20_00	2	214	10,2012	29	10,2063	-0,0051
80_20_00	3	186	10,3457	105	10,3860	-0,0402
80_20_00	4	17	10,1297	17	10,1323	-0,0026
80_20_00	5	191	10,8143	179	11,1235	-0,3092
85_05_10	1	17	11,1250	17	10,8239	0,3011
85_05_10	2	194	10,5361	133	10,2622	0,2739
85_05_10	3	266	10,5786	220	10,3714	0,2072
85_05_10	4	275	10,3929	17	10,1323	0,2605
85_05_10	5	269	10,8926	151	10,7581	0,1345
85_10_05	1	17	11,1250	17	10,8239	0,3011
85_10_05	2	29	10,0437	29	10,2063	-0,1626
85_10_05	3	266	10,5786	268	10,8020	-0,2234
85_10_05	4	17	10,1297	17	10,1323	-0,0026
85_10_05	5	233	10,8143	281	11,1300	-0,3157
90_05_05	1	17	11,1250	17	10,8239	0,3011
90_05_05	2	74	10,5361	171	10,3218	0,2143
90_05_05	3	280	10,3457	275	10,7453	-0,3996
90_05_05	4	207	10,4380	17	10,1323	0,3057
90_05_05	5	279	10,8143	198	10,7306	0,0836
90_10_00	1	159	11,2282	17	10,8239	0,4043
90_10_00	2	147	10,2012	29	10,2063	-0,0051
90_10_00	3	252	10,8934	105	10,3860	0,5074
90_10_00	4	96	10,3387	17	10,1323	0,2064
90_10_00	5	237	10,7845	139	11,2319	-0,4475

A.3 Estudo de Caso III: Proteínas Cinases

Como mencionado no Capítulo 5, essa família de proteínas contém duas subfamílias: Serina/Treonina Cinases e Tirosina Cinases. Melo-Minardi et al. (2010) reportaram também um subgrupo das Tirosina Cinases rotulados como EGFRs, do inglês *Epidermal Growth Factor Receptor*. Para o conjunto atualizado de proteínas, o ASMC (Melo-Minardi et al., 2010) produz, com os mesmos parâmetros usados pelos autores para o conjunto original, um agrupamento hierárquico cujo primeiro nível divide essa família em três grupos, e cujo segundo nível a divide em sete. Por essas razões, o sistema de GP foi executado com dois a sete grupos. Os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados nas Tabelas A.17 a A.22.

Tabela A.17: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em dois grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	243	109,7282	245	109,5696	0,1587
70_10_20	2	4	109,2974	4	109,2974	0,0000
70_10_20	3	222	109,3256	247	109,0210	0,3046
70_10_20	4	15	109,2974	186	109,6239	-0,3265
70_10_20	5	135	109,8928	84	109,6227	0,2701
70_20_10	1	244	109,7282	87	109,2974	0,4308
70_20_10	2	4	109,2974	4	109,2974	0,0000
70_20_10	3	268	109,4410	270	109,8944	-0,4534
70_20_10	4	263	109,7164	15	109,2974	0,4190
70_20_10	5	135	109,8928	84	109,6227	0,2701
80_05_15	1	178	109,8944	270	109,8944	0,0000
80_05_15	2	4	109,2974	4	109,2974	0,0000
80_05_15	3	267	109,5824	245	109,3815	0,2009
80_05_15	4	239	109,5572	117	109,6614	-0,1041
80_05_15	5	284	109,8928	84	109,6227	0,2701
80_15_05	1	246	109,8944	253	109,9902	-0,0958
80_15_05	2	4	109,2974	4	109,2974	0,0000
80_15_05	3	252	109,6010	281	109,1808	0,4202
80_15_05	4	260	109,7164	15	109,2974	0,4190
80_15_05	5	158	109,8928	84	109,6227	0,2701
80_20_00	1	273	109,7282	243	109,3703	0,3579
80_20_00	2	4	109,2974	103	109,6227	-0,3253
80_20_00	3	244	109,8928	213	109,0198	0,8730
80_20_00	4	264	109,8928	177	109,6614	0,2314
80_20_00	5	191	109,8928	229	109,6614	0,2314
85_05_10	1	252	109,8944	229	109,8944	0,0000

Tabela A.17: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	2	4	109,2974	143	109,6227	-0,3253
85_05_10	3	296	109,7021	220	109,6614	0,0407
85_05_10	4	171	109,7282	15	109,2974	0,4308
85_05_10	5	175	109,8928	84	109,6227	0,2701
85_10_05	1	260	109,9003	205	109,8944	0,0059
85_10_05	2	101	109,8944	140	109,5574	0,3370
85_10_05	3	241	109,3930	230	109,6614	-0,2684
85_10_05	4	157	109,7282	184	109,6614	0,0668
85_10_05	5	112	109,7282	246	109,7282	0,0000
90_05_05	1	242	109,7649	101	109,8944	-0,1295
90_05_05	2	4	109,2974	4	109,2974	0,0000
90_05_05	3	255	109,8928	294	109,0719	0,8209
90_05_05	4	295	109,7282	229	109,6614	0,0668
90_05_05	5	237	109,8928	97	109,6227	0,2701
90_10_00	1	204	109,7282	101	109,8944	-0,1662
90_10_00	2	4	109,2974	145	109,5574	-0,2600
90_10_00	3	287	109,7282	200	109,4864	0,2418
90_10_00	4	284	109,8928	266	109,7282	0,1646
90_10_00	5	234	109,8928	132	109,6227	0,2701

Tabela A.18: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em três grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	218	102,3506	274	102,7412	-0,3906
70_10_20	2	159	102,9445	172	102,8691	0,0753
70_10_20	3	287	103,1394	192	102,8009	0,3385
70_10_20	4	105	102,9408	237	102,9373	0,0036
70_10_20	5	46	101,9447	251	102,6676	-0,7229
70_20_10	1	227	102,5444	227	102,7196	-0,1752
70_20_10	2	118	102,9408	267	102,8691	0,0717
70_20_10	3	253	102,8804	215	103,1629	-0,2825
70_20_10	4	250	102,9816	274	102,8691	0,1125
70_20_10	5	245	102,5468	75	102,7477	-0,2009
80_05_15	1	265	102,2702	276	102,7175	-0,4473
80_05_15	2	166	102,9408	206	102,9184	0,0225
80_05_15	3	226	102,8787	277	102,8833	-0,0046
80_05_15	4	167	102,9408	270	102,9184	0,0225

Tabela A.18: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_05_15	5	251	102,6222	266	102,7477	-0,1255
80_15_05	1	268	102,6539	263	102,7142	-0,0603
80_15_05	2	18	101,9447	260	102,9184	-0,9737
80_15_05	3	173	102,8804	267	102,8691	0,0112
80_15_05	4	247	102,9816	3	101,6886	1,2930
80_15_05	5	276	102,7329	75	102,7477	-0,0149
80_20_00	1	254	102,6832	259	102,4953	0,1879
80_20_00	2	227	102,9445	207	102,8691	0,0753
80_20_00	3	256	102,8804	192	102,8009	0,0794
80_20_00	4	118	102,9408	267	101,9039	1,0370
80_20_00	5	272	102,5438	75	102,7477	-0,2040
85_05_10	1	255	102,4351	288	102,4728	-0,0377
85_05_10	2	143	102,9408	276	102,9184	0,0225
85_05_10	3	256	102,9445	242	102,9219	0,0225
85_05_10	4	125	102,9408	250	102,9184	0,0225
85_05_10	5	269	102,8263	75	102,7477	0,0786
85_10_05	1	276	102,2992	247	102,6505	-0,3513
85_10_05	2	271	102,7230	188	102,0206	0,7024
85_10_05	3	209	102,8787	267	102,9214	-0,0427
85_10_05	4	191	102,9408	215	102,8691	0,0717
85_10_05	5	249	102,5708	243	102,7865	-0,2157
90_05_05	1	246	102,3885	291	102,7366	-0,3481
90_05_05	2	203	102,9408	231	102,8691	0,0717
90_05_05	3	112	102,8787	285	103,0458	-0,1671
90_05_05	4	237	103,1297	268	102,9837	0,1460
90_05_05	5	247	102,7531	264	102,9888	-0,2358
90_10_00	1	261	102,4485	216	102,7531	-0,3046
90_10_00	2	18	101,9447	222	102,8691	-0,9244
90_10_00	3	265	102,9162	146	103,1629	-0,2466
90_10_00	4	141	102,9408	259	102,9371	0,0037
90_10_00	5	46	101,9447	279	102,8836	-0,9389

Tabela A.19: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em quatro grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	271	77,4473	144	76,1640	1,2833
70_10_20	2	242	75,7914	259	75,6544	0,1370

Tabela A.19: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	3	249	77,4952	254	76,3708	1,1244
70_10_20	4	3	76,2402	3	77,3378	-1,0976
70_10_20	5	15	76,1779	253	76,3210	-0,1431
70_20_10	1	244	77,4299	238	76,5182	0,9117
70_20_10	2	260	75,8621	149	76,1389	-0,2768
70_20_10	3	32	76,1630	274	76,2802	-0,1173
70_20_10	4	222	77,5067	3	77,3378	0,1689
70_20_10	5	256	77,4866	237	76,1554	1,3312
80_05_15	1	179	77,4818	251	76,2454	1,2364
80_05_15	2	256	76,7872	284	75,6489	1,1382
80_05_15	3	197	77,4138	220	76,1640	1,2498
80_05_15	4	164	78,5358	232	78,2894	0,2464
80_05_15	5	240	77,3912	258	75,9737	1,4175
80_15_05	1	246	77,4811	271	76,3011	1,1800
80_15_05	2	155	77,1893	256	75,7823	1,4070
80_15_05	3	241	77,1346	278	76,2949	0,8397
80_15_05	4	152	79,2250	255	78,4779	0,7472
80_15_05	5	273	77,3415	249	75,9320	1,4094
80_20_00	1	238	77,4397	187	76,2315	1,2082
80_20_00	2	252	76,6716	210	76,1389	0,5327
80_20_00	3	157	77,2433	204	75,9184	1,3249
80_20_00	4	195	77,4216	3	77,3378	0,0838
80_20_00	5	257	77,3345	210	76,1465	1,1879
85_05_10	1	230	77,5244	283	76,2454	1,2789
85_05_10	2	264	76,6019	125	75,3479	1,2540
85_05_10	3	260	77,4055	271	76,2992	1,1063
85_05_10	4	207	78,5358	3	77,3378	1,1980
85_05_10	5	286	77,4479	287	76,0974	1,3506
85_10_05	1	210	77,5244	279	76,3625	1,1618
85_10_05	2	267	75,7308	267	76,1389	-0,4081
85_10_05	3	233	77,2792	269	76,1640	1,1151
85_10_05	4	231	79,2250	242	78,4597	0,7653
85_10_05	5	255	77,4647	267	76,1469	1,3177
90_05_05	1	269	77,5241	260	76,2532	1,2710
90_05_05	2	289	76,5300	269	75,6486	0,8814
90_05_05	3	203	77,4658	279	76,1466	1,3192
90_05_05	4	237	77,9436	241	78,2894	-0,3458
90_05_05	5	237	77,2587	251	76,0118	1,2469

Tabela A.19: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	1	269	77,5244	242	76,5251	0,9993
90_10_00	2	266	76,7820	214	75,8216	0,9604
90_10_00	3	253	77,4957	250	76,1921	1,3036
90_10_00	4	137	78,5358	249	78,3177	0,2181
90_10_00	5	250	77,4167	229	75,8864	1,5303

Tabela A.20: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em cinco grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	99	61,4858	221	62,2840	-0,7983
70_10_20	2	85	61,8051	248	60,9996	0,8055
70_10_20	3	232	61,8255	233	61,5764	0,2491
70_10_20	4	129	60,2682	260	60,8878	-0,6196
70_10_20	5	279	61,2579	15	60,4777	0,7803
70_20_10	1	176	61,8255	252	60,9848	0,8408
70_20_10	2	85	61,8051	257	60,9996	0,8055
70_20_10	3	261	61,8255	32	59,7171	2,1085
70_20_10	4	196	61,3458	17	59,6647	1,6811
70_20_10	5	177	62,1763	15	60,4777	1,6986
80_05_15	1	134	61,5328	217	62,0744	-0,5416
80_05_15	2	85	61,8051	289	61,0026	0,8025
80_05_15	3	145	61,1656	202	61,5764	-0,4108
80_05_15	4	235	61,4860	17	59,6647	1,8213
80_05_15	5	292	61,3470	270	61,0271	0,3198
80_15_05	1	264	61,5089	226	61,4326	0,0763
80_15_05	2	275	62,7687	29	60,4777	2,2910
80_15_05	3	216	61,8255	271	62,0855	-0,2600
80_15_05	4	68	60,4121	183	60,7336	-0,3215
80_15_05	5	193	61,2579	15	60,4777	0,7803
80_20_00	1	157	61,2967	217	61,4326	-0,1359
80_20_00	2	249	62,6656	29	60,4777	2,1879
80_20_00	3	216	61,8255	261	62,0855	-0,2600
80_20_00	4	17	60,2653	17	59,6647	0,6006
80_20_00	5	185	61,2579	15	60,4777	0,7803
85_05_10	1	194	61,9238	240	61,0475	0,8763
85_05_10	2	85	61,8051	241	61,0026	0,8025
85_05_10	3	32	60,0625	259	62,0048	-1,9423

Tabela A.20: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	4	17	60,2653	17	59,6647	0,6006
85_05_10	5	205	61,6284	84	62,3218	-0,6934
85_10_05	1	270	61,6284	242	61,0475	0,5808
85_10_05	2	29	60,6218	29	60,4777	0,1442
85_10_05	3	32	60,0625	230	61,5764	-1,5139
85_10_05	4	172	60,7805	241	60,7336	0,0469
85_10_05	5	284	61,2579	15	60,4777	0,7803
90_05_05	1	134	61,5328	148	62,2901	-0,7573
90_05_05	2	29	60,6218	29	60,4777	0,1442
90_05_05	3	32	60,0625	243	61,5764	-1,5139
90_05_05	4	71	62,3892	17	59,6647	2,7245
90_05_05	5	178	61,2579	15	60,4777	0,7803
90_10_00	1	191	61,3547	233	62,0855	-0,7308
90_10_00	2	226	61,1948	29	60,4777	0,7171
90_10_00	3	32	60,0625	172	61,5764	-1,5139
90_10_00	4	17	60,2653	17	59,6647	0,6006
90_10_00	5	257	61,4592	15	60,4777	0,9815

Tabela A.21: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em seis grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	226	53,5591	103	54,7189	-1,1597
70_10_20	2	18	53,9077	18	53,7061	0,2016
70_10_20	3	272	55,2030	271	56,5966	-1,3936
70_10_20	4	3	53,4959	274	52,1331	1,3628
70_10_20	5	277	58,2916	263	58,2180	0,0735
70_20_10	1	93	56,5434	65	56,5699	-0,0265
70_20_10	2	18	53,9077	18	53,7061	0,2016
70_20_10	3	121	54,6474	229	57,3382	-2,6909
70_20_10	4	155	54,1347	269	55,5082	-1,3736
70_20_10	5	234	58,2290	254	58,4269	-0,1979
80_05_15	1	173	53,4959	82	54,5061	-1,0102
80_05_15	2	18	53,9077	18	53,7061	0,2016
80_05_15	3	244	56,6274	63	56,5699	0,0575
80_05_15	4	235	54,1347	275	55,5082	-1,3736
80_05_15	5	252	58,1008	232	58,2795	-0,1787
80_15_05	1	243	52,1190	221	55,7768	-3,6578

Tabela A.21: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_15_05	2	18	53,9077	87	53,8133	0,0944
80_15_05	3	96	56,6274	280	56,8919	-0,2645
80_15_05	4	79	54,1347	204	55,4992	-1,3646
80_15_05	5	255	58,0406	269	58,1734	-0,1328
80_20_00	1	247	55,0550	244	56,7057	-1,6507
80_20_00	2	18	53,9077	87	53,8133	0,0944
80_20_00	3	243	56,8194	200	56,7100	0,1093
80_20_00	4	89	53,7793	218	52,2911	1,4882
80_20_00	5	265	58,1918	228	58,1891	0,0027
85_05_10	1	155	53,4959	286	56,7319	-3,2360
85_05_10	2	18	53,9077	18	53,7061	0,2016
85_05_10	3	290	55,6871	63	56,5699	-0,8829
85_05_10	4	89	54,1347	198	55,5082	-1,3736
85_05_10	5	258	58,1097	248	58,4592	-0,3495
85_10_05	1	203	53,5718	189	54,7189	-1,1470
85_10_05	2	156	54,9121	18	53,7061	1,2060
85_10_05	3	282	55,1926	63	56,5699	-1,3773
85_10_05	4	79	54,1347	279	55,2691	-1,1344
85_10_05	5	260	58,1005	206	58,1734	-0,0728
90_05_05	1	228	54,5430	124	54,7189	-0,1758
90_05_05	2	18	53,9077	18	53,7061	0,2016
90_05_05	3	103	56,6274	63	56,5699	0,0575
90_05_05	4	237	54,1347	237	53,8364	0,2983
90_05_05	5	252	56,6750	11	57,5695	-0,8945
90_10_00	1	115	53,1193	128	54,5061	-1,3868
90_10_00	2	18	53,9077	87	53,8133	0,0944
90_10_00	3	103	56,6274	258	56,6205	0,0069
90_10_00	4	3	53,4959	226	53,9870	-0,4910
90_10_00	5	234	56,6750	282	57,6769	-1,0018

Tabela A.22: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família Proteínas Cinases em sete grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	256	49,9020	176	50,4225	-0,5206
70_10_20	2	261	48,8298	18	48,3159	0,5138
70_10_20	3	236	51,9059	226	50,4563	1,4496
70_10_20	4	17	47,7466	17	47,4240	0,3226

Tabela A.22: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	5	149	50,4456	96	51,6303	-1,1848
70_20_10	1	137	49,2668	213	48,9539	0,3129
70_20_10	2	148	48,5169	18	48,3159	0,2009
70_20_10	3	63	49,7334	111	50,0601	-0,3267
70_20_10	4	120	48,2904	17	47,4240	0,8665
70_20_10	5	266	50,2809	169	51,4415	-1,1607
80_05_15	1	228	49,0882	212	49,0906	-0,0024
80_05_15	2	285	48,5075	18	48,3159	0,1916
80_05_15	3	282	50,6998	207	51,5856	-0,8858
80_05_15	4	245	48,9012	17	47,4240	1,4773
80_05_15	5	236	50,2063	246	50,3479	-0,1416
80_15_05	1	131	49,5268	236	50,6654	-1,1386
80_15_05	2	124	48,1744	175	48,5792	-0,4048
80_15_05	3	113	51,6869	276	50,1051	1,5818
80_15_05	4	157	48,2904	79	48,2885	0,0019
80_15_05	5	249	50,2828	231	51,6303	-1,3475
80_20_00	1	95	49,5268	141	48,9539	0,5730
80_20_00	2	205	48,8258	175	48,5792	0,2467
80_20_00	3	255	52,0297	215	51,5242	0,5055
80_20_00	4	171	48,0079	17	47,4240	0,5840
80_20_00	5	267	50,4456	205	51,6754	-1,2299
85_05_10	1	290	49,2668	281	48,8809	0,3859
85_05_10	2	29	47,7825	18	48,3159	-0,5335
85_05_10	3	262	52,2069	156	49,8595	2,3475
85_05_10	4	286	48,9638	17	47,4240	1,5398
85_05_10	5	281	50,5978	214	50,1263	0,4715
85_10_05	1	95	49,5268	86	48,2885	1,2383
85_10_05	2	178	48,1413	18	48,3159	-0,1747
85_10_05	3	160	51,6869	281	52,4204	-0,7335
85_10_05	4	17	47,7466	79	48,2885	-0,5419
85_10_05	5	267	50,2226	178	51,6303	-1,4077
90_05_05	1	95	49,5268	294	50,4225	-0,8957
90_05_05	2	286	49,2495	18	48,3159	0,9336
90_05_05	3	281	51,6869	192	52,4204	-0,7335
90_05_05	4	137	48,0428	265	48,2885	-0,2457
90_05_05	5	202	50,4582	141	51,6303	-1,1721
90_10_00	1	207	50,0807	160	50,4537	-0,3730
90_10_00	2	212	48,8258	18	48,3159	0,5099

Tabela A.22: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	3	222	51,6869	211	52,4204	-0,7335
90_10_00	4	17	47,7466	17	47,4240	0,3226
90_10_00	5	276	50,2226	286	51,4415	-1,2189

A.4 Estudo de Caso IV: DUF849

Essa família de proteínas de função desconhecida foi estudada por Bastard et al. (2014), que obtiveram 84 grupos com uma aplicação automática do ASMC (Melo-Minardi et al., 2010), sete grupos com uma manipulação manual do agrupamento hierárquico produzido pelo ASMC e 32 grupos com uma abordagem de *ensemble clustering* para integrar resultados de diferentes estratégias de agrupamento. Por isso, o sistema GP foi executado para produzir 7, 32 e 84 grupos para a família DUF849, e os valores de MI para os melhores agrupamentos encontrados em cada execução são apresentados, respectivamente, nas Tabelas A.23, A.24 e A.25.

Tabela A.23: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família DUF849 em sete grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	280	36,4800	268	36,4700	0,0100
70_10_20	2	255	36,3378	254	34,7473	1,5905
70_10_20	3	250	36,1348	274	36,5913	-0,4565
70_10_20	4	92	34,3354	55	33,7960	0,5394
70_10_20	5	256	35,4881	277	35,7840	-0,2959
70_20_10	1	229	36,3417	212	36,3562	-0,0145
70_20_10	2	247	36,3582	262	33,7253	2,6329
70_20_10	3	260	35,8160	56	36,2143	-0,3983
70_20_10	4	92	34,3354	55	33,7960	0,5394
70_20_10	5	169	35,3704	235	35,7334	-0,3631
80_05_15	1	265	36,5131	287	36,5434	-0,0303
80_05_15	2	258	36,2111	293	33,2584	2,9527
80_05_15	3	198	36,1348	150	36,2936	-0,1588
80_05_15	4	55	34,3354	55	33,7960	0,5394
80_05_15	5	121	35,3862	216	36,2514	-0,8652
80_15_05	1	160	36,5131	254	36,5656	-0,0525
80_15_05	2	217	36,3207	243	33,5775	2,7432
80_15_05	3	192	36,1527	96	36,5913	-0,4386
80_15_05	4	55	34,3354	55	33,7960	0,5394
80_15_05	5	250	35,3427	224	35,7914	-0,4488
80_20_00	1	180	35,4427	238	36,4443	-1,0016
80_20_00	2	250	36,3378	270	34,5569	1,7809
80_20_00	3	240	36,1473	117	36,5913	-0,4440
80_20_00	4	169	34,5757	55	33,7960	0,7797
80_20_00	5	194	35,3744	271	35,7888	-0,4144
85_05_10	1	288	36,1612	290	36,3505	-0,1893
85_05_10	2	279	36,4962	278	34,7172	1,7790
85_05_10	3	295	35,3143	251	36,2945	-0,9802

Tabela A.23: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
85_05_10	4	55	34,3354	55	33,7960	0,5394
85_05_10	5	290	35,3704	220	35,7888	-0,4184
85_10_05	1	160	36,5131	274	36,5633	-0,0502
85_10_05	2	217	36,3207	209	34,6441	1,6766
85_10_05	3	282	35,9846	252	36,5973	-0,6127
85_10_05	4	55	34,3354	55	33,7960	0,5394
85_10_05	5	256	35,4091	270	35,7686	-0,3595
90_05_05	1	118	36,4291	273	36,5084	-0,0793
90_05_05	2	281	36,4634	289	34,4250	2,0384
90_05_05	3	233	36,1167	257	36,6185	-0,5017
90_05_05	4	55	34,3354	55	33,7960	0,5394
90_05_05	5	174	35,3704	247	35,7045	-0,3342
90_10_00	1	282	35,9436	270	36,5154	-0,5718
90_10_00	2	223	36,4774	239	34,2418	2,2356
90_10_00	3	243	36,1672	274	36,2945	-0,1273
90_10_00	4	55	34,3354	55	33,7960	0,5394
90_10_00	5	207	35,3704	253	35,7840	-0,4136

Tabela A.24: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família DUF849 em 32 grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	133	11,7440	288	11,6823	0,0617
70_10_20	2	275	11,8022	184	11,7161	0,0861
70_10_20	3	227	11,4602	206	11,4284	0,0317
70_10_20	4	244	11,9082	137	11,5681	0,3401
70_10_20	5	283	11,7984	200	11,4537	0,3447
70_20_10	1	204	11,9066	201	11,7920	0,1146
70_20_10	2	23	11,6793	109	11,8120	-0,1328
70_20_10	3	258	11,4353	214	11,5530	-0,1176
70_20_10	4	179	11,9612	220	11,7374	0,2238
70_20_10	5	56	11,5014	277	11,3750	0,1265
80_05_15	1	193	11,8069	251	11,7006	0,1062
80_05_15	2	23	11,6793	266	11,8502	-0,1709
80_05_15	3	137	11,7847	247	11,8905	-0,1057
80_05_15	4	159	12,0662	233	11,7374	0,3288
80_05_15	5	234	11,6821	237	11,4513	0,2309
80_15_05	1	69	11,7440	188	11,7172	0,0269

Tabela A.24: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
80_15_05	2	106	11,7927	212	11,7676	0,0251
80_15_05	3	154	11,7146	234	11,4815	0,2331
80_15_05	4	135	11,8597	283	11,7374	0,1223
80_15_05	5	186	11,7254	239	11,4406	0,2848
80_20_00	1	126	11,7440	209	11,8406	-0,0966
80_20_00	2	108	11,7927	154	11,8120	-0,0193
80_20_00	3	242	11,8783	245	11,4950	0,3832
80_20_00	4	192	11,9612	31	11,5207	0,4405
80_20_00	5	242	11,9914	255	11,4682	0,5232
85_05_10	1	122	11,7440	242	11,8397	-0,0957
85_05_10	2	23	11,6793	170	11,8120	-0,1328
85_05_10	3	207	11,7052	257	11,6004	0,1048
85_05_10	4	247	12,0662	31	11,5207	0,5455
85_05_10	5	265	11,9503	212	11,3112	0,6391
85_10_05	1	107	11,7959	166	11,8153	-0,0194
85_10_05	2	106	11,7927	199	11,8477	-0,0550
85_10_05	3	273	11,8192	214	11,5651	0,2540
85_10_05	4	143	11,8597	31	11,5207	0,3389
85_10_05	5	56	11,5014	250	11,4550	0,0464
90_05_05	1	70	11,7440	127	11,6945	0,0495
90_05_05	2	186	11,7927	159	11,7676	0,0251
90_05_05	3	159	11,6886	256	11,4479	0,2407
90_05_05	4	247	12,0527	31	11,5207	0,5320
90_05_05	5	231	11,9336	216	11,4349	0,4987
90_10_00	1	70	11,7440	137	11,6945	0,0495
90_10_00	2	23	11,6793	207	11,8120	-0,1328
90_10_00	3	164	11,6680	231	11,4101	0,2580
90_10_00	4	187	11,9807	176	11,5681	0,4126
90_10_00	5	264	11,8011	260	11,4110	0,3901

Tabela A.25: Valores de informação mútua para os melhores indivíduos produzidos pelo sistema de GP para divisão da família DUF849 em 84 grupos.

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	1	255	5,0458	192	5,0734	-0,0276
70_10_20	2	277	4,9471	280	4,9126	0,0344
70_10_20	3	69	4,8371	50	4,8410	-0,0039
70_10_20	4	154	5,2693	251	5,0140	0,2553

Tabela A.25: (continuação)

Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
70_10_20	5	273	5,1853	180	5,1438	0,0415
70_20_10	1	128	5,0204	172	5,0392	-0,0188
70_20_10	2	248	4,7610	256	4,8536	-0,0925
70_20_10	3	190	4,9392	50	4,8410	0,0982
70_20_10	4	157	5,0357	263	5,0283	0,0075
70_20_10	5	176	5,0762	267	5,1087	-0,0325
80_05_15	1	199	5,0915	189	5,1587	-0,0671
80_05_15	2	291	4,7903	245	4,8200	-0,0297
80_05_15	3	209	4,9154	262	4,9387	-0,0233
80_05_15	4	236	5,2693	265	5,0245	0,2447
80_05_15	5	99	5,0519	280	5,1665	-0,1145
80_15_05	1	260	5,0387	199	5,0437	-0,0051
80_15_05	2	207	4,8066	274	4,8506	-0,0440
80_15_05	3	244	4,9437	249	4,8469	0,0968
80_15_05	4	213	5,0215	63	4,9899	0,0316
80_15_05	5	138	5,0955	152	5,0520	0,0435
80_20_00	1	261	5,2048	245	5,1234	0,0814
80_20_00	2	192	4,7395	275	4,7768	-0,0373
80_20_00	3	250	4,9388	218	4,9598	-0,0210
80_20_00	4	99	5,2693	277	5,0021	0,2671
80_20_00	5	55	5,0408	229	5,1012	-0,0604
85_05_10	1	257	5,0763	210	5,1058	-0,0294
85_05_10	2	258	4,9212	194	4,7547	0,1665
85_05_10	3	272	4,8951	254	4,9176	-0,0225
85_05_10	4	210	5,0416	269	5,0855	-0,0439
85_05_10	5	230	5,0739	101	5,1301	-0,0562
85_10_05	1	253	5,0876	262	5,1508	-0,0631
85_10_05	2	239	4,8581	227	4,8120	0,0461
85_10_05	3	197	4,9932	169	4,8784	0,1148
85_10_05	4	168	4,9946	238	5,0237	-0,0291
85_10_05	5	223	5,0680	160	5,0335	0,0345
90_05_05	1	148	5,0801	264	5,0506	0,0295
90_05_05	2	266	4,9045	273	4,7889	0,1156
90_05_05	3	268	4,8465	223	4,8488	-0,0023
90_05_05	4	228	5,0213	270	5,1246	-0,1033
90_05_05	5	197	5,0993	247	5,1166	-0,0173
90_10_00	1	287	5,0862	265	5,0273	0,0588
90_10_00	2	264	4,8635	244	4,7942	0,0692

Tabela A.25: (continuação)

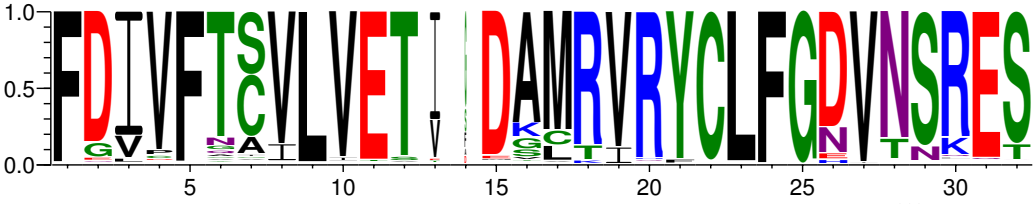
Parâmetros	Repetição	Grafo com Positivos		Grafo com Todos		Diferença
		Indivíduo	MI	Indivíduo	MI	
90_10_00	3	51	4,8363	276	4,8855	-0,0492
90_10_00	4	230	5,2693	275	5,0632	0,2061
90_10_00	5	236	5,0964	205	5,2123	-0,1159

Apêndice B
Resultados do ASMC para as Famílias Proteicas Originais

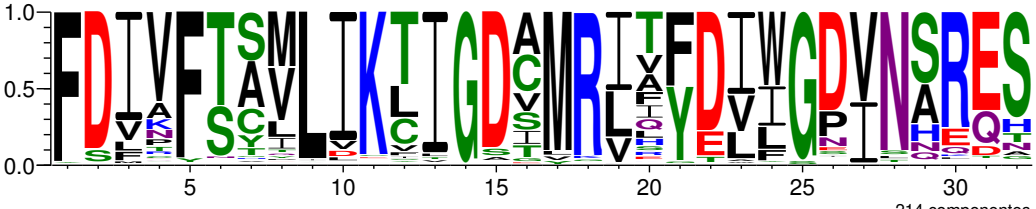
Neste capítulo são apresentados os agrupamentos gerados pelo ASMC (Melo-Minardi et al., 2010) para as famílias proteicas como utilizadas por eles, antes de serem retiradas as proteínas que, desde então, foram excluídas do UniProt.

B.1 Nucleotidil Ciclases

Antes da remoção das 75 proteínas obsoletas no UniProt, essa família continha 536 proteínas, sendo 213 com rótulo de Adenilato Ciclases e 323, de Guanilato Ciclases. A árvore gerada pelo ASMC divide as Nucleotidil Ciclases em dois grupos já no primeiro nível. A Figura B.1 mostra os logotipos dos grupos gerados.



(a) Grupo I: 322 Guanilato Ciclases



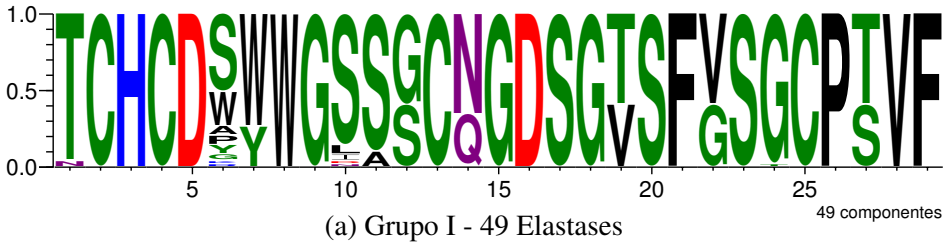
(b) Grupo II: 213 Adenilato Ciclases e 1 Guanilato Ciclase

Figura B.1: Divisão do conjunto original de Nucleotidil Ciclases em dois grupos no primeiro nível do agrupamento hierárquico do ASMC.

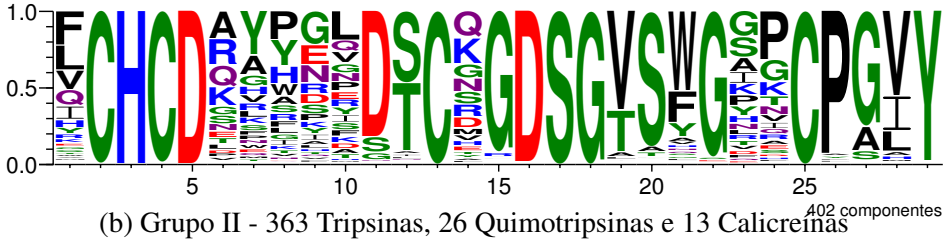
B.2 Serino Proteases

Antes da remoção das 140 proteínas obsoletas, essa família continha 1.673 proteínas, sendo 1.598 com rótulo de Tripsinas, 49, de Elastases e 26, de Quimotripsinas. Melo-Minardi et al. (2010) re-

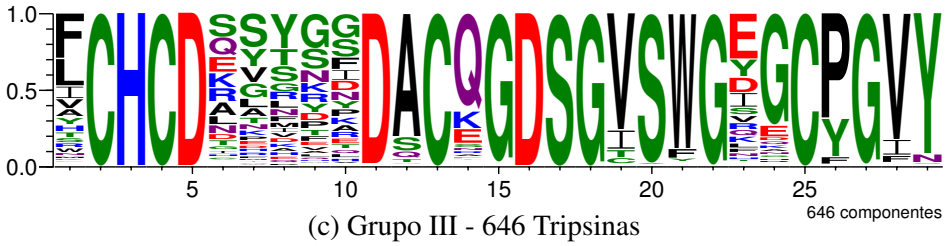
portaram também um subgrupo de 13 Tripsinas rotuladas como Calicreínas. No primeiro nível do agrupamento hierárquico gerada pelo ASMC, as Serino Proteases são divididas em quatro grupos, cujos logotipos são mostrados na Figura B.2.



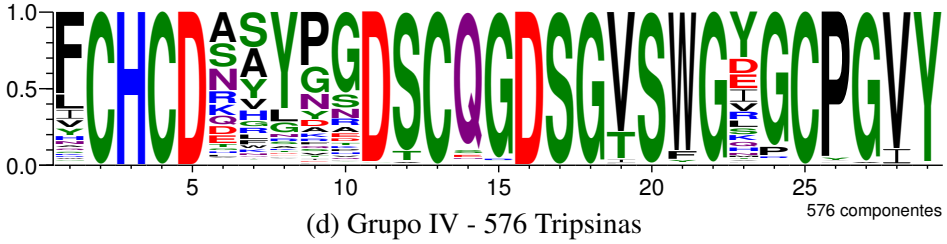
(a) Grupo I - 49 Elastases



(b) Grupo II - 363 Tripsinas, 26 Quimotripsinas e 13 Calicreínas



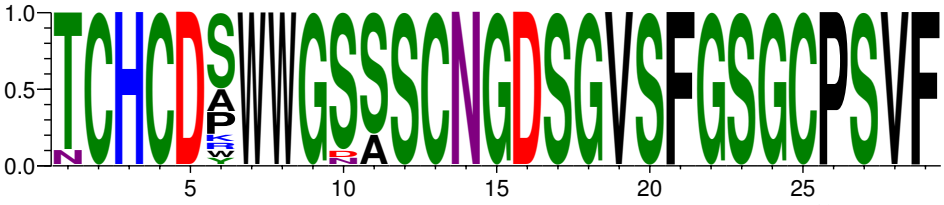
(c) Grupo III - 646 Tripsinas



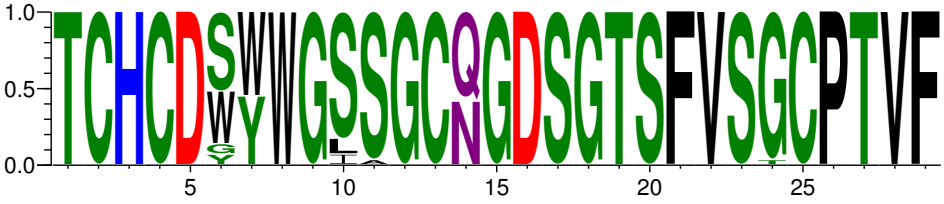
(d) Grupo IV - 576 Tripsinas

Figura B.2: Divisão do conjunto original de Serino Proteases em quatro grupos no primeiro nível do agrupamento hierárquico do ASMC.

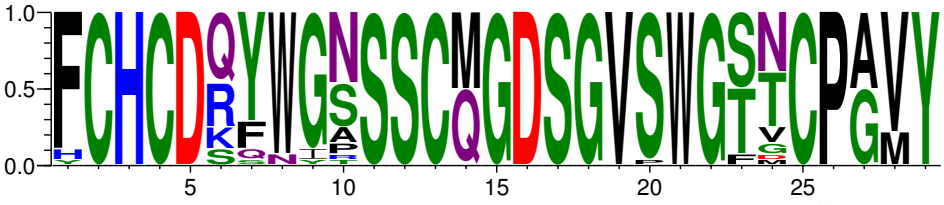
No segundo nível do agrupamento hierárquico do ASMC, o grupo de Elastases foi subdividido em dois, o Grupo II foi subdividido em quatro, e os Grupos III e IV foram subdivididos em dois cada um, como mostra a Figura B.3.



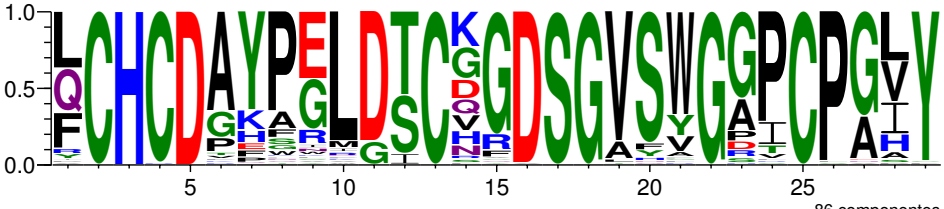
(a) Grupo I.1 - 20 Elastases



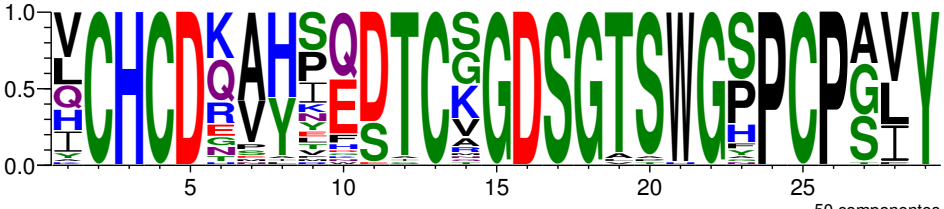
(b) Grupo I.2 - 29 Elastases



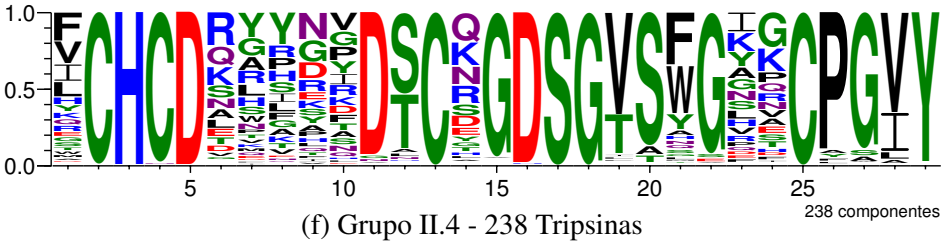
(c) Grupo II.1 - 26 Quimotripsinas e 2 Tripsinas



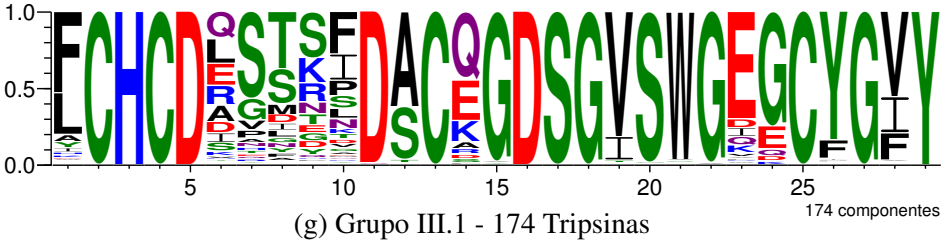
(d) Grupo II.2 - 86 Tripsinas



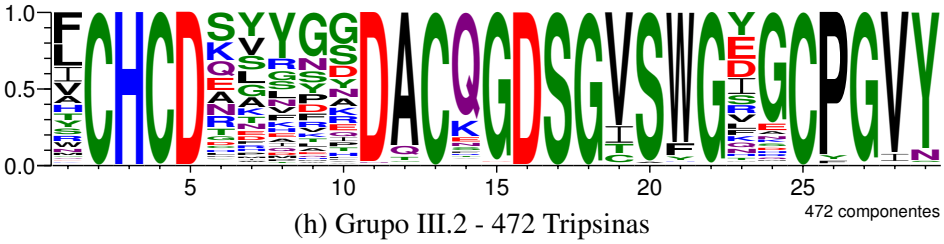
(e) Grupo II.3 - 37 Tripsinas e 13 Calicreínas



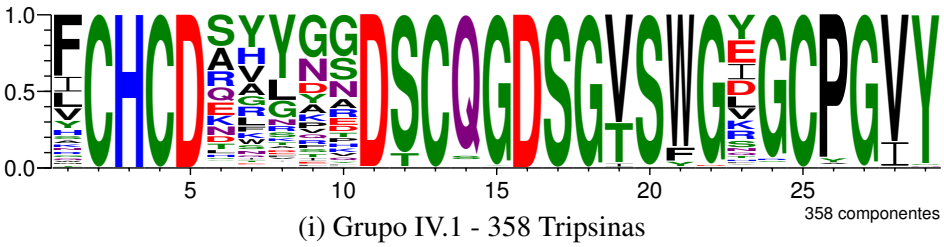
(f) Grupo II.4 - 238 Tripsinas



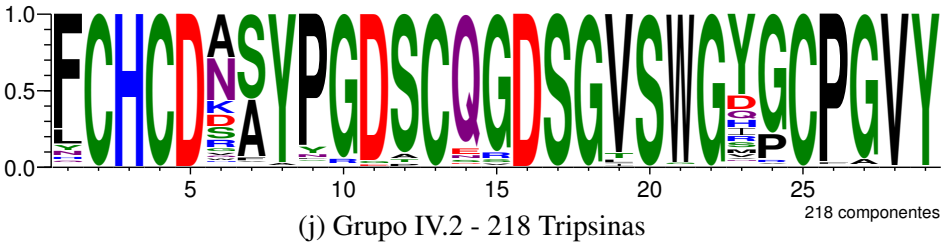
(g) Grupo III.1 - 174 Tripsinas



(h) Grupo III.2 - 472 Tripsinas



(i) Grupo IV.1 - 358 Tripsinas

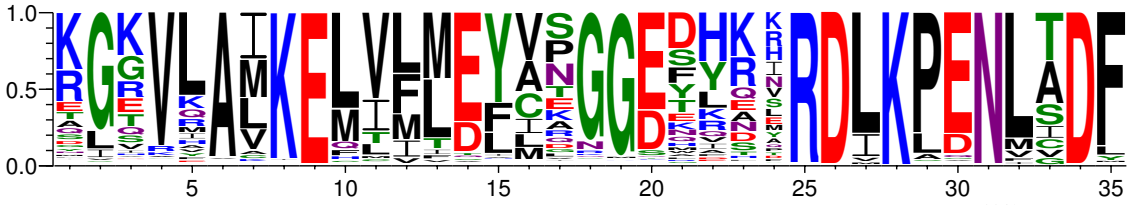


(j) Grupo IV.2 - 218 Tripsinas

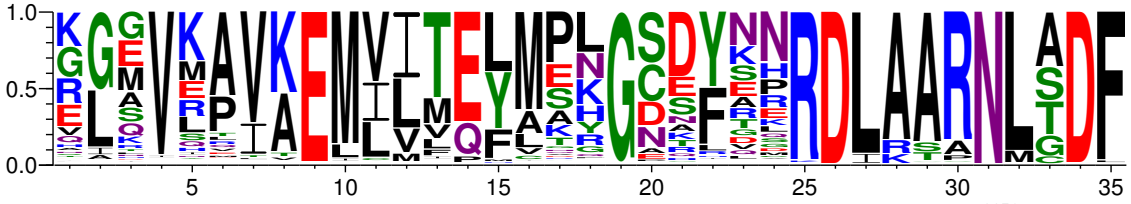
Figura B.3: Divisão do conjunto original de Serino Proteases em dez grupos no segundo nível do agrupamento hierárquico do ASMC.

B.3 Proteínas Cinases

Essa família continha 3.401 proteínas antes da remoção das 314 proteínas obsoletas no UniProt, sendo 2.250 com rótulo de Serina/Treonina Cinases e 1.151, de Tirosina Cinases. Melo-Minardi et al. (2010) reportaram também um subgrupo das Tirosina Cinases contendo 237 EGFRs (do inglês *Epidermal Growth Factor Receptor*). No primeiro nível do agrupamento hierárquico gerada pelo ASMC, a família foi dividida em dois grupos, cujos logotipos são apresentados na Figura B.4.



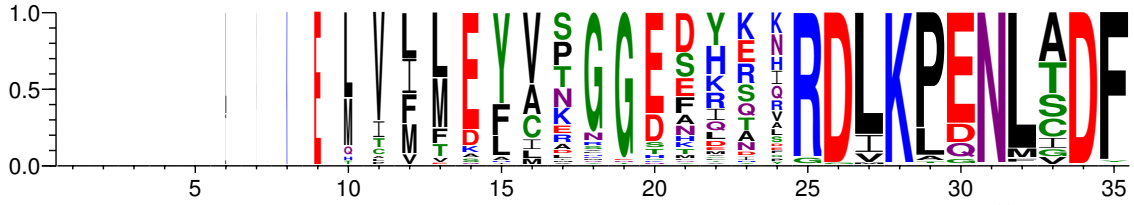
(a) Grupo I: 2249 Serina/Treonina Cinases



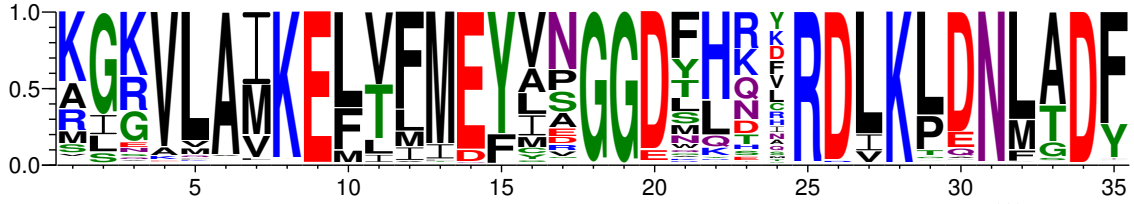
(b) Grupo II: 914 Tirosina Cinases, 237 EGFRs e 1 Serina/Treonina Cinase

Figura B.4: Divisão do conjunto original de Proteínas Cinases em dois grupos no primeiro nível do agrupamento hierárquico do ASMC.

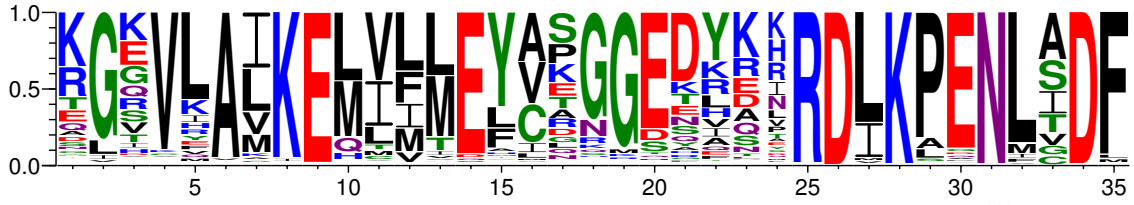
Os logotipos dos grupos do segundo nível do agrupamento hierárquico do ASMC são apresentados na Figura B.5. O grupo de Serina/Treonina Cinases foi subdividido em quatro (Subfiguras (a) a (d)), enquanto o das Tirosina Cinases foi subdividido em dois (Subfiguras (e) e (f)).



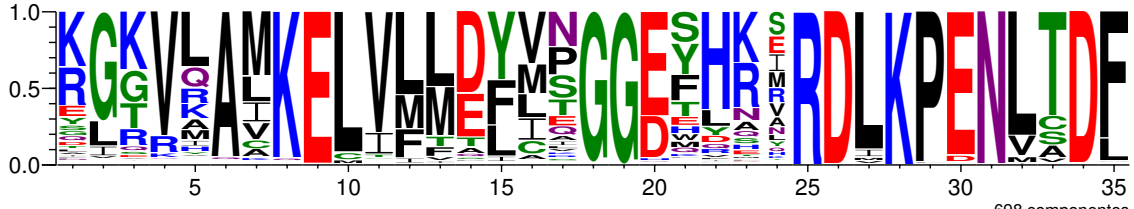
(a) Grupo I.A: 110 Serina/Treonina Cinases



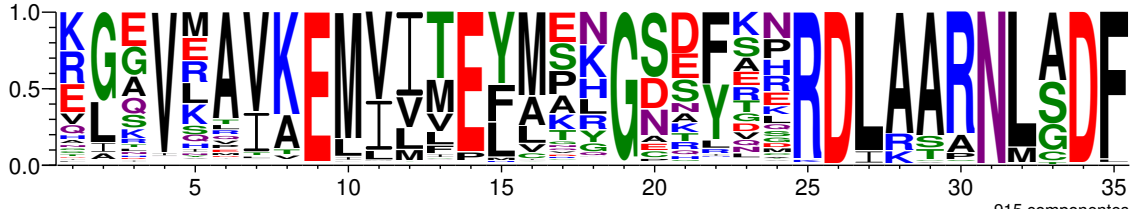
(b) Grupo I.B: 313 Serina/Treonina Cinases



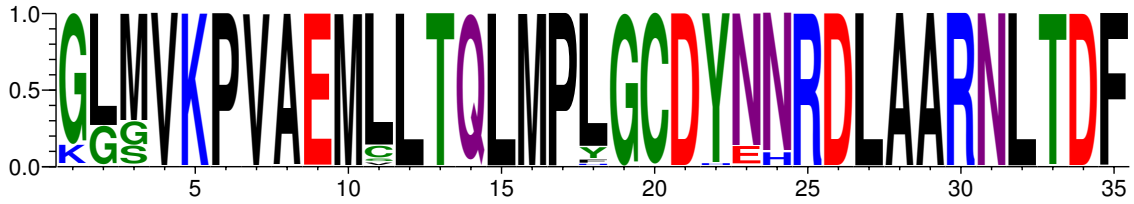
(c) Grupo I.C: 1128 Serina/Treonina Cinases



(d) Grupo I.D: 698 Serina/Treonina Cinases



(e) Grupo II.A: 914 Tirosina Cinases e 1 Serina/Treonina Cinase



(f) Grupo II.B: 237 EGFRs

Figura B.5: Divisão do conjunto original de Proteínas Cinases em seis grupos no segundo nível do agrupamento hierárquico do ASMC.