CSM: UMA ASSINATURA PARA GRAFOS BIOLÓGICOS BASEADA EM PADRÕES DE DISTÂNCIAS

DOUGLAS EDUARDO VALENTE PIRES

CSM: UMA ASSINATURA PARA GRAFOS BIOLÓGICOS BASEADA EM PADRÕES DE DISTÂNCIAS

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

Orientador: Wagner Meira Jr.
Co-Orientador: Raquel Cardoso de Melo Minardi
Co-Orientador: Carlos Henrique da Silveira

Belo Horizonte
01 de outubro de 2012

Pires, Douglas Eduardo Valente
CSM: Uma Assinatura para Grafos Biológicos Baseada em Padrões de Distâncias / Douglas Eduardo Valente Pires. — Belo Horizonte, 2012
xxix, 123 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Wagner Meira Jr.Co-Orientador: Raquel Cardoso de Melo MinardiCo-Orientador: Carlos Henrique da Silveira

1. Assinatura Estrutural. 2. Cutoff Scanning Matrix. 3. Contatos. 4. Classificação Estrutural. 5. Predição de Função Proteica. 6. Predição de Ligantes. 7. Predição de Toxicidade. 8. Grafos. 9. Mineração de Dados. I. Título.

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

À Deus,
aos meus pais, Aparecida e Dinarte,
ao meu grande amor Grasiela,
aos meus amigos, professores
colegas de curso e de laboratório,
dedico este trabalho.

Agradecimentos

Sou grato primeiramente a Deus, pelo dom da vida e por todas as conquistas alcançadas. Obrigado, Senhor, por ter me sustentado e iluminado meus caminhos até aqui.

Agradeço também aos meus pais, Aparecida e Dinarte, pelo amor incondicional, pela paciência, incentivo e presença em cada momento. Vocês são minha referência e meu alicerce. Eu amo vocês.

À minha namorada Grasiela agradeço, por seu imenso amor, dedicação e por todas as palavras de carinho e incentivo. Meu amor, suas palavras me acalmam e confortam, seus olhinhos me inspiram e cativam, seu sorriso me alegra e fortalece. Eu amo você.

Aos meus professores, agradeço pelos conhecimentos adquiridos. Em particular, agradeço ao meu orientador Wagner Meira Jr. pelo apoio, confiança e ensinamentos que, desde a Iniciação Científica, moldaram meu perfil acadêmico.

Agradeço ainda a meus co-orientadores e amigos Carlos Silveira e Raquel Minardi. Silveira, você me apresentou à Bioinformática e me incentivou desde o início. Raquel, agradeço por ter me acompanhado de perto nessa caminhada. Obrigado pelo apoio. Saibam que vocês são pra mim exemplos de dedicação e amor à ciência.

Agradeço também a todos os colegas dos laboratórios LBS e e-Speed, em especial ao Fernando (FH) e ao Sandro pela troca de experiências e pelas várias e animadas conversas. Em particular também agradeço a Valdete pelo companheirismo diário, pelas palavras de apoio. Você é uma grande amiga, e não é só porque é flamenguista e nasceu no mesmo dia que eu. Obrigado por tudo.

Gostaria de agradecer aos meus amigos Felipe Ferré e Luiz Fernando (Fernandinho), que me mostraram que mesmo em momentos difíceis temos que rir da vida. Agradeço também a meus grandes amigos Janaína, Rodrigo e Thatyene. A amizade e apoio de vocês foi, é e sempre será fundamental.

Também agradeço a todos que torceram por mim e que, mesmo indiretamente, contribuíram para que esse sonho fosse alcançado.

"É melhor tentar e falhar,
que preocupar-se e ver a vida passar.
É melhor tentar, ainda que em vão,
que sentar-se fazendo nada até o final.
Eu prefiro na chuva caminhar,
que em dias tristes em casa me esconder.
Prefiro ser feliz, embora louco,
que em conformidade viver."
(Martin Luther King)

Resumo

O ritmo acelerado de geração e disponibilização de dados biológicos tem criado diversos desafios computacionais no que diz respeito ao processamento e extração de informações relevantes, bem como padrões não-triviais a partir desses grandes volumes de dados.

Nesse contexto, uma modelagem computacional apropriada, eficiente e escalável faz-se necessária. É cada vez maior o número de sistemas reais que podem ser modelados computacionalmente como redes ou grafos, representações abstratas de entidades e seus relacionamentos que se mostraram muito eficazes na modelagem de fenômenos, sistemas e processos naturais.

Em muitos cenários reais atuais entretanto, modelos de grafos tradicionais não são aplicáveis ou falham em virtude de falta de adequação, escalabilidade ou em virtude da dinamicidade da informação, o que tem criado uma demanda relevante por novos paradigmas, modelos e algoritmos para que redes biológicas em larga escala sejam devidamente analisadas e os fenômenos que as governam, compreendidos.

Neste trabalho, apresentamos um novo modelo para geração de assinaturas para grafos biológicos denominada Cutoff Scanning Matrix (CSM). O CSM gera vetores de atributos que representam padrões de distâncias entre nós de um grafo, que são então usados como evidência em tarefas de classificação. Adicionalmente, a Decomposição em Valores Singulares (SVD) é empregada como um passo de pré-processamento para reduzir a dimensionalidade e o ruído inerente aos dados. A metodologia proposta é instanciada com sucesso em diversos cenários, tais como anotação automática proteica, predição de ligantes e de atividade de pequenas moléculas.

Os resultados obtidos nas diferentes tarefas mostram que os padrões de distâncias em grafos correspondem a um componente robusto e conservado, sendo uma importante fonte de informação topológica. Adicionalmente, as assinaturas derivadas do conceito CSM mostraram-se eficazes e eficientes na resolução das diversas tarefas propostas sendo comparáveis ou superiores aos principais trabalhos concorrentes. Por fim, a aplicabilidade do conceito CSM em três diferentes grafos biológicos mostra, além de sua generalidade, um grande potencial ainda a ser explorado.

Abstract

The unforgiving pace of growth of available biological data has generated several computational challenges regarding processing and extracting relevant, informative and non-trivial information from such large volumes of data.

In this context, an appropriate, efficient and scalable computational approach is necessary. An increasing number of real systems may be computationally modeled as networks or graphs, abstract representations of entities and their relationships that have proved very effective in modeling natural phenomena, processes and systems.

In many current real scenarios however, traditional graph models are not applicable or fail due to inadequacy, lack of scalability, or due to information dynamics, which has created a relevant demand for new paradigms, models and algorithms that properly analyze large-scale biological networks and characterize the phenomena that govern them.

In this thesis, we present a new methodology for generating signatures for biological graphs called Cutoff Scanning Matrix (CSM). The CSM generates vectors of attributes that represent distance patterns between nodes of a graph, which are then used as evidence in classification tasks. In addition, Singular Value Decomposition (SVD) is used as a preprocessing step to reduce the dimensionality and inherent noise in the data. The proposed methodology was successfully instantiated in various scenarios, such as protein automatic annotation, receptor-based ligand prediction and anti-cancer activity, mutagenesis and toxicity prediction for small molecules.

The results obtained in the different tasks show that distances patterns in graphs correspond to a robust and conserved component, being an important source of topological information. Additionally, the signatures derived from the CSM concept proved to be effective and efficient in solving different tasks, being comparable or superior to the main competitors works. Finally, the CSM concept applicability in different scenarios and biological networks shows, besides its generality, a great potential still to be explored.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Vivemos em tempos de alta disponibilidade e geração contínua de grandes volumes de dados biológicos, os quais estão sujeitos a alterações constantes, o que os confere uma natureza altamente dinâmica. De um lado, o UniprotKB/TrEMBL [Consortium, 2010] possui 24.000.000 de sequências proteicas. No mês de setembro, mais de 800.000 novas sequências foram adicionadas a esse repositório, e em torno de 4.900.000 registros de anotação foram revisados. Por outro lado, a base de dados de famílias proteicas PFam [Finn et al., 2010] cobre em torno de 13.600 famílias, sendo 25% domínios de função desconhecida (*Domains of Unkowns Function* - DUFs), revelando que o estado-da-arte em métodos de anotação baseados em sequência e mesmo aqueles baseados em perfil tem tido um sucesso limitado na tarefa de atribuir função a proteínas recém descobertas.

Bases de dados de classificação estrutural proteica, como o SCOP [Murzin et al., 1995], também apresentam dificuldades para manterem-se atualizados dado o número crescente de estruturas de proteínas resolvidas e depositadas em repositórios públicos. Aproximadamente 53% dos registros do Protein Data Bank (PDB) [Berman et al., 2002] estão classificados pela versão 1.75 do SCOP, e após remover-se redundância (90% de similaridade de sequência), a cobertura cai para em torno de 41%. À medida que iniciativas internacionais de genômica estrutural têm produzido um grande número de estruturas sem função conhecida, tentativas de assinalamento automático de função a essas proteínas tornam-se cada vez mais necessárias, e um esforço significativo tem sido empregado nessa tarefa [Laskowski et al., 2005b,a; Watson et al., 2005, 2007]. No entanto, o estudo de novas abordagens eficientes e escaláveis para a anotação automática proteica ainda são necessárias nesse contexto.

O problema torna-se ainda mais evidente quando consideramos que dados proteicos correspondem apenas a uma dentre as várias possíveis fontes de informação

biológica. O crescente número de projetos genoma e metagenoma, juntamente com iniciativas de genômica estrutural [Chandonia & Brenner, 2006], também tem contribuído de forma significativa para aumento da produção de dados biológicos.

O ritmo acelerado de geração e disponibilização de dados biológicos tem gerado, nesse sentido, diversos desafios computacionais no que diz respeito ao processamento e extração de informações relevantes, inéditas, bem como padrões não-triviais a partir desses grandes volumes de dados.

Nesse contexto, uma modelagem computacional apropriada, eficiente e escalável faz-se necessária. Na natureza, entidades biológicas relacionam-se de modo a formar sistemas complexos, sendo o caráter e intensidade desses relacionamentos responsáveis por definir o comportamento tanto dos objetos quanto do sistema como um todo. Por exemplo, funções biológicas podem ser descritas e compreendidas a partir de interações entre biomoléculas. Assim, a modelagem de objetos biológicos na forma de redes parece ser uma estratégia bastante pertinente.

Redes ou grafos são representações abstratas de entidades e seus relacionamentos que se mostraram muito eficazes na modelagem de fenômenos, sistemas e processos naturais, sendo cada vez maior o número de sistemas reais que podem ser modelados computacionalmente como grafos. Em muitos cenários reais atuais entretanto, modelos de grafos tradicionais não são aplicáveis ou falham em virtude de falta de adequação, escalabilidade ou em virtude da dinamicidade da informação. Tais desafios tem criado uma demanda relevante por novos paradigmas, modelos e algoritmos para que redes biológicas em larga escala sejam devidamente analisadas e os fenômenos que as governam, compreendidos.

Um possível abordagem refere-se à busca por assinaturas topológicas. Assinaturas podem ser vistas como um conjunto de características que define, agrupa ou discrimina um conjunto de objetos dos demais. Um assinatura para grafos, por conseguinte, tem por objetivo agregar atributos de naturezas diversas que sejam comuns a grafos similares, seja do ponto de vista estrutural (perfil topológico global ou local do grafo) ou mesmo em relação a seus rótulos. A identificação de um conjunto de características que compõe uma assinatura adequada, no entanto, é uma tarefa não trivial. Deseja-se que uma assinatura de grafos seja:

- **Conservada:** coerente e consistente com a definição do grupo formado por grafos similares;
- **Concisa:** de modo a facilitar sua geração, utilização, análise e armazenamento;
- **Eficiente:** espera-se que a assinatura seja escalável para bases de dados em

crescimento acelerado;

- **Generalizável:** deseja-se, em última instância, que uma assinatura estrutural seja útil em vários contextos e domínios de aplicação.

À luz desses requisitos, é proposto no presente trabalho uma assinatura para grafos baseada em padrões de distâncias entre seus nós chamada *Cutoff Scanning Matrix* (CSM). O CSM gera vetores de atributos que representam padrões de distâncias entre nós de um grafo, que são então usados como evidência em tarefas de classificação.

Dado o grande volume de informação a ser tratada, bem como os requisitos de escalabilidade, uma etapa que reduza o ruído inerente aos dados bem como os custos computacionais de processamento das assinaturas faz-se necessária. Para tal, a Decomposição em Valores Singulares (SVD) é empregada como um passo de pré-processamento para reduzir a dimensionalidade e o ruído existente nos dados. A metodologia proposta é instanciada com sucesso em diversos cenários, tais como anotação automática proteica, predição de ligantes e de atividade de pequenas moléculas.

A metodologia aqui proposta é oportuna ao passo que aborda questões atuais e relevantes para o desenvolvimento de áreas como a Mineração em Grafos, Bioinformática Estrutural e Quimioinformática. Além disso propomos soluções prontamente aplicáveis a cenários e bases de dados reais, em um contexto de alta disponibilidade e volume de informação a ser tratada.

Dentre as principais contribuições do presente trabalho está a proposta de uma nova abordagem para definição de assinaturas para grafos. Essa abordagem mostrou-se útil em tarefas de classificação estrutural e inferência de função proteica e pode ainda contribuir direta ou indiretamente, como subsídio, para o desenvolvimento de outras aplicações como estudos de interação proteína-proteína, interação ligante-ligante, que consequentemente podem levar à descoberta de novos alvos terapêuticos e fármacos. Relatamos também o sucesso da aplicação do conceito CSM na definição de assinaturas para tarefas de predição de ligantes baseada no receptor e inferência de atividade de pequenas moléculas.

Os resultados obtidos nas diferentes tarefas mostram que os padrões de distâncias em grafos correspondem a um componente robusto e conservado, sendo uma importante fonte de informação topológica. Adicionalmente, as assinaturas derivadas do conceito CSM mostraram-se eficazes e eficientes na resolução das diversas tarefas propostas sendo comparáveis ou superiores aos principais trabalhos concorrentes. Por fim, a aplicabilidade do conceito CSM em diferentes grafos biológicos mostra, além de sua generalidade, um grande potencial ainda a ser explorado.

1.1 Objetivos

1.1.1 Objetivos Gerais

O foco principal do trabalho consiste na proposta, implementação, validação e avaliação de uma assinatura para grafos biológicos baseada em padrões de distâncias aplicada a tarefas tais como anotação automática proteica, predição de ligantes e de atividade de pequenas moléculas.

1.1.2 Objetivos Específicos

- **Projetar, implementar e validar uma assinatura de grafos baseada em padrões de distâncias (aqui denominada CSM):** diz respeito ao ponto focal do trabalho que consiste na proposta da utilização do conceito CSM para geração de descritores de grafos biológicos.

- **Instanciar a assinatura proposta em diferentes tarefas:** corresponde aos três cenários de aplicação das assinaturas estudadas, a saber, a anotação automática de estruturas proteicas (que envolve classificação estrutural e predição de função), predição de ligantes baseada no receptor proteico e predição de atividade anticâncer, toxicidade e mutagênese de pequenas moléculas. Cada uma possui características particulares tanto em termos da modelagem em grafos quanto da geração das assinaturas. Nesse sentido, foram derivadas três assinaturas para cada respectiva tarefa (CSM, aCSM e gCSM) a partir do conceito CSM.

- **Avaliar as diferentes estratégias de mineração de dados sobre essas assinaturas:** trata do problema de definir quais algoritmos utilizar nas diferentes tarefas de mineração de dados, além da definição de passos de pré-processamento que visam a redução de dimensionalidade dos dados (visando escalabilidade), selecionando apenas informações úteis (de modo a maximizar a eficácia da metodologia).

- **Analisar a semântica por trás das assinaturas, bem como realizar sua caracterização para grupos de dados similares:** refere-se à análise dos padrões de distâncias encontrados para variados grupos de dados de modo a atribuir semântica a esses. Um possível exemplo de aplicação seria a análise de padrões em famílias de estruturas proteicas.

O restante do texto está dividido da seguinte forma: No Capítulo 2, é realizada uma revisão acerca da modelagem via grafos e das abordagens atuais para extração de descritores e tarefas de classificação. O Capítulo 3 introduz o conceito CSM, passo fundamental na definição da assinatura de grafos aqui proposta, baseada em padrões de distâncias, bem como os requisitos para sua utilização e fluxo metodológico. No Capítulo seguinte (Capítulo 4), o conceito CSM é instanciado na forma de assinatura estrutural proteica, em nível de resíduos, e os resultados de sua aplicação em tarefas como a classificação estrutural e predição de função proteica são reportados. Em seguida, no Capítulo 5, a tarefa de predição de ligantes baseada no receptor é abordada a partir de uma assinatura CSM para *pockets* proteicos em nível atômico. Finalmente, no Capítulo 6, é documentada a utilização bem sucedida do conceito CSM na definição de uma assinatura para grafos de pequenas moléculas com o intuito de predizer características, tais como, atividade anticâncer, toxicidade e mutagênese. Por fim, o Capítulo 7 apresenta as conclusões, trabalhos e perspectivas futuras.

Capítulo 2

Estudo e Modelagem de Redes Biológicas

Redes ou grafos são representações abstratas de entidades e seus relacionamentos que se mostraram muito eficazes na modelagem de fenômenos, sistemas e processos naturais. Parte do sucesso de tais modelos matemáticos se deve a uma teoria subjacente robusta e madura, desenvolvida ao longo dos anos. De fato, a primeira referência ao termo *grafo* em seu sentido moderno data de 1.878 [Gross & Yellen, 2004].

Cada vez um número maior de sistemas reais podem ser modelados computacionalmente como grafos. Dentre os exemplos mais proeminentes encontram-se:

- **Redes biológicas:** como as redes metabólicas, cadeias alimentares, redes de interações proteicas e de regulação genética;
- **Redes tecnológicas:** como a internet, as redes de transmissão de energia e de comunicação e as malhas viárias;
- **Redes sociais:** como as redes sociais *online* Facebook e Twitter, além de redes reais como as de crime organizado;
- **Redes informacionais:** como as redes P2P, redes de citações de artigos científicos e páginas de sites colaborativos.

Em muitos cenários reais atuais, entretanto, modelos de grafos tradicionais não são aplicáveis ou falham em virtude de falta de adequação (por não serem consistentes com a semântica biológica da rede) ou escalabilidade. Desafios surgem da dinamicidade da informação, bem como do ritmo acelerado de produção e depósito de novos dados em repositórios públicos, o que tem criado uma demanda relevante por novos paradigmas,

modelos e algoritmos para que redes em larga escala sejam devidamente analisadas e os fenômenos que as governam, compreendidos.

Nesse contexto, a mineração em grafos [Chakrabarti & Faloutsos, 2006] surge como uma disciplina-chave nesse processo. Dentre os objetivos da mineração em grafos, alguns dos quais abordados no presente trabalho, podemos destacar a classificação de grafos em categorias, o agrupamento de grafos similares, a comparação de grafos por meio de índices de similaridade e a recuperação de grafos por similaridade.

Nosso foco inicial refere-se à *classificação de grafos*. Em linhas gerais, nessa tarefa temos uma base de dados de grafos rotulados por categoria, e desejamos transformar cada um dos grafos em vetores de atributos que serão, em seguida, utilizados como insumo para treinamento de classificadores. A Seção 2.2 aborda detalhadamente os aspectos e desafios relacionados a tal tarefa.

2.1 Definição de Conceitos

Antes, porém, de entrarmos no mérito das tarefas e desafios abordados pelo presente trabalho, definimos formalmente a seguir um conjunto de conceitos-chave que serão utilizados ao longo de todo o texto:

Grafo: Um grafo $G = (V, E)$, consiste de um conjunto de vértices $V = \{v_0, v_1, ..., v_n\}$ e um conjunto de arestas $E = \{(v_i, v_j) : v_i, v_j \in V\}$. Sejam L_V e L_E conjuntos de rótulos/pesos de vértices e arestas, respectivamente, e sejam $\mathcal{V} : V \to L_V$ e $\mathcal{E} : E \to L_E$ funções que mapeiam rótulos/pesos para cada vértice e aresta. Um grafo é dito *ponderado* se possui pesos associados às suas arestas.

Subgrafo: Um grafo $G' = (V', E')$ é um subgrafo de $G = (V, E)$ se $V' \subseteq V$ e $E' \subseteq E$.

Isomorfismo de grafos: Dois grafos $G = (V, E)$ e $G' = (V', E')$ são isomorfos se existir uma função bijetora $f : V \to V'$ tal que a aresta $(u, v) \in E$ se e somente se $(f(u), f(v)) \in E'$. Em outras palavras, é possível mapear os vértices de G para os de G' mantendo a adjacência das arestas correspondentes.

Isomorfismo de subgrafos: Um grafo $G' = (V, E)$ é subgrafo isomorfo de $G = (V, E)$ se G' é isomorfo a algum subgrafo de G. O isomorfismo de subgrafos pertence à classe de problemas NP-Completo.

Caminho: Um caminho de comprimento k de um vértice x a um vértice y em um grafo $G = (V, E)$ é uma sequência de vértices $(v_0, v_1, ..., vk)$ tal que $x = v_0$ e $y = v_k$, e $(v_{i-1}, v_i) \in E$ para $i = 1, 2, ..., k$. O *comprimento* de um caminho é o

número de arestas nele. O *custo* de um caminho para um grafo ponderado é a soma dos pesos das arestas dele. Um caminho mínimo entre dois vértices é aquele com o menos custo.

Distância entre vértices: Dado um grafo $G = (V, E)$, a distância entre dois vértices corresponde ao número absoluto ou somatório dos pesos das arestas de um caminho mínimo que os conecta.

Matriz de adjacência: A matriz de adjacência A de um grafo $G = (V, E)$ contendo n vértices é uma matriz binária nxn, onde $A[i, j]$ é 1 (ou verdadeiro) se e somente se existe uma aresta do vértice i para o vértice j. Para grafos ponderados $A[i, j]$ contém o rótulo ou peso associado à aresta e, neste caso, a matriz não é binária.

Matriz de distância: A matriz de distância D de um grafo $G = (V, E)$ contendo n vértices é uma matriz numérica nxn, onde $D[i, j]$ corresponde ao custo do caminho mínimo entre os vértices i e j. Caso não exista um caminho entre i e j, dizemos que j não é alcançável a partir de i e, nesse caso, $D[i, j] = \infty$.

2.2 Classificação de Grafos

Em linhas gerais, aprendizado supervisionado diz respeito à inferência de uma função a partir de uma base de treino composta por um conjunto de exemplos e seu respectivo rótulo ou classe. Cada exemplo pode ser visto, então, como um par formado pelo objeto de entrada em sua representação (*e.g.*, conjunto de descritores) e o rótulo de saída esperado. Um algoritmo de aprendizado supervisionado analisa os dados de treino e produz uma função (ou modelo) de inferência, chamada de classificador (caso o rótulo de saída esperado seja discreto) ou função de regressão (caso seja contínuo).

Deseja-se que a função de inferência gerada consiga predizer o rótulo de saída para novos objetos válidos fornecidos como entrada, cujos rótulos são desconhecidos. Nesse sentido, é necessário que o algoritmo de aprendizado seja capaz de realizar generalizações a partir do conjunto de treinamento. Em suma, a classificação é um processo executado em dois passos, que correspondem à construção do modelo a partir de um conjunto de dados de treinamento e, em seguida, a classificação de elementos desconhecidos.

Dentre os principais fatores que influenciam no sucesso ou insucesso de uma tarefa de classificação estão os algoritmos de aprendizado supervisionado empregados e os dados utilizados como descritores das entidades a serem classificadas.

Uma vez que diversos algoritmos de aprendizado supervisionado foram desenvolvidos ao longo dos anos, e esses mostraram-se muito eficazes na descoberta e utilização de padrões para inferência para bons descritores de dados, o maior desafio reside na busca por tais atributos, de modo que sejam mais discriminativos e eficazes na representação das entidades consideradas e dos relacionamentos que norteiam sua classificação.

No caso da classificação de grafos, as abordagens tradicionais descritas na literatura dividem-se tipicamente em:

- Abordagens baseadas na busca de padrões;
- Abordagens baseadas em *funções kernel*.

A busca de padrões tipicamente envolve a mineração de subgrafos frequentes, abordagem de custo computacional muito elevado (exponencial), o que restringe sua aplicação. Já os *graph kernels* são funções, em geral de complexidade de tempo polinomial, que tentam calcular medidas de similaridade entre grafos de modo a compará-los.

Existem ainda abordagens que não se enquadram nas restrições de uma função kernel, como as baseadas em descritores topológicos e de rótulos que podem ser utilizados para descrever grafos ou porções de um grafo, formando *fingerprints* ou assinaturas.

A seguir, essas abordagens serão descritas mais detalhadamente.

2.2.1 Mineração de Subgrafos Frequentes

Na Bioinformática, a mineração de subgrafos frequentes tem sido utilizada em diversas tarefas, desde a mineração de resíduos específicos a certas famílias proteicas [Huan et al., 2004], à predição atividade anticâncer [Yan et al., 2008] e anti-HIV [Kong et al., 2011] de pequenas moléculas.

A Figura 2.1 exemplifica o processo geral de aplicações baseadas na mineração de subgrafos frequentes. A partir de uma base de grafos $D = \{G_1, G_2, ..., G_n\}$ e um valor de suporte mínimo θ, os subgrafos frequentes são aqueles que estão contidos em pelo menos $\theta|D|$ grafos em D. Após a extração dos subgrafos frequentes, são selecionados os padrões mais significativos, de acordo com uma métrica ou função objetivo definida pelo usuário. Uma vez selecionados, os padrões relevantes podem ser utilizados como descritores dos grafos em diversas aplicações.

Entretanto, a principal limitação de tal processo encontra-se em sua falta de escalabilidade. Para encontrarmos os subgrafos mais relevantes é necessário a

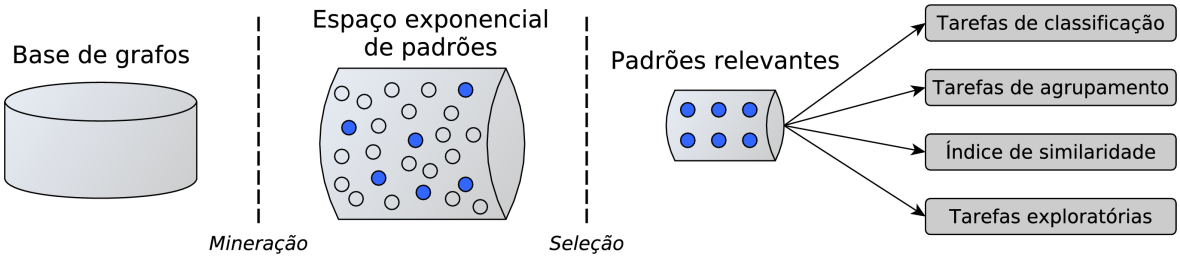

Figura 2.1. Processo geral de aplicações baseadas na mineração de subgrafos frequentes. A partir de uma base de dados de grafos, diversos algoritmos podem ser utilizados na mineração de um conjunto (potencialmente exponencial) de padrões ou subgrafos frequentes. Os padrões mais discriminativos ou relevantes podem ser selecionados para utilização em tarefas variadas.

enumeração de todos os subgrafos frequentes. Adicionalmente, em muitos casos é necessário que o valor suporte seja baixo para que nenhum padrão significativo seja perdido, o que muitas vezes pode levar a um conjunto exponencial de padrões. Aliado à redundância entre os padrões, esse fator limita enormemente a aplicação do processo, fazendo com que seja necessário abrir mão da qualidade dos padrões encontrados para que a computação seja finalizada em tempo hábil.

Nesse sentido, a comunidade científica tem mudado seu enfoque para a construção de algoritmos amostrais que identifiquem um conjunto reduzido, mas representativo, de subgrafos frequentes ao invés da enumeração completa de todos os padrões [Al Hasan & Zaki, 2009]. Outras abordagens de sumarização e amostragem de subgrafos frequentes foram relatadas por [Al Hasan et al., 2007] e [Al Hasan & Zaki, 2009].

2.2.2 Funções *Kernel*

Uma função *Kernel* $\mathcal{K}$ mapeia dois objetos de entrada x e x' em um espaço de atributos $\mathcal{H}$. Através de uma função de mapeamento ϕ, uma medida de similaridade entre os objetos nesse espaço é calculada a partir do produto interno:

$$\mathcal{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

A Figura 2.2 exemplifica o processo de transformação de um espaço não-linearmente separável em um domínio linearmente separável, mapeamento esse realizado por uma função *Kernel*.

Faz-se necessário que a função de mapeamento ϕ pertença a um domínio onde o cálculo do produto interno seja possível. Essas funções também satisfazem o Teorema de Mercer [Mercer, 1909].

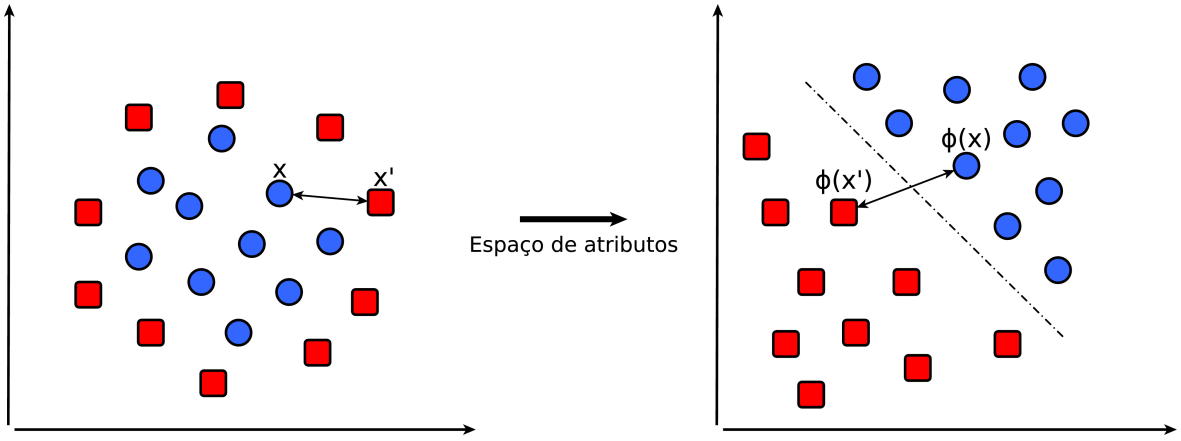

Figura 2.2. Processo de transformação de domínios não-linearmente separáveis para um espaço de atributos linearmente separável através da aplicação de uma função de mapeamento ϕ em cada objeto de entrada.

Teorema de Mercer. *Uma função é considerada kernel, se a matriz K é positivamente definida (seus autovalores são maiores que zero), sendo K:*

$$K = K_{x,x'} = \mathcal{K}(x, x')$$

Funções *kernel* podem ser executadas em pares de grafos de modo que a similaridade entre eles pode ser estimada. Diferentes estratégias de definição de novas funções *Kernels* para grafos tem sido utilizadas em Bioinformática e Quimioinformática, contemplando uma ampla gama de propósitos. Uma breve revisão sobre aplicações de kernels para cálculo de similaridade molecular por ser encontrado em [Rupp & Schneider, 2010].

Em [Borgwardt et al., 2005], os autores tratam o problema de predição de função de proteínas a partir de uma representação simplificada de estruturas proteicas a partir de seus elementos de estruturas secundárias (e.g., os nós dos grafos) e seus vizinhos mais próximos (e.g., as arestas dos grafos). A partir dessa modelagem as autores utilizam uma modificação de um *kernel* baseado em *random walks* [Kashima et al., 2003], para mensurar a similaridade estrutural de proteínas.

Tarefas de predição de toxicidade, mutagênese e atividade anticâncer de pequenas moléculas são abordadas em [Swamidass et al., 2005] a partir de *kernels* baseados em caminhos (*paths*). Alguns *kernels* podem ainda sofrer de um fenômeno conhecido como *toterring* [Mahé et al., 2004]. Em um *kernel* baseado em *random walks*, por exemplo,

uma aresta pode ser visitada várias vezes. Uma vez que a similaridade entre os grafos seja calculada a partir dos *walks* comuns, uma pequena similaridade estrutural pode onerar significativamente o cálculo de similaridade.

2.2.3 Assinaturas de Grafos

Uma assinatura ou *fingerprint* pode ser vista como um conjunto de características que define, agrupa ou discrimina um conjunto de objetos dos demais. Um assinatura para grafos, por conseguinte, tem por objetivo agregar atributos de naturezas diversas que sejam comuns a grafos similares, seja do ponto de vista estrutural (perfil topológico global ou local do grafo) ou mesmo em relação a seus rótulos.

Na literatura, abordagens que não se enquadram nas restrições de uma função *kernel*, como as baseadas em descritores topológicos e de rótulos que podem ser utilizados para descrever grafos ou porções de um grafo são chamadas de *fingerprints* ou assinaturas. Tipicamente uma *fingerprint* corresponde a um vetor binário de características, ou seja, um vetor que reflete a existência ou ausência de um atributo em um determinado objeto. No presente trabalho, consideramos assinaturas vetores de atributos genéricos, que podem contemplar diferentes tipos de dados. A identificação de um conjunto de características que compõe uma assinatura adequada a tarefas como a classificação de grafos, e que assim quantifique a similaridade entre esses objetos, é um grande desafio.

Fingerprints tem sido utilizadas de forma recorrente em tarefas de triagem virtual (*virtual screening*). Os autores de [Willett, 2006] discutem acerca da utilização de fragmentos de pequenas moléculas como descritores na formação de uma *fingerprint* binária. Essa, por sua vez, é utilizada na busca de moléculas que são estruturalmente semelhantes a compostos bioativos em bases de dados de compostos. Diferentes métricas de comparação desses descritores são relatadas.

Em [Harper et al., 2004], métricas de similaridade para triagem virtual a partir de grafos reduzidos de compostos são descritas e, segundo os autores, capturam muitas características importantes relativas ao reconhecimento molecular entre ligante e receptor e permitem a composição de consultas às bases de dados que são mais flexíveis, permitindo um agrupamento de moléculas de caráter similar.

Dentre os principais trabalhos descritos na literatura, podemos ainda destacar a utilização de descritores topológicos e de rótulos para a classificação de grafos proteicos, celulares e de pequenos compostos químicos em relação à atividade mutagênica e anticâncer [Li et al., 2011]. Os autores argumentam em favor da utilização de atributos topológicos (e.g., métricas de centralidade) e de rótulos (e.g., entropia dos rótulos), o

que confere à abordagem descrita não apenas eficácia compatível em comparação com diversos métodos baseados em *kernels*, mas também um ganho substancial em termos de tempo de execução.

Recaptulando, deseja-se que uma assinatura para grafos, seja:

- **Conservada:** coerente e consistente com a definição do grupo de grafos;
- **Concisa:** de modo a facilitar sua geração, utilização, análise e armazenamento;
- **Eficiente:** espera-se que a assinatura seja escalável para bases de dados em crescimento acelerado;
- **Generalizável:** deseja-se, em última instância, que uma assinatura estrutural seja útil em vários contextos e domínios de aplicação.

À luz desses requisitos, é proposto no presente trabalho uma assinatura para grafos baseada em padrões de distâncias entre seus nós chamada *Cutoff Scanning Matrix* (CSM). O Capítulo a seguir (Capítulo 3) apresenta a assinatura proposta, os requisitos para sua aplicação e fluxo metodológico.

Capítulo 3

CSM: *Cutoff Scanning Matrix*

O termo *Cutoff Scanning* foi originalmente utilizado em [da Silveira et al., 2009] na comparação de diferentes metodologias para propspecção de contatos inter-resíduo em estruturas proteicas, considerando a representação dos resíduos via centróides (carbonos-alfa, carbonos-beta e o centro geométrico da cadeia lateral).

No trabalho supracitado, foi conduzida uma análise comparativa entre duas metodologias clássicas para aferição de contatos. A primeira é baseada em aspectos geométricos e independe de um valor de corte (*cutoff*) que define a distância máxima entre resíduos para existência do contato. Essa metodologia é baseada na geração de um triangulação de Delaunay [Delaunay, 1934], problema dual à geração de um diagrama de Voronoi [Voronoi, 1908]. A segunda abordagem define contatos a partir de uma distância de corte. Diversas distâncias foram, então, analisadas realizando uma varredura (*scanning*) dentre as distâncias possíveis. Foi realizada um comparação entre os resultados obtidos por essa variação e aqueles obtidos pelo método geométrico de modo a obter-se uma maneira robusta e confiável de definição desses contatos.

A assinatura para grafos aqui proposta utiliza o conceito de varredura de modo a gerar uma distribuição do perfil de distâncias entre os nós de um grafo. Os aspectos mais relevantes para definição da assinatura proposta, e requisitos para sua aplicação em um conjunto qualquer de grafos são:

Modelagem das entidades como grafos: diz respeito à modelagem computacional das entidades de interesse por meio da definição dos objetos (nós) e da relação entre esses (arestas), bem como da definição de rótulos e pesos a serem considerados.

Métrica de distância entre os nós do grafo: para que padrões de distâncias entre nós de um grafo sejam extraídos faz-se necessária a definição de uma métrica

de distância entre todos os pares de nós de um grafo. O exemplo mais natural (porém não limitando-se a apenas esse caso) é considerar o número (ou somatória de pesos) das arestas que ligam um par de nós por um caminho mínimo, ou seja, o custo do caminho mínimo.

Definição de parâmetros para varredura das distâncias: corresponde à definição do intervalo de distâncias a ser considerado na varredura bem como da granularidade da mesma. Maiores detalhes serão dados na Seção 3.2.

A Figura 3.1 exemplifica o processo de aplicação do conceito CSM para cálculo de assinaturas para grafos. O conceito pode ser aplicado à uma base de grafos de maneira elegante e genérica a partir do cálculo de uma matriz de distâncias. O algoritmo CSM extrai padrões de distâncias entre os nós pela varredura de um conjunto de distâncias possíveis dentro de um intervalo definido pelo usuário, formando uma distribuição acumulada dessas distâncias.

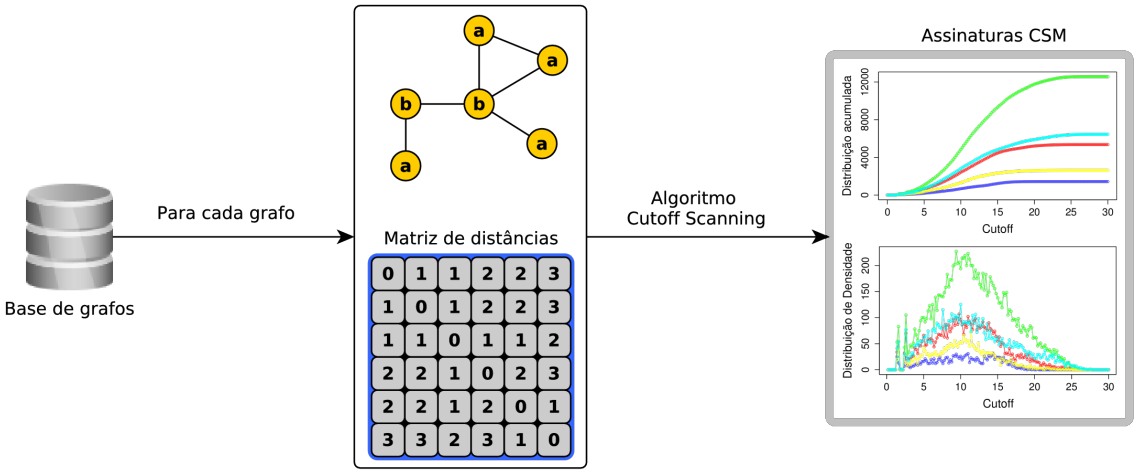

Figura 3.1. Extração de informações topológicas de grafos a partir do conceito CSM. A presente Figura ilustra o processo de extração da assinatura proposta para uma base de grafos, sendo necessária a definição de uma métrica de distância entre os nós e consequente cálculo de uma matriz de distâncias entre os pares de nós. O algoritmo CSM extrai padrões de distâncias entre os nós pela varredura de um conjunto de distâncias possíveis dentro de um intervalo definido, o que gera uma distribuição acumulada dessas distâncias.

3.1 Fluxo Metodológico

A Figura 3.2 proporciona uma visão esquemática da abordagem baseada no conceito CSM para geração, utilização e avaliação de assinaturas para grafos, que podem ser

instanciadas, por exemplo, em tarefas de classificação. As etapas são divididas em pré-processamento de dados, modelagem computacional dos grafos, geração das matrizes CSM, redução de dimensionalidade e ruído, utilização das assinaturas em tarefas de aprendizado e avaliação, validação e visualização dos resultados.

Após aquisição e filtragem dos conjuntos de dados, é realizada a modelagem computacional dos objetos da base de dados via grafos, as matrizes CSM por sua vez são geradas (mais detalhes do procedimento na Seção 3.2). Uma matriz CSM define um conjunto de vetores de atributos que são, então, processados na etapa de redução de dimensionalidade. Finalmente, a matriz CSM reduzida é submetida a diferentes algoritmos de aprendizado para avaliação. Métricas de avaliação são calculadas e comparadas para quantificar a adequação e sucesso de cada algoritmo.

O intuito da metodologia proposta é de maximizar a quantidade e diversidade de informação extraída, cobrindo um amplo espectro de valores de distâncias (*scanning*), deixando a cargo das etapas de pré-processamento e classificação a elucidação da informação significativa e descarte do ruído e da redundância dos dados.

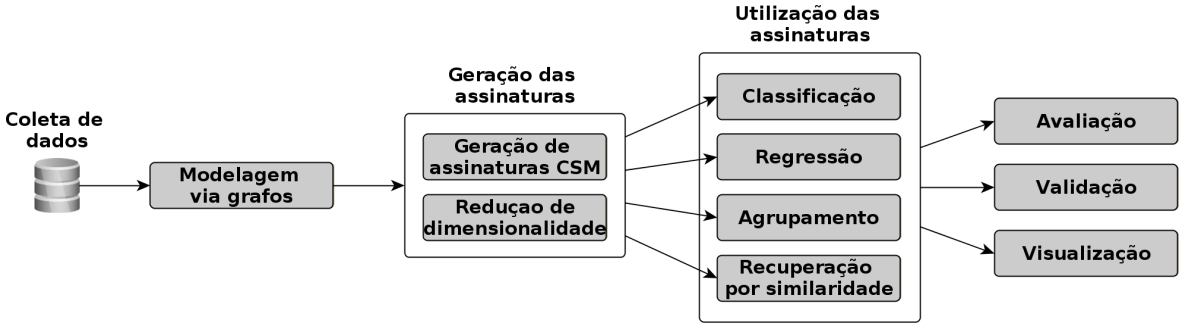

Figura 3.2. *Workflow* da metodologia proposta: é exibida uma visão esquemática da abordagem CSM para geração de assinatura para grafos aplicada à uma tarefa preditiva ou de classificação. O *workflow* é dividido em etapas de pré-processamento de dados, modelagem dos grafos, geração das matrizes CSM, redução de dimensionalidade e ruído, utilização das assinaturas em tarefas de aprendizado e avaliação, validação e visualização dos resultados.

3.2 Algoritmo CSM

Uma matriz CSM é gerada da seguinte forma: para cada grafo do conjunto de dados, geramos um vetor de atributos. Primeiro, calculamos a distância entre todos os pares de nós do grafo e definimos um intervalo de distâncias (*cutoffs*) a ser considerado e um passo. As distâncias contidas nesse intervalos são avaliadas e a frequência de pares de nós dentro dessas distâncias, para cada tipo de aresta dado por seu rótulo, computados.

Como resultado, temos uma distribuição acumulada do número de arestas no grafo para um dado intervalo de distâncias. A contagem da frequência de arestas por distância é discretizada pelos tipos de rótulos dos nós. O Algoritmo 1 exibe a função que calcula assinaturas CSM. São fornecidos como parâmetros de entrada para a função: uma base de grafos ($BaseDeGrafos$), um conjunto dos tipos possíveis de rótulos ($TipoRotulo$), distâncias mínima (D_{MIN}) e máxima (D_{MAX}) e um passo (D_{PASSO}). Os três últimos parâmetros formam o intervalo de distâncias considerado, bem como a granularidade da variação de distâncias.

Algorithm 1 Cálculo de assinaturas CSM

```
1: function CSM(BaseDeGrafos, TipoRotulo, D_MIN, D_MAX, D_PASSO)
2:     for all grafo i ∈ (BaseDeGrafos) do
3:         j = 0
4:         distMatriz ← calculaDistancias(grafo)
5:         for dist ← D_MIN; até D_MAX; passo D_PASSO do
6:             for all tipo ∈ (TipoRotulo) do
7:                 CSM[i][j] ← obtemFrequencia(distMatriz, dist, tipo)
8:                 j++
9:     return CSM
```

3.3 Redução de Ruído e Dimensionalidade

De modo a reduzir o ruído inerente aos dados gerados e também o custo de execução dos algoritmos de classificação, tanto em termos de requisitos de memória quanto de tempo, utilizamos a Decomposição em valores singulares (SVD) para reduzir dimensionalidade.

A Decomposição em valores singulares é uma técnica de álgebra linear amplamente utilizada em tarefas de redução de dimensionalidade. No presente trabalho usamos a SVD também para reduzir ou eliminar o ruído inerente das assinaturas geradas e, por consequência, melhorar a eficácia e reduzir processamento de custo dos algoritmos de classificação, em termos de tempo de execução e requisitos de memória.

A SVD estabelece relacionamentos relevantes e não-triviais entre grupos de elementos [Eldén, 2007, 2006; Berry et al., 1995]. A lógica por trás da SVD diz que uma matriz A, composta por m linhas e n colunas, pode ser representada por um conjunto de matrizes derivadas [Berry et al., 1995] que permitem uma representação numérica diferente dos dados sem perda de significado semântico. Em outras palavras,

$$A = TSD^T$$

Onde T é uma matriz ortonormal de dimensões m x m, S é uma matriz diagonal de dimensões m x n e D é uma matriz ortonormal de dimensões n x n. Os valores da diagonal de S são os valores singulares de A, e eles são ordenados em ordem decrescente de significância.

Quando consideramos somente um subconjunto de valores singulares de tamanho $k < p$, onde p é o posto de A, podemos obter A_k, uma matriz aproximada da matriz original A:

$$A \approx A_k = T_k S_k D_k^T$$

Assim, a aproximação dos dados depende do número de valores singulares utilizados [del Castillo-Negrete et al., 2007]. Nesse caso, o número k de valores singulares é também o posto da matriz A_k. A possibilidade de extração de informação relevante a partir de um volume menor de dados é parte do sucesso dessa técnica, ao passo que ela pode permitir a compressão/descompressão dos dados dentro de um tempo de execução factível, tornando a análise de grandes volumes de dados viável [del Castillo-Negrete et al., 2007]. Uma base de dados representada por um conjunto menor de valores singulares em relação à base original tende a gerar certos agrupamentos que não seriam notados se utilizássemos os dados originais [Berry et al., 1995]. Esse agrupamento pode explicar porque grupos derivados da SVD podem expor relacionamentos não-triviais entre os objetos da base original [Deerwester et al., 1989]. Nesse trabalho, utilizamos A_k, aproximação da matriz original com posto k, mas somente com dois componentes da decomposição, que geram a matriz V_k como descrito em [Eldén, 2006]:

$$A_k = T_k S_k D_k^T = T_k (S_k D_k^T) = T_k V_k$$

A justificativa para somente utilizar V_k é que os relacionamentos entre as colunas de A_k são preservadas em V_k porque T_k é uma base para as colunas de A_k.

Avaliamos a distribuição dos valores singulares obtidos nos experimentos em um esforço de melhor caracterizar um bom valor de corte para reduzir o número de dimensões da base sem perda de informação relevante. Deseja-se encontrar um valor de corte para aproximação e recomposição dos dados de modo que a contribuição dos demais valores singulares na descrição da matriz original é insignificante, e portanto pode ser vista como ruído. Esse passo, bem como toda geração de gráficos, foi desenvolvida via *scripts* na linguagem de programação R.

3.4 Tarefas de Classificação

Diversos algoritmos de aprendizado supervisionado foram empregados de modo a avaliar o poder de predição das assinaturas baseadas no conceito CSM. O funcionamento e principais características de alguns desses algoritmos são descritos a seguir. Mais informações sobre outros algoritmos utilizados podem ser encontradas em [Breiman, 2001; Lewis, 1998].

- **K-Nearest Neighbor (KNN)** [Cover & Hart, 1967]:

  Em linhas gerais, o KNN delimita o grupo de k objetos (vizinhos) no conjunto de treino que estão mais próximos de um objeto de teste, de acordo com uma métrica de distância ou similaridade, e assinala um rótulo de classe ao objeto de teste baseado na classe predominante na sua vizinhança (voto da maioria).

  Classificadores KNN são baseados em exemplos e fazem parte de um conjunto de algoritmos de aprendizagem dita *lazy* no sentido de que um modelo não é construído/treinado explicitamente, postergando o custo computacional para a fase de classificação. Muito embora classificar um objeto desconhecido em geral requer o cálculo de distância entre esse e todos os outros objetos no conjunto de treino, a utilização de algoritmos de busca de vizinhos apropriados torna o KNN tratável até mesmo para bases de dados de larga escala.

  Levando em consideração um compromisso com eficiência, o KNN foi a abordagem de classificação em geral utilizada em experimentos de larga escala.

- **Regressão Logística Multinomial** [Landwehr et al., 2005]:

  Um classificador construído sobre uma regressão logística multinomial considera uma variável de classe C, também chamada variável dependente, assumindo valores no intervalo $1, ..., N_c$, onde N_c denota o número de classes. Em seguida, um modelo de regressão logística é construído para representar as probabilidades de classe $p(C = c|X = x)$, onde X é um vetor de atributos independentes para uma instância, para N_c classes. Dadas as estimativas para as probabilidades de classes, instâncias desconhecidas podem ser classificadas por:

  $$j^* = argmax_j p(C = c|X = x) \tag{3.1}$$

  A regressão logística estima tais probabilidades utilizando uma função linear em x garantindo que essas permaneçam no intervalo $[0, 1]$ e que sua soma seja 1. O

modelo é especificado em termos de $N_c - 1$ classes e a classe base N_c através da função *logit*:

$$log\frac{p(C = c|X = x)}{p(C = N_c|X = x)} = \beta_c x, c = 1, ..., N_c - 1 \tag{3.2}$$

Onde β_c é um vetor de parâmetros estimado dos dados de treino para cada classe c, ou de forma equivalente:

$$p(C = c|X = x) = \frac{e_j^{\beta_c}}{1 + \sum_{l=1}^{N_c-1} e_j^{\beta_c}}, c = 1, ..., N_c - 1 \tag{3.3}$$

$$p(C = N_c|X = x) = \frac{1}{1 + \sum_{l=1}^{N_c-1} e_j^{\beta_c}} \tag{3.4}$$

3.5 Metodologia de Avaliação dos Resultados

No que diz respeito a tarefas de classificação, uma série extensa de experimentos foram projetados para avaliar a eficácia da assinatura CSM como fonte de informação.

Cada experimento de classificação gera uma matriz de confusão [Provost & Kohavi, 1998] como resultado, onde cada coluna representada as instâncias em uma classe predita, enquanto cada linha representa as instância em uma classe real. A partir de uma matriz de confusão são extraídos, para cada classe sendo considerada, valores de verdadeiro-positivos (VP), falso-positivos (FP),verdadeiro-negativos (VN) e falso-negativos (FN). A Tabela 3.1 exemplifica o conceito de matriz de confusão para uma tarefa de classificação binária (duas classes consideradas, p′ e n′, com P′ e N′ entidades, respectivamente).

O desempenho dos classificadores foi avaliado, entre outras alternativas, por meio de diversas métricas. Abaixo temos uma sumarização dos conceitos e métricas de avaliação utilizados.

- **Verdadeiro-positivos (VP)** é um item corretamente predito como sendo da classe positiva.
- **Falso-positivos (FP)** é um item incorretamente predito como sendo da classe positiva
- **Verdadeiro-negativos (VN)** é um item corretamente predito como sendo da classe negativa.

Tabela 3.1. Exemplo de matriz de confusão para classificação binária: algumas métricas consideradas são derivadas a partir dessa matriz.

		Classe predita		
		p	**n**	**total**
Classe real	**p′**	Verdadeiro-positivos	Falso-negativos	P′
	n′	Falso-positivos	Verdadeiro-negativos	N′
	total	P	N	

- **Falso-negativos (FN)** é um item incorretamente predito como sendo da classe negativa.

- **Taxa de VP:** é a razão de itens corretamente preditos como sendo da classe positiva.

- **Taxa de FP:** é a razão de itens incorretamente preditos como sendo da classe positiva.

- **Precisão:** é o número de verdadeiro-positivos dividido pelo total de elementos preditos como pertencentes à classe positiva ou

$$\frac{VP}{VP + FP}$$

.

- **Revocação:** é a fração de instâncias relevantes recuperadas, ou seja, o número de verdadeiro-positivos divido pelo tamanho da classe positiva:

$$\frac{VP}{VP + FN}$$

- **F-measure (F1-score):** é a média harmônica entre a precisão e a revocação ou

$$\frac{2 \times precisao \times revocacao}{precisao + revocacao}$$

- **Curva ROC:** o termo curva ROC refere-se à *Receiver Operating Characteristic Curve* e é derivada de uma matriz de confusão tradicional. Curvas ROC são técnicas para analisar o desempenho de classificadores [Swets, 1988]. Um gráfico ROC é um gráfico com a taxa de falso-positivo no eixo-X e a taxa de verdadeiro-positivo representada no eixo-Y. No espaço ROC, o ponto (0,1) indica um classificador perfeito: ele classifica todos os casos positivos e negativos corretamente. O ponto (0,1) indica que a taxa de falso-positivos é 0, e a taxa de verdadeiro-positivos é de 1. O ponto (0,0) representa um classificador que prevê todos os casos como sendo negativos, enquanto que o ponto (1,1) corresponde a um classificador que prevê todos os casos como sendo positivos. O ponto (1,0) representa um classificador que é incorreto para todas as instâncias. Essas curvas fornecem uma ferramenta visual para examinar o compromisso entre a capacidade de um classificador de identificar corretamente os casos positivos e o número de casos negativos que foram incorretamente classificados.

- **Área sobre a curva(AUC):** é a área sob a curva ROC. Pode ser utilizada como medida de precisão em muitas aplicações. Os valores de AUC vão de 0 a 1 e um classificador aleatório teria uma AUC de 0,5.

- **Acurácia:** é a fração de itens cuja classe foi devidamente assinalada ou

$$\frac{VP + VN}{VP + VN + FP + FN}$$

- **Validação cruzada:** é uma abordagem estatística tradicional utilizada para estimar o desempenho de modelos preditivos. Ela consiste no particionamento dos dados de entrada em n conjuntos. $n - 1$ partes serão utilizadas para a construção do modelo e denominamos tais partes como conjunto de treino. A parte restante é utilizada então para avaliar a adequação do modelo, parte essa denominada conjunto de teste. Esse procedimento é repetido n vezes variando sistematicamente os conjuntos de treino e teste e as médias das métricas de qualidade dos classificadores para as n execuções são calculadas. Nesse caso, dizemos que executamos a validação em *n-partições*.

Em alguns cenários, as variações em precisão ($\Delta Precisao$) também foram utilizados para medir o ganho obtido pelo processamento com a SVD, e variações de revocação ($\Delta Revocacao$) foram avaliadas para comparar os resultados com aqueles obtidos por métodos competidores, utilizando bases desses trabalhos.

Também correlacionamos a taxa de sucesso obtida pelos classificadores, de acordo com diversas métricas, com o número de valores singulares considerados e a comparamos com os resultados utilizando a matriz CSM completa.

3.6 Aplicação do Conceito CSM

Na presente tese, o conceito CSM é instanciado em três classes de problemas:

Anotação Automática: (Capítulo 4) Predição de função (números EC) e classificação estrutural proteica (classificação nos vários níveis do SCOP) baseada em assinaturas CSM para grafos proteicos em nível de resíduo.

Predição de Ligantes: (Capítulo 5) Predição de ligantes para *pockets* proteicos com base em assinaturas CSM para grafos proteicos em nível atômico.

Predição de Toxicidade, Mutagênese e Atividade Anti-câncer: (Capítulo 6) Predição de características de pequenas moléculas com base em assinaturas CSM geradas para grafos moleculares compostos dos átomos e das ligações covalentes que formam os compostos.

Capítulo 4

Anotação Automática

A anotação automática proteica é uma tarefa desafiadora ao passo que remete à questões complexas como a elucidação e entendimento do relacionamento entre a estrutura e função de proteínas. Além disso, com o grande número de estruturas proteicas sendo resolvidas e disponibilizadas por diversas iniciativas como as de genômica estrutural [Chandonia & Brenner, 2006], tem sido criada uma demanda cada vez maior por paradigmas, modelos e metodologias eficientes e escaláveis para essa tarefa.

Um problema inicial enfrentado ao se lidar com a anotação automática, particularmente com função proteica, é a definição do escopo de função, como bem ressaltado por [Punta & Ofran, 2008]. A predição de função de proteínas pode ser entendida sob diferentes perspectivas. Pode ser tratada como uma tarefa de predição do processo celular no qual a proteína está envolvida, sua atividade enzimática ou mesmo um papel fisiológico por ela desempenhado. Por exemplo, a atividade enzimática pode ser descrita por um Enzyme Comission number (EC), enquanto o papel fisiológico pode estar relacionado com a sua sub-localização celular. Neste trabalho, o aspecto da função proteica considerado é a atividade enzimática. Entretanto, o estudo pode ser estendido, sem perda de generalidade, para outros atributos funcionais, como a anotação de termos do Gene Ontology (GO) [Ashburner et al., 2000].

Muito embora a função não possa ser diretamente implicada a partir de uma enovelamento específico adotado por uma proteína, dados estruturais podem ser utilizados para detectar proteínas com função similar cujas sequências divergiram ao longo da evolução [Lee et al., 2007]. Uma possível estratégia para abordar esse problema é a definição de *assinaturas estruturais*, que correspondem a um conjunto de características que são aptas a identificar univocamente um determinado enovelamento e a natureza das interações que esse pode estabelecer com outras proteínas ou ligantes.

Esse conjunto de características é uma representação concisa de estruturas proteicas, e acreditamos que sua descoberta seja um importante marco no campo da predição de função proteica, sendo um passo além dos métodos baseados em homologia de sequência. Deseja-se que uma assinatura estrutural para uma família de proteínas, estruturalmente ou funcionalmente relacionadas, seja:

- **Conservada:** coerente e consistente com a definição do grupo de proteínas similares estruturalmente ou de mesma função;
- **Concisa:** de modo a facilitar sua geração, utilização, análise e armazenamento;
- **Eficiente:** espera-se que a assinatura seja escalável para bases de dados em crescimento acelerado (requisito motivado por iniciativas de genômica estrutural [PSI, 2011]);
- **Generalizável:** deseja-se, em última instância, que uma assinatura estrutural seja útil em vários contextos e domínios de aplicação (como em diversas tarefas de anotação automática).

Neste trabalho, investigamos um tipo especial de atributo que pode fazer parte de assinaturas estruturais: o padrão de distâncias (ou contatos) inter-resíduo. Proteínas com diferentes enovelamentos e funções apresentam uma diferença significativa na distribuição de distâncias entre seus resíduos como consequência da rede de interação e empacotamento subjacente, que é fundamental na definição do enovelamento [Soundararajan et al., 2010]. Utilizamos o conceito CSM de modo a modelar estruturas proteicas como grafos e extrair padrões de distâncias inter-resíduos que sejam específicos de famílias estruturais ou funcionais e aplicáveis como evidências em tarefas de classificação;

A motivação para o uso desse tipo de informação reside na ideia de que os resíduos de proteínas com diferentes enovelamentos e funções estejam arranjados no espaço de maneira significativamente diferente, e por outro lado, similaridade de estrutura seja refletida nessas distâncias, informações que são capturadas pelo CSM.

No entanto, essa abordagem gera um grande volume de dados que, por construção, podem apresentar redundâncias ou ruído. Assim, após a geração desse dado estrutural, como etapa de pré-processamento, aplicamos uma técnica de redução de ruído e dimensionalidade de modo a aumentar a eficácia e conferir escalabilidade à metodologia. Nesse caso utilizamos a Decomposição em Valores Singulares (SVD). Finalmente, a matriz processada é, então, submetida a diferentes algoritmos de classificação.

Tendo em mente essas considerações, demonstramos que padrões derivados de CSMs podem ser utilizados de forma eficaz e eficiente na predição de função e classificação estrutural automática. Os primeiros resultados mostram que, para o caso de predição de função enzimática, o método proposto atingiu 95,1% de precisão e revocação para um conjunto de enzimas com seus respectivos números EC (os 950 números EC mais populosos em termos de estruturas resolvidas). Adicionalmente, considerando os níveis hierárquicos da classificação SCOP [Murzin et al., 1995], fomos capazes de atingir índices de precisão e revocação de até 95,4%. Em comparação com métodos estado-da-arte nesse contexto, como o proposto por Jain & Hirst, utilizando uma base bastante similar como entrada, nossa metodologia apresentou resultados mais robustos e homogêneos, com uma precisão média um pouco abaixo da atingida por aqueles autores mas com uma melhora expressiva na revocação. Maiores detalhes são apresentados na Seção 4.3.

Dentre as principais contribuições do presente trabalho está a proposta de uma nova abordagem para busca de assinaturas estruturais capaz de detectar invariantes de famílias proteicas. Essa abordagem mostrou-se útil em tarefas de classificação estrutural e inferência de função proteica e pode ainda contribuir direta ou indiretamente, como subsídio, para o desenvolvimento de outras aplicações como estudos de interação proteína-proteína, interação ligante-ligante, que consequentemente podem levar a descoberta de novos alvos terapêuticos e fármacos. Mostramos ainda que nossa metodologia é, em geral, independente do classificador utilizado, proporcionando resultados comparáveis para diferentes estratégias de classificação.

4.1 Modelagem Computacional

Na presente Seção descrevemos a modelagem computacional realizada que corresponde à representação de estruturas proteicas como grafos de interações entre resíduos de aminoácidos.

- **Grafo de contatos inter-resíduos:** grafo simples, completo (clique), não-direcionado, ponderado, não-rotulado;
- **Nós:** correspondem a um centroide representativo do resíduo. Nesse caso, utilizamos o carbono-α;
- **Arestas:** todos os nós são conectados, formando um grafo completo ou clique;
- **Peso das arestas:** corresponde à distância Euclidiana entre os resíduos de aminoácidos (nós), representados pelo carbono-α.

Em suma, as estruturas proteicas foram modeladas como cliques onde, os nós representam os aminoácidos e o peso das arestas a distância entre esses. A Figura 4.1 exemplifica a modelagem realizada. São exibidos grafos obtidos pela imposição de uma distância máxima ao comprimento das arestas a partir do grafo completo original. A contagem de arestas de cada um desses, representa uma dimensão na assinatura CSM gerada.

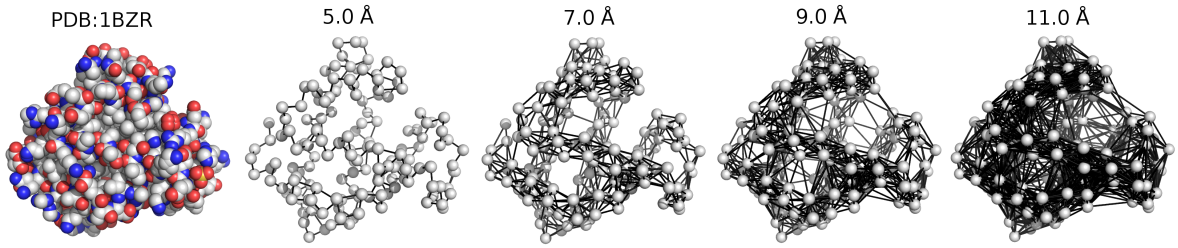

Figura 4.1. Grafos de contatos inter-resíduos. Na figura é exibida à esquerda uma mioglobina de baleia (código PDB:1BZR) considerando a visualização dos átomos em *spacefill*. Em seguida são exibidos os grafos de contatos inter-resíduo para essa proteína considerando como nós (centroide dos resíduos) os carbonos-α. São exibidos grafos cujas arestas foram definidas a partir de diferentes distâncias máximas de corte (*cutoffs*), obtidas a partir do grafo completo inicial. A contagem de arestas de cada um desses grafos representa uma dimensão na assinatura CSM.

4.1.1 Métodos

A Figura 4.2 proporciona uma visão esquemática da abordagem baseada em CSM para geração e avaliação de assinaturas estruturais, aqui utilizada em tarefas de predição de função e classificação estrutural, a qual pode ser divida em etapas de pré-processamento, geração das matrizes CSM, redução de ruído e dimensionalidade (via SVD), avaliação e validação (por tarefas de classificação).

Após aquisição e filtragem dos conjuntos de dados (montados tanto para predição de função quanto classificação estrutural), as matrizes CSM são geradas. Uma matriz CSM define um conjunto de vetores de atributos que são, então, processados na etapa de redução de dimensionalidade. Finalmente, a matriz CSM reduzida é submetida para diferentes algoritmos de classificação para avaliação. Métricas como precisão e revocação são calculadas e comparadas para quantificar o poder de predição dos classificadores.

O intuito da metodologia proposta é de maximizar a quantidade e diversidade de informação extraída, cobrindo um amplo espectro de valores de distâncias (*scanning*),

deixando a cargo das etapa de pré-processamento e classificação a elucidação da informação significativa e descarte do ruído e da redundância dos dados.

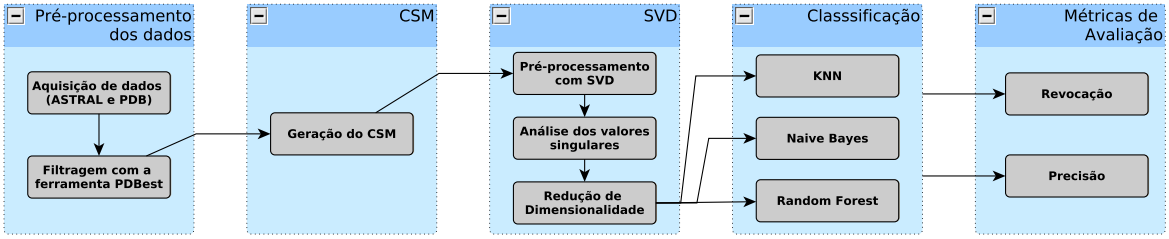

Figura 4.2. *Workflow* da metodologia proposta: é exibida uma visão esquemática da abordagem CSM empregada na predição de função e classificação estrutural proteica. O *workflow* é dividido em etapas de pré-processamento de dados, geração das matrizes CSM, redução de dimensionalidade e ruído (via SVD) e avaliação dos resultados (via tarefas de classificação).

No presente trabalho, utilizamos a abordagem baseada em *cutoff scanning* em tarefas de classificação, o que é a base das matrizes CSM. A motivação em favor do uso desse tipo de informação reside no fato de proteínas com diferentes enovelamentos e funções apresentarem diferenças significativas na distribuição de distâncias entre seus resíduos. Por outro lado, pode-se esperar que similaridade estrutural seja refletida nesse padrão de distâncias, informação que é capturada em CSMs.

Uma matriz CSM é gerada da seguinte forma: para cada proteína do conjunto de dados, geramos um vetor de atributos. Primeiro, calculamos a distância Euclidiana entre todos os pares dos centróides que representam os resíduos e definimos um intervalo de distâncias (*cutoffs*) a ser considerado e um passo. As distâncias contidas nesse intervalos são avaliadas e a frequência de pares de resíduos dentro dessas distâncias, cada um representado por seu centroide (carbonos-α, por exemplo), computados. O Algoritmo 2 exibe a função que calcula a CSM. Nesse trabalho utilizamos os carbonos-α como centróides representantes dos resíduos de aminoácidos.

Inicialmente variamos a distância de corte de 0,0 Å a 30,0 Å, com um passo de 0,2 Å, o que gerou para cada proteína um vetor com 151 valores. Juntos, esses vetores compõem a matriz CSM. Resumidamente, cada linha da matriz representa uma proteína, e cada coluna representa a frequência de pares de resíduos a uma distância menor ou igual a uma certa distância. De forma alternativa, essa frequência pode ser vista como o número de contatos na proteína para uma dada distância de corte ou o número de arestas de um grafo de contatos gerado por esses *cutoffs*. Essa etapa foi implementada utilizando a linguagem *scripting* Perl.

A motivação para o uso das CSMs advém das diferenças entre as distribuições

Algorithm 2 Cálculo do CSM

Entrada: $Proteinas, Dist_{MIN}, Dist_{MAX}, Dist_{PASSO}$
Saida: CSM

```
1: function geraCSM
2:     for all proteína i ∈ (Proteinas) do
3:         j = 0
4:         Calcula as distâncias entre todos os pares de centróides da proteína
5:         for dist ← Dist_MIN; to Dist_MAX; passo Dist_STEP do
6:             CSM[i][j] ← Obtém frequência de pares centróides
7:             a uma distância menor ou igual a dist
8:             j++
9:     retorna CSM
```

de contatos entre proteínas de diferentes classes estruturais, como pode ser visto na Figura 4.3, que mostra a distribuição de densidade normalizada da contagem de arestas por *cutoff* para proteínas de diferentes classes do SCOP, a saber: *all alpha*, *all beta*, *alpha+beta* e *alpha/beta*. É possível notar que as diferenças aparecem em diferentes intervalos de distâncias. Por exemplo, os primeiros picos para a proteína *all alpha* indicam a primeira camada de contatos de suas hélices e as diferenças em distâncias maiores podem ocorrer devido ao diâmetro das proteínas. Ressaltamos que essas variações na contagem de arestas não são unicamente um fenômeno da composição da estrutura secundária das proteínas mas um fenômeno do próprio empacotamento proteico. É importante ainda explicar a variação de *cutoffs*. Essa variação (*scanning*) agrega informações relevantes relativas ao empacotamento e captura, implicitamente, a forma da macromolécula. Acreditamos que bolsões na superfície e mesmo cavidades internas sejam contabilizadas por esse dado estrutural.

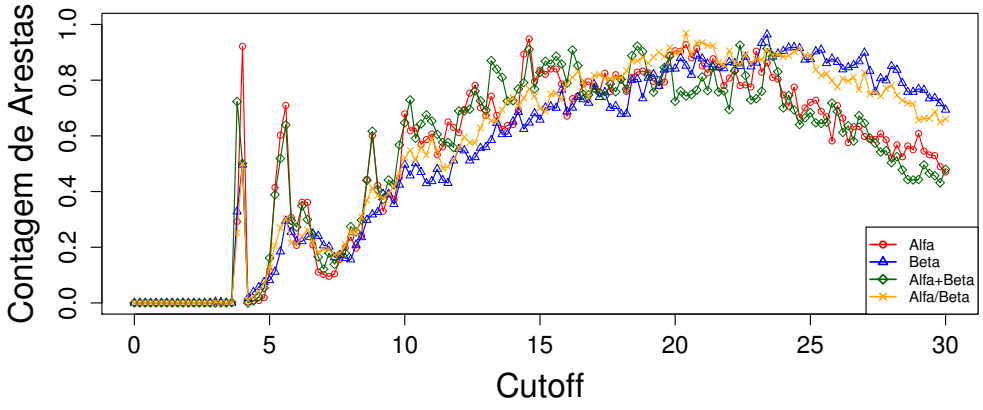

Figura 4.3. Distribuição de densidade de vetores de atributos para proteínas de diferentes classes do SCOP: cada curva representa os valores médios para dez representantes de cada classe selecionados ao acaso.

Outro exemplo de distribuição de contatos é exibido na Figura 4.4. Três proteínas com estruturas bastante dissimilares foram selecionadas (uma globina, PDB:1A6M; uma porina, PDB:2ZFG; e uma molécula de colágeno, PDB:1BKV), e a topologia do grafo de contatos obtida para diferentes distâncias de corte são exibidas (6,0 Å, 9,0 Å e 12 Å). Adicionalmente, as distribuições acumulada e de densidade dos vetores de atributos de CSM para esses representantes foram calculadas. Podemos perceber nesses exemplos que a diferença expressiva no formato das proteínas é contabilizada pelo CSM. No perfil de contatos, os picos indicam alta frequência de padrões de distância recorrentes que estão presentes nas estruturas. Um pico mais alto em torno de 3,8-4,0 Å é evidência de contatos entre carbonos-α consecutivos. Essas distâncias tendem a ser independentes da classe estrutural da proteína em virtude da propriedade planar que caracteriza a ligação peptídica que faz intermédio entre dois carbonos-α contíguos na cadeia principal.

Em adição a esse padrão, em proteínas ricas em hélices, encontraremos picos sugestivos entre 5,0 Å e 7,0 Å, representando principalmente as distâncias recorrentes entre as posições locais dos carbonos-α (em sequência) $(i, i+2)$, $(i, i+3)$ e $(i, i+4)$ que compõem as voltas de uma hélice, além de alguns contatos não-locais (em relação à sequência primária). Por outro lado, em proteínas ricas em fitas-β, picos importantes serão notados em torno de 6,0 Å e 5,0 Å, que se referem não somente a distâncias de carbonos-α locais em sequência mas também contatos entre fitas. Isso sugere que o CSM está manipulando informações de dois níveis estruturais essenciais: contatos locais e não-locais relevantes. Podemos ver também que a forma da proteína interfere diretamente na rede de contatos subjacente, o que reflete no enovelamento proteico, como apontado por [Soundararajan et al., 2010]. Essas propriedades fazem do CSM uma fonte rica e importante de informação ao lidar com problemas como predição de função e classificação estrutural.

4.1.2 Tarefas de classificação

Uma série extensa de experimentos foram projetados para avaliar a eficácia da assinatura CSM como fonte de informação nas tarefas de classificação estrutural e predição de função proteica.

Para as tarefas de classificação, a ferramenta Weka [Hall et al., 2009], *developer version 3.7.2* foi utilizada. Três algoritmos de classificação foram utilizados, e seus desempenhos comparados: KNN (*K-Nearest Neighbors*) [Cover & Hart, 1967], Random Forest [Breiman, 2001] e Naive Bayes [Lewis, 1998]. Os parâmetros de entrada dos algoritmos foram variados, quando aplicável, e os melhores resultados computados.

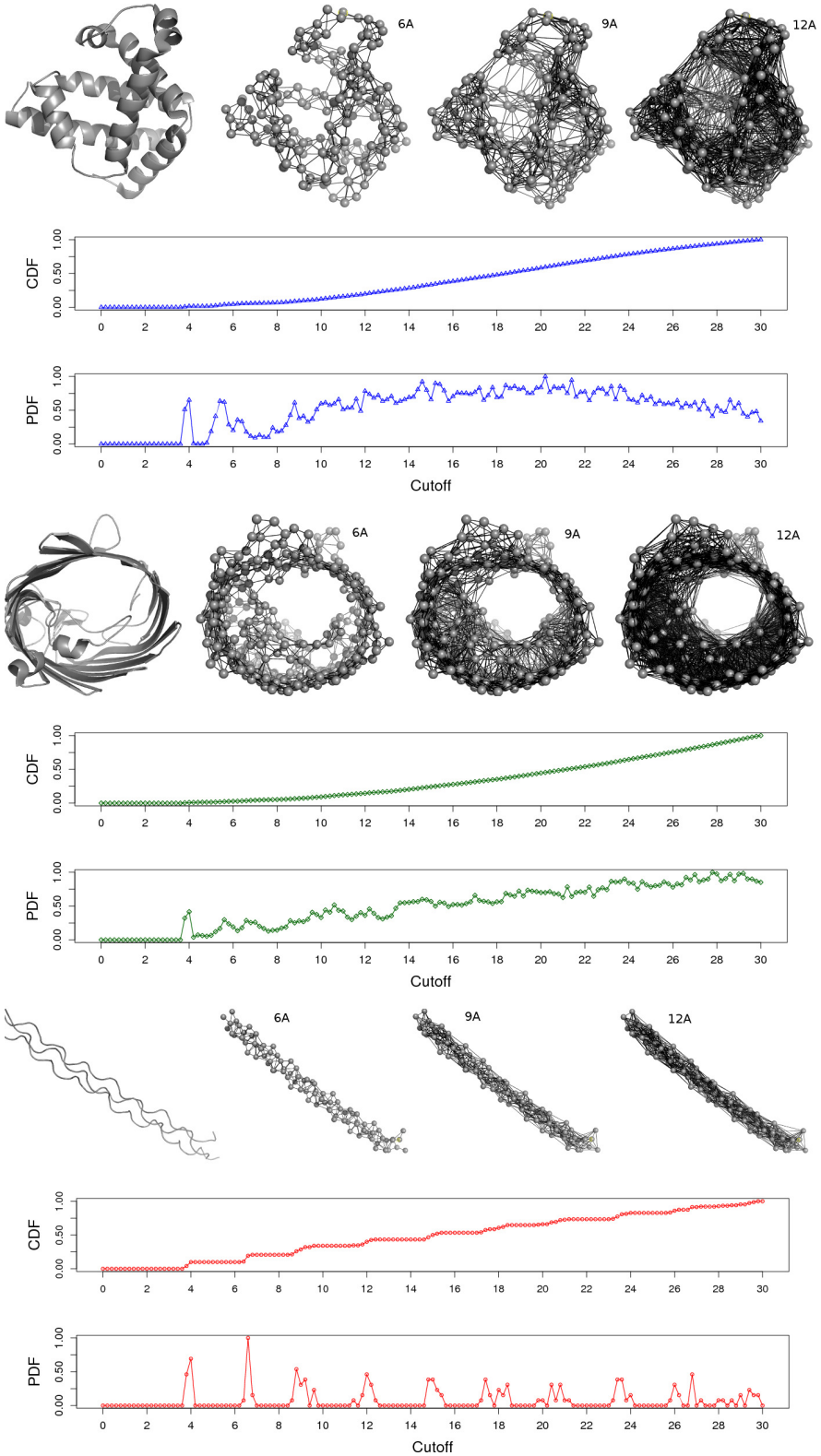

Figura 4.4. Topologia dos grafos de contatos de três estruturas distintas (de cima para baixo: globina, porina e colágeno) para diferentes valores de *cutoff*: 6,0Å, 9,0Å e 12,0Å. A contagem de arestas para cada grafo representante denota uma entrada no vetor de atributos gerado pelo CSM. As distribuições cumulativas e de densidade normalizadas dos vetores de atributos de cada proteína são exibidas.

Em todos os cenários foi aplicada validação cruzada de 10 partições.

Esses algoritmos de classificação foram selecionados por serem bem estabelecidos na literatura e por pertencerem a diferentes paradigmas de classificação (um baseado em exemplos (*lazy learning*), outro baseado em árvores de decisão e outro probabilístico, respectivamente). Não utilizamos o algoritmo considerado estado-da-arte em classificação, *Support Vector Machine* (SVM Cortes & Vapnik [1995]), dado que o tempo de execução necessário para executá-lo em bases de grande porte, como as utilizadas nesse trabalho, o torna praticamente inviável.

O desempenho dos classificadores foi avaliado por meio de métricas como *Precisão*, *Revocação*, *F1-score* (a média harmônica entre precisão e revocação) e a área sob a curva ROC (Area Under ROC Curve - AUC). As variações em precisão ($\Delta Precisao$) foram utilizados para medir o ganho obtido pelo processamento com SVD, e variações de revocação ($\Delta Revocacao$) foram avaliadas para comparar os resultados com aqueles obtidos por [Jain & Hirst, 2010], utilizando uma base derivada desse trabalho.

Também correlacionamos a precisão obtida pelos classificadores com o número de valores singulares considerados e a comparamos com os resultados utilizando a matriz CSM completa.

4.1.3 Bases de Dados

Nossas bases de dados consistem de estruturas proteicas disponíveis no Protein Data Bank [Berman et al., 2002]. Os domínios cobertos pelo SCOP (versão 1.75) foram obtidos através do ASTRAL [Brenner et al., 2000]. Conjuntos de estruturas foram organizadas de acordo com o propósito do experimento, a saber, predição de função e classificação estrutural. Para estruturas resolvidas via NMR (Ressonância Nuclear Magnética), somente foi considerado o primeiro modelo. As cadeias foram separadas em diferentes arquivos e as coordenadas dos carbonos-α extraídas com auxílio da ferramenta PDBest [Pires et al., 2007].

A primeira base de dados refere-se a um padrão-ouro de superfamílias de enzimas que utilizam mecanismos distintos para executar suas respectivas funções [Brown et al., 2006]. Consideramos *seis superfamílias* (*amidohydrolase*, *crotonase*, *haloacid dehalogenase*, *isoprenoid synthase type I* e *vicinal oxygen chelate*), compreendendo 47 famílias distribuídas em 566 diferentes *cadeias*.

A segunda base contem enzimas com seus números EC. Consideramos os 950 números EC mais populosos em termos de estruturas disponíveis, com pelo menos 9 representantes por classe, totalizando 55.474 cadeias, que cobrem aproximadamente 95% de todas as enzimas revisadas do Uniprot [Consortium, 2010], i.e., as anotações

experimentalmente validadas dessa base.

A terceira base, montada com o intuito de classificação estrutural, é originada da versão 1.75 do SCOP. Selecionamos todos os identificadores PDB cobertos pelo SCOP com pelo menos 10 resíduos de aminoácidos e 10 representantes por nó na hierarquia de classificação do SCOP. Esses identificadores representam no total 110.799, 108.332, 106.657 e 102.100 domínios nos níveis *classe*, *enovelamento*, *super-família* e *família*, respectivamente. É importante enfatizar que esse é um conjunto de dados consideravelmente grande e que não encontramos nenhum outro trabalho na literatura que utiliza uma base tão completa em tarefas de classificação estrutural.

O último conjunto de dados foi derivado de [Jain & Hirst, 2010] para comparação de resultados em classificação estrutural. Selecionamos todos os domínios descritos nos arquivos adicionais desse artigo com um mínimo de 10 representantes por nó na hierarquia do SCOP. Não foi possível identificar exatamente os domínios utilizados a partir desses arquivos e somente aqueles pares com identidade de sequência inferior a 35% foram retidos. É fundamental ressaltar que o trabalho de Jain & Hirst somente contempla proteínas pequenas, com 3, 4, 5 ou 6 elementos de estrutura secundária.

4.2 Trabalhos Relacionados

4.2.1 Assinatura Estrutural

Uma possível definição para assinaturas estruturais seria: representações concisas das características das proteínas de mesmo enovelamento. São um conjunto de características inerentes às sequências que são determinantes do seu enovelamento e função [Melo-Minardi, 2008]. Em outras palavras, um conjunto de atributos que identifica univocamente uma família de proteínas e a distingue das demais. Desse modo, deseja-se entender que variantes mas, principalmente, que invariantes estão envolvidos na formação de uma família de proteínas funcionalmente ou estruturalmente relacionadas e quais invariantes formariam um conjunto que melhor discriminasse uma família proteica das demais.

A busca por tais características, que é uma passo além de uma simples comparação de estruturas por meio de uma métrica de similaridade, tem sido abordada sob diversas perspectivas. Uma delas é a extração automática de regras, via *Inductive Logic Programming* (ILP) [Turcotte et al., 2001], técnica limitada devido a necessidade de se pré-definir o conjunto de predicados de primeira ordem da base de conhecimento na indução de hipóteses. Encontramos também, trabalhos que utilizam *Cadeias Ocultas de Markov* (HMM) [Kersting et al., 2003] sobre dados de estruturas secundária.

Existem ainda abordagens baseadas na busca por regiões conservadas, ou motivos estruturais [Richards & Kundrot, 1988; Hutchinson & Thornton, 1996; Koehl, 2001; Ausiello et al., 2008].

No trabalho aqui proposto exploramos uma nova componente de assinaturas estruturais, formada por padrões de distâncias entre nós de grafos compostos por resíduos de aminoácidos.

4.2.2 Distâncias Inter-resíduo

A distância Euclidiana inter-resíduo tem sido estudada e aplicada em diversos contextos em Bioinformática. São estudadas na definição de contatos [da Silveira et al., 2009], particularmente no estudo das relações entre a rede formada pelos contatos e o enovelamento proteico [Soundararajan et al., 2010], a caracterização topológica dessas redes [Altigan et al., 2004] e na mineração de subgrafos frequentes [Huan et al., 2004], na mineração de padrões em mapas de contatos [Hu et al., 2002] ou em sua utilização para cálculo de similaridade entre proteínas [Melo-Minardi et al., 2006].

Existem ainda, relatos na literatura da proposta de métodos para detecção de similaridades locais em estruturas proteicas [Zemla, 2003] ou mesmo de funções de pontuação para a avaliação da qualidade de *templates* de estruturas [Zhang & Skolnick, 2004].

4.2.3 Predição de Função e Classificação Estrutural Proteica

Em um contexto de geração massiva de dados biológicos, novos paradigmas, modelos e metodologias para anotação automática precisam ser investigados. Uma vez que estrutura e função proteicas são mais conservadas que a sequência [Chothia & Lesk, 1986], a identificação de similaridades entre novas sequências e estruturas conhecidas poderia melhorar significativamente a caracterização dessas sequências. Dentre as possíveis tarefas de anotação automática está o reconhecimento de enovelamento (*fold recognition*) que refere-se à identificação dos principais atributos estruturais entre elementos de estrutura secundária e suas inter-conexões. Reciprocamente, de acordo com Murzin et al. [Murzin et al., 1995], classificação estrutural pode ser realizada em níveis hierárquicos (*classe*, *enovelamento*, *super-família*, *família*) que denotam relações estruturais e evolucionárias. Nesse trabalho, focamos na classificação estrutural, que engloba o problema do *fold recognition*. Tanto a classificação estrutural quanto o reconhecimento de enovelamento são importantes passos na direção da predição de função proteica.

Ao longo dos anos, reconhecimento de enovelamento tem sido abordado sob diferentes perspectivas. Os autores de [Ding & Dubchak, 2001] focaram na extração de características da sequência e utilizaram *support vector machines* (SVM) e redes neurais artificiais como classificadores para uma base de dados composta por enovelamentos do SCOP. Mais tarde, *ensemble classifiers* [Shen & Chou, 2006] foram aplicados ao mesmo conjunto de atributos extraídos no trabalho anterior, melhorando a taxa de sucesso. A utilização de informação de sequência e estrutura de forma combinada trouxe um ganho considerável à tarefa de reconhecimento de enovelamento, como mencionado na abordagem de recuperação de informação proposta em [Cheng & Baldi, 2006].

Adicionalmente, vários esforços na predição de função baseada em estrutura foram realizados. Podemos citar, por exemplo, a busca por motivos estruturais [Barker & Thornton, 2003; Goyal et al., 2007; Stark & Russell, 2003] e resíduos funcionais (como sítios que ligam DNA [Shazman et al., 2007] e íons metálicos [Babor et al., 2008]), o uso de *templates* 3D [Laskowski et al., 2005b] e a comparação de enovelamento por alinhamentos estruturais [Holm & Sander, 1993; Kolodny et al., 2005]. Também existem tentativas de inferir função a partir da estrutura sem o uso de técnicas de alinhamentos, como aqueles no contexto de classificação de enzimas [Dobson & Doig, 2005; Alvarez & Yan, 2010]. De forma similar, no presente trabalho, não utilizamos técnicas de alinhamento ou qualquer informação extraída da sequência primária, valendo-se unicamente de fundamentos estruturais.

4.3 Resultados

De modo a testar a capacidade de nossa metodologia predizer classe estrutural e função proteica, executamos dois conjuntos de experimentos, projetados para essas duas diferentes tarefas.

Para predição de função, como mencionado na Seção 4.1.3, montamos uma base de dados de superfamílias de proteínas curada manualmente e outra baseada em números EC de modo a verificar se o método proposto poderia ser útil no contexto de anotação automática baseada em estrutura.

Para classificação estrutural, executamos experimentos para verificar nossa habilidade de assinalar a classificação SCOP em nível de *classe*, *enovelamento*, *superfamília* e *família* a domínios proteicos. Adicionalmente, de modo a colocar o trabalho em um contexto da literatura, comparamos o desempenho de nossa metodologia com o trabalho de Jain & Hirst. Até onde sabemos, esse trabalho apresenta os maiores níveis de precisão em tarefas de classificação estrutural até então publicados.

Finalmente, apresentamos alguns experimentos que objetivaram a avaliação da estratégia de redução de ruído baseada em SVD, bem como a análise de desempenho de um centroide alternativo, os carbonos-β, em tarefas de predição de função.

4.3.1 Predição de Função

Nos experimentos de predição de função, nosso objetivo era avaliar quão bem três diferentes algoritmos de classificação iriam se desempenhar em uma tarefa de predição de função de acordo com números EC e de acordo com um padrão-ouro de famílias de enzimas [Brown et al., 2006]. Utilizamos validação cruzada de 10 partições em todos os experimentos.

Para a base com os 950 ECs mais populosos, as assinaturas CSM foram capazes de atingir 95,1% de precisão e revocação após o pré-processamento com SVD, utilizando o algoritmo KNN (*K-Nearest Neighbors*). Os quatro níveis do número EC foram utilizados em conjunto como as classes das proteínas para treinar e testar o classificador, mas em etapas futuras pode-se expandir o estudo para classificação dos demais níveis dos números EC.

A Figura 4.5 apresenta a variação das métricas de desempenho pelas classes (números EC) consideradas. Muito embora o número de proteínas assinalada a cada número EC seja bastante desbalanceado, a maioria das classes foi devidamente classificada, com altos índices de qualidade, de acordo com as métricas extraídas.

Em relação aos números EC que não foram bem classificados, deseja-se verificar tanto a fração de entidades da base cobertas por eles quanto os possíveis motivos tornaram esses grupos de proteínas mais difíceis de classificar. Podemos supor dois casos particularmente difíceis. Em um deles teríamos números EC que possuem proteínas com estruturas bastante dissimilares. Em outras palavras, são grupos formados por proteínas com estruturas significativamente diferentes cuja função convergiu. Podemos denominar esses casos de forma genérica de *evolução convergente* e ela pode ser quantificada analisando a diversidade estrutural dos números EC, por exemplo, computando a variação da classificação SCOP (em seus quatro níveis) para cada grupo de enzimas de mesma função. De fato, é possível encontrar números EC com proteínas de diferentes *classes* do SCOP, o nível mais diverso, que diz respeito à composição em termos de estruturas secundárias.

Em outro caso teríamos enzimas com estruturas muito semelhantes, mas com números EC diversos. Podemos imaginar que houve uma diferenciação da função proteica (alterações no sítio ativo, por exemplo) sem muitas modificações globais na estrutura e de forma análoga denominamos esses casos de *evolução divergente*.

Em ambos os casos podemos imaginar que uma estratégia de análise mais local seja mais apropriada. Nesse sentido, desejamos verificar o desempenho de assinaturas CSM geradas para porções das proteínas, que possam ser mais conservadas do que a estrutura como um todo (como o caso de sítios catalíticos).

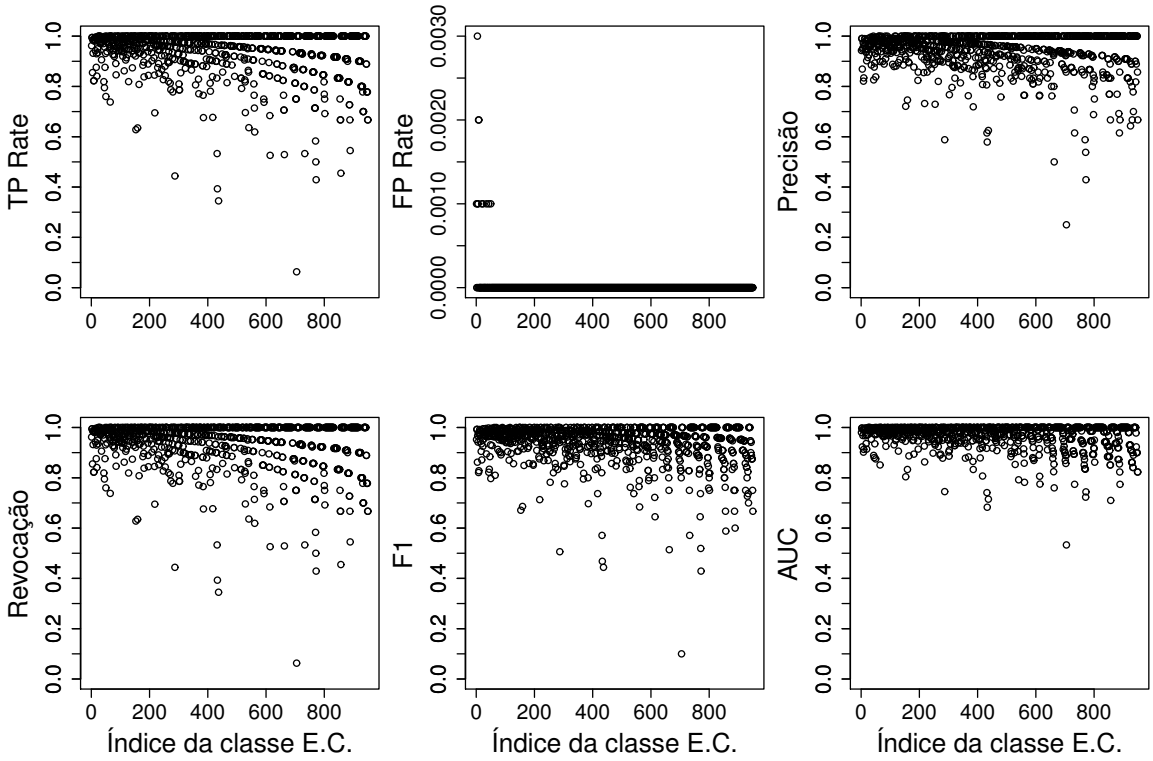

Figura 4.5. Métricas de desempenho para as diversas classes de números EC: a maioria dos números EC foram adequadamente classificados de acordo com as métricas exibidas.

Considerando o padrão-ouro, sem a utilização da SVD e executando o KNN, nosso método atingiu uma precisão média de 94,2% (± 5,5) e uma revocação de 94,5% (± 5,5) (Tabela 4.1). Para o Naive Bayes, atingiu 82,3% (± 13,8) de precisão e 79,2% (± 15,4) de revocação (Tabela 4.2), e para o Random Forest, 92,0% (± 6,9) de precisão e 91,6% (± 7,2) de revocação (Tabela 4.3). Também mostramos que, com a utilização da SVD, pudemos melhorar significativamente esses resultados, e no pior caso obtivemos 94,6% de precisão e 93,1% de revocação para a superfamília *enolase* utilizando Naive Bayes. Os métodos KNN e Random Forest foram capazes de detectar a superfamília *isoprenoid synthase type I* com 100% de precisão e revocação. Adicionalmente, executamos experimentos utilizando todas as seis superfamílias para treinar um único classificador. Nesse cenário, mesmo com um número maior de famílias

Tabela 4.1. Desempenho para predição de superfamílias no padrão-ouro utilizando KNN: os experimentos foram executados intra-família, um classificador por superfamília, logo as *classes* para predição eram as famílias das enzimas. As métricas de precisão e revocação são médias ponderadas. Validação cruzada de 10 partições foi empregada.

Superfamília	**Antes da SVD**		**Depois da SVD**		**ΔPrec.**	**ΔRev.**
	Prec.	*Revoc.*	*Prec.*	*Revoc.*		
Amidohydrolase	0.983	0.983	1.000	1.000	+1.7%	+1.7%
Crotonase	0.955	0.953	0.979	0.977	+2.4%	+2.4%
Enolase	0.876	0.853	0.971	0.967	+9.5%	+11.4%
Haloacid Dehalogenase	0.881	0.925	0.984	0.981	+10.3%	+5.6%
Isoprenoid Synthase I	1.000	1.000	1.000	1.000	+0.0%	+0.0%
Vicinal Oxygen Chelate	1.000	1.000	1.000	1.000	+0.0%	+0.0%
All	0.901	0.903	0.991	0.989	+9.0%	+8.6%

Tabela 4.2. Desempenho para predição de superfamílias no padrão-ouro utilizando Naive Bayes: os experimentos foram executados intra-família, um classificador por superfamília, logo as *classes* para predição eram as famílias das enzimas. As métricas de precisão e revocação são médias ponderadas. Validação cruzada de 10 partições foi empregada.

Superfamília	**Antes da SVD**		**Depois da SVD**		**ΔPrec.**	**ΔRev.**
	Prec.	*Revoc.*	*Prec.*	*Revoc.*		
Amidohydrolase	0.985	0.983	1.000	1.000	+1.5%	+1.7%
Crotonase	0.754	0.698	0.979	0.977	+22.5%	+27.9%
Enolase	0.596	0.580	0.946	0.931	+35.0%	+35.1%
Haloacid Dehalogenase	0.863	0.830	0.971	0.962	+10.8%	+13.2%
Isoprenoid Synthase I	0.970	0.966	0.970	0.966	+0.0%	+0.0%
Vicinal Oxygen Chelate	0.855	0.836	0.983	0.982	+12.8%	+14.6%
All	0.741	0.655	0.946	0.933	+20.5%	+27.8%

nas fases de treino e teste, foi possível atingir até 99% de precisão com KNN e Random Forest, após redução de dimensionalidade com SVD.

É importante ressaltar o fato de redução de dimensionalidade proporcionada pela SVD ter gerado resultados comparáveis para diferentes estratégias de classificação. Os classificadores que pior tinham se desempenhado com a base sem o tratamento foram os que obtiveram os maiores ganhos após a SVD. Uma vez que existe uma grande diferença de requisitos entre os algoritmos, tanto de tempo de execução quanto de memória, essa característica permite que escolhamos um classificador mais simples e eficiente e obtenhamos resultados equiparáveis àqueles obtidos com abordagens mais complexas e dispendiosas. Nesse trabalho sequer utilizamos o algoritmo considerado estado-da-arte em classificação, *Support Vector Machine* (SVM Cortes & Vapnik [1995]), uma vez que

Tabela 4.3. Desempenho para predição de superfamílias no padrão-ouro utilizando Random Forest: os experimentos foram executados intra-família, um classificador por superfamília, logo as *classes* para predição eram as famílias das enzimas. As métricas de precisão e revocação são médias ponderadas. Validação cruzada de 10 partições foi empregada.

Superfamília	**Antes da SVD**		**Depois da SVD**		**ΔPrec.**	**ΔRev.**
	Prec.	*Revoc.*	*Prec.*	*Revoc.*		
Amidohydrolase	0.983	0.983	0.996	0.996	+1.3%	+1.3%
Crotonase	0.844	0.837	0.979	0.977	+13.5%	+14.0%
Enolase	0.815	0.807	0.977	0.973	+16.2%	+16.6%
Haloacid Dehalogenase	0.982	0.981	0.986	0.981	+0.4%	+0.0%
Isoprenoid Synthase I	0.970	0.966	1.000	1.000	+3.0%	+3.4%
Vicinal Oxygen Chelate	0.947	0.945	0.984	0.982	+3.7%	+3.7%
All	0.898	0.892	0.991	0.991	+9.4%	+9.9%

o tempo de execução necessário para executá-lo em uma base de grande porte o torna praticamente inviável.

4.3.2 Classificação Estrutural

Não é de nosso conhecimento nenhum teste de classificação estrutural automática em um conjunto de proteínas tão grande, como o SCOP completo (que possui em torno de 110.000 domínios), que tenha sido publicado. Devido à habilidade da SVD em reduzir a dimensionalidade dos dados, representando as proteínas com um conjunto menor de atributos, apresentamos um método que pode eficientemente manipular tão grande volume de dados.

Fomos capazes de reconhecer o nível de *enovelamento* com 92,2% de precisão e 92,3% de revocação utilizando KNN (Tabela 4.4). Mesmo categorias proteicas com estruturas bastante diversas, como o nível de *classe* do SCOP, puderam ser separadas utilizando a assinatura CSM com valores de precisão e revocação muito bons (95,4% para ambos). A metodologia proposta foi capaz de classificar adequadamente proteínas nos quatro níveis da hierarquia do SCOP de forma bem sucedida, conforme reforçado pelas métricas de desempenho extraídas, mostrando que o CSM é adequado para classificação estrutural e também que o CSM é um componente promissor na definição de assinaturas estruturais proteicas.

Ademais, o fato do CSM ter tido bom desempenho em todos os níveis do SCOP, do mais geral ao mais específico, pode indicar que a assinatura gerada agrega informações de vários níveis, desde da composição de estruturas secundárias de proteínas até, em um nível mais fino, informações que diferenciam famílias de uma mesma super-

Tabela 4.4. Desempenho da classificação estrutural, utilizando KNN, para o conjunto completo de domínios do SCOP: o experimento foi executado para cada nível da hierarquia de classificação. As métricas de precisão e revocação são médias ponderadas. Validação cruzada de 10 partições foi empregada.

Nível SCOP	**Antes da SVD**		**Depois da SVD**		**ΔPrec.**	**ΔRev.**
	Prec.	*Revoc.*	*Prec.*	*Revoc.*		
Classe	0.927	0.926	0.954	0.954	+2.7%	+2.8%
Enovelamento	0.868	0.869	0.922	0.923	+5.4%	+5.4%
Super-família	0.871	0.872	0.926	0.927	+5.5%	+5.5%
Família	0.888	0.889	0.938	0.938	+5.0%	+4.9%

família. Uma hipótese a ser investigada é de que os níveis de informação possam estar relacionados com as faixas de *cutoff* considerados.

Adicionalmente, verificamos o impacto da imposição de um número mínimo de entidades por nó na hierarquia do SCOP na precisão da predição realizada. A Figura 4.6 mostra uma correlação aproximadamente linear entre essas variáveis para os níveis de *enovelamento*, *super-família* e *família* após processamento com SVD. Essa correlação não foi analisada para o nível *classe* porque todas as classes consideradas possuíam mais de 100 entidades.

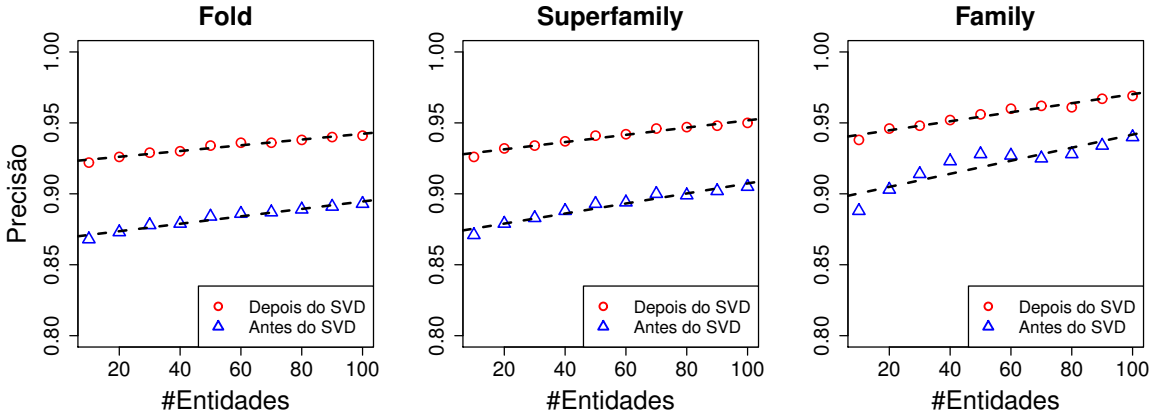

Figura 4.6. Correlação entre precisão e o número mínimo de representantes nas classes: avaliamos a base completa do SCOP, com e sem a utilização da SVD. Nesse contexto, *classe* deve ser entendida como o grupo de entidades com a mesma classificação SCOP para um dado nível: *enovelamento*, *super-família* ou *família* nesse caso.

Tabela 4.5. Comparativo de desempenho entre o estudo atual e o método introduzido por Jain & Hirst: as métricas de precisão e revocação representam médias ponderadas. Os resultados compreendem a execução de validação cruzada em 10 partições para o KNN.

Base	Nível SCOP	CSM+SVD			Jain et al.			ΔPrec.	ΔRev.
		Prec.	*Rev.*	*F1*	*Prec.*	*Rev.*	*F1*		
3SSE	Classe	0.991	0.991	0.991	0.890	0.840	0.864	+10.1%	+15.1%
	Enovelamento	0.956	0.957	0.956	0.860	0.450	0.591	+9.6%	+50.7%
	Super-família	0.956	0.957	0.956	0.800	0.550	0.652	+15.6%	+40.7%
	Família	0.935	0.935	0.935	0.820	0.870	0.844	+11.5%	+6.5%
4SSE	Classe	0.961	0.962	0.961	0.990	0.990	0.990	-2.9%	-2.8%
	Enovelamento	0.939	0.939	0.938	0.960	0.830	0.890	-2.1%	+10.9%
	Super-família	0.938	0.937	0.937	0.880	0.690	0.774	+5.8%	+24.7%
	Família	0.935	0.934	0.933	0.980	0.920	0.949	-4.5%	+1.4%
5SSE	Classe	0.985	0.985	0.985	0.980	1.000	0.990	+0.5%	-1.5%
	Enovelamento	0.969	0.969	0.969	1.000	0.690	0.817	-3.1%	+27.9%
	Super-família	0.970	0.969	0.969	0.980	0.650	0.782	-1.0%	+31.9%
	Família	0.967	0.965	0.965	0.980	0.920	0.949	-1.3%	+4.5%
6SSE	Classe	0.966	0.965	0.965	0.970	1.000	0.985	-0.4%	-3.5%
	Enovelamento	0.943	0.943	0.942	0.950	0.510	0.664	-0.7%	+43.3%
	Super-família	0.937	0.939	0.937	0.950	0.570	0.713	-1.3%	+36.9%
	Família	0.932	0.932	0.930	0.980	0.840	0.905	-4.8%	+9.2%

4.3.3 Análise Comparativa

Em Jain & Hirst [2010], os autores apresentam um método baseado no algoritmo Random Forest para predizer a classificação SCOP utilizando descritores de elementos de estrutura secundária, que obteve precisão de até 99%. Utilizando uma base similar à usada naquele trabalho, tentamos comparar nosso resultado com o deles. Consideramos que esse seja o estado-da-arte em métodos de classificação estrutural automática. Eles utilizaram um subconjunto do SCOP para classificação. Em nossa comparação de resultados, fomos capazes de atingir níveis equiparáveis de precisão mas com um ganho considerável de revocação (superando-os em até 50%) na maioria dos casos. Em apenas 3 dos 16 experimentos realizados, obtivemos valores inferiores de revocação e, como consequência, obtivemos valores superiores de F1. O conjunto completo de informações acerca desse experimento encontra-se na Tabela 4.5.

A Figura 4.7 mostra o comparativo de desempenho para cada experimento em termos de precisão e revocação. O CSM pôde superar significativamente a revocação obtida pelo estudo escolhido para comparação enquanto preservou um nível comparável de precisão. Ressaltamos que nosso método não é limitado a pequenas proteínas. Esse

resultado mostra que nosso método não é somente comparável ao de Jain & Hirst mas também apresenta uma melhora considerável de revocação.

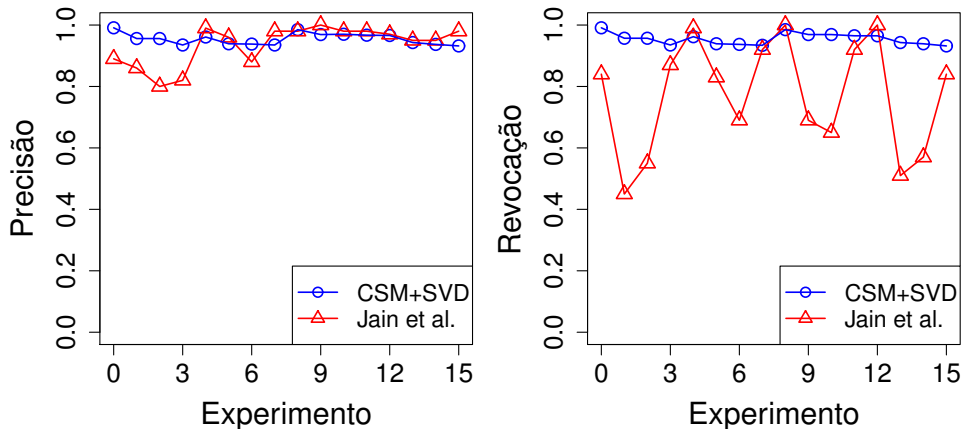

Figura 4.7. Comparativo de desempenho: um comparativo de desempenho da abordagem CSM (após SVD) e do trabalho de Jain & Hirst é exibido em termos de precisão e revocação. O CSM atingiu níveis compatíveis de precisão, apresentando uma melhora considerável de revocação.

4.3.4 Estratégia de Redução de Ruído

Como já mencionado, a redução de ruído baseada em SVD foi capaz de melhorar tanto níveis de precisão quanto de revocação nos experimentos realizados. Obtivemos ganhos de até 10,3% quando o KNN foi utilizado, 35,0% para o Naive Bayes e 16,2% para o Random Forest. É interessante notar que diferentes classificadores obtiveram resultados semelhantes após a utilização da SVD para redução de dimensionalidade (todas as métricas acima de 90%). A habilidade de reduzir o número de dimensões utilizadas é importante para escalabilidade nesse cenário por conta do grande volume de dados sendo gerado. Existem em torno de 110.000 domínios, i.e., instâncias a serem classificadas na base do SCOP e cada uma dessas pode ser representada por 151 atributos (dimensões) no caso do CSM com um *cutoff* de até 30Å.

Para encontrar o ponto de corte que maximiza a redução do ruído, realizamos um estudo da distribuição de valores singulares obtida para a base de dados do padrão-ouro de enzimas. A Figura 4.8 mostra a contribuição de cada valor singular na recomposição da informação original da matriz. Utilizando em torno de 9 dimensões foi possível representar a mesma informação (reduzindo o ruído) e obtendo altos índices de precisão na classificação com um volume de dados consideravelmente menor a ser tratado. Como mostrado no Figura 4.9, a precisão máxima pode ser alcançada com apenas 9 valores singulares para todos os experimentos.

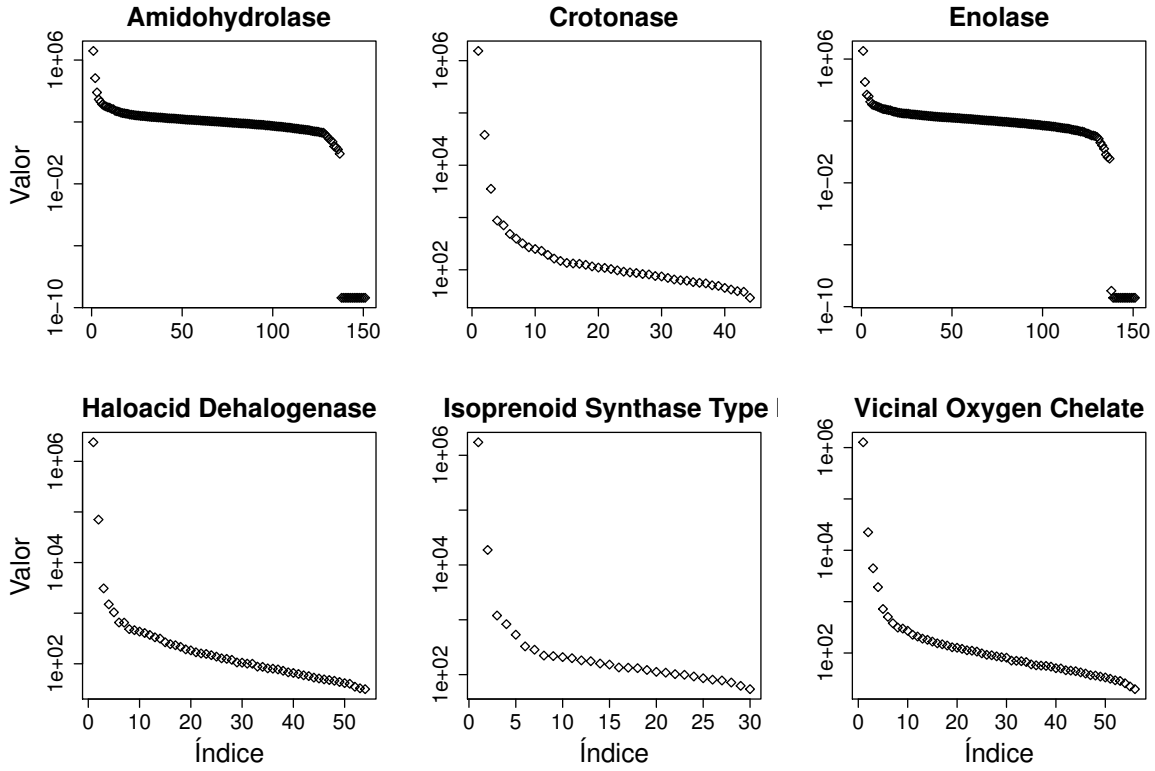

Figura 4.8. Distribuição de valores singulares: distribuição obtida após execução da rotina da SVD para cada superfamília considerada no padrão-ouro. Uma queda abrupta nessa distribuição denota um ponto de corte (critério do cotovelo) para redução de dimensionalidade. O eixo Y é exibido em escala logarítmica.

É importante enfatizar que a redução de dimensionalidade propiciou um aumento de eficácia mas também de eficiência nos experimentos. Em testes com a base composta pelos 950 números EC mais populosos o tempo de execução com todas as 151 dimensões (base original) foi de aproximadamente 45m01,42s (± 20,97s), para a etapa de classificação com Naive Bayes (com validação cruzada de 10 partes), enquanto que para a base reduzida após tratamento com a SVD (15 dimensões) foi de 6m33,08s(± 3,93s), uma diferença de quase 7 vezes. Em todos os testes utilizamos uma máquina *quad-core*, com 8GB de memória.

4.3.5 Avaliação da Utilização de Centróides

A metodologia proposta no presente trabalho tem por base o cálculo de distância Euclidiana entre resíduos de aminoácidos intra-cadeia. Cada resíduo considerado é representado por um ponto representativo ou centróide. Realizamos uma análise comparativa inicial de desempenho entre dois centróides (carbono-α e carbono-β) para

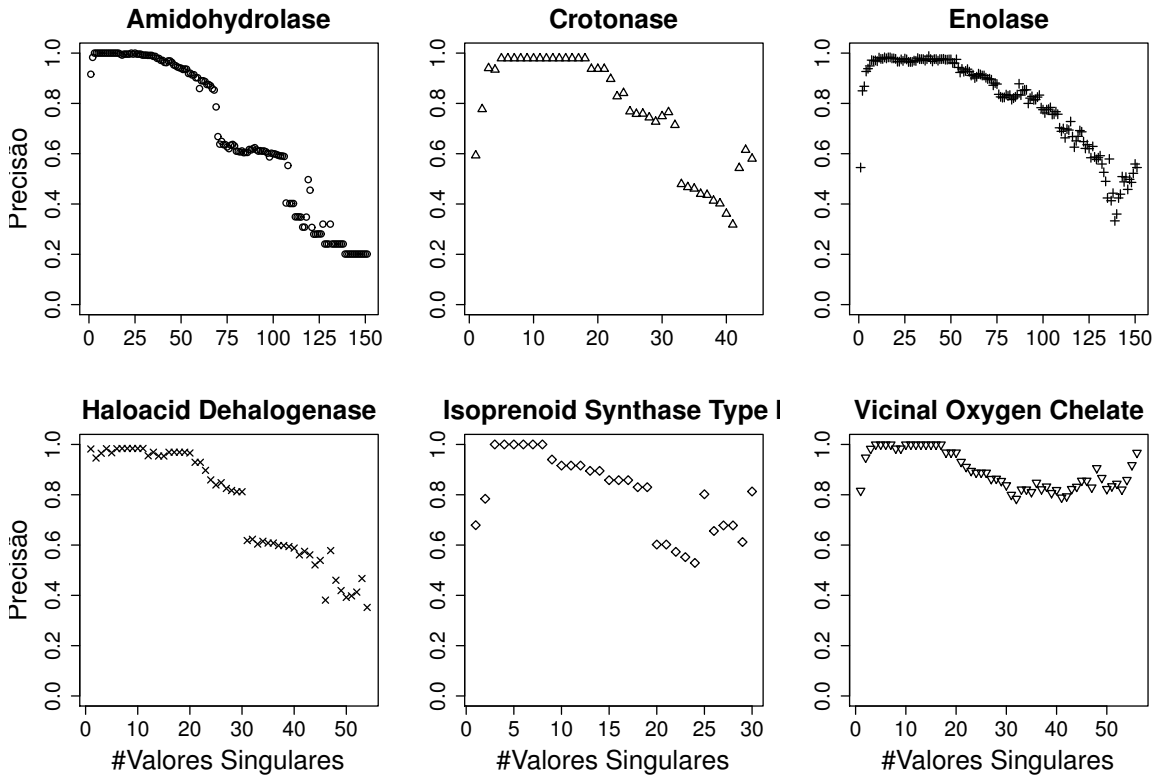

Figura 4.9. Influência do ponto de corte para redução de dimensionalidade na precisão (média ponderada) para as superfamílias do padrão-ouro: Uma queda repentina na precisão após um certo número de valores singulares pode indicar o ponto onde componentes com ruído começam a aparecer.

a base de dados de números EC. A Figura 4.10 mostra um comparativo de desempenho entre os centróides carbono-α e carbono-β para a base de dados de números EC. O C_α desempenhou-se melhor em todos os experimentos, fato que demanda investigação mais aprofundada.

É importante destacar que outros centróides podem ser utilizados ao invés do carbono-α ou carbono-β, como o último átomo pesado da cadeia lateral (Last Heavy Atom - LHA) ou mesmo um ponto artificial como o centro geométrico da cadeia lateral.

4.4 Conclusões

Predição de função e classe estrutural proteicas, enquanto meios de entendimento acerca da composição, operação, interação e evolução de proteínas, são ainda grandes desafios em face ao explosivo crescimento de geração e armazenamento de dados proteicos em repositórios públicos. Para acompanhar o ritmo frenético imposto pela crescente

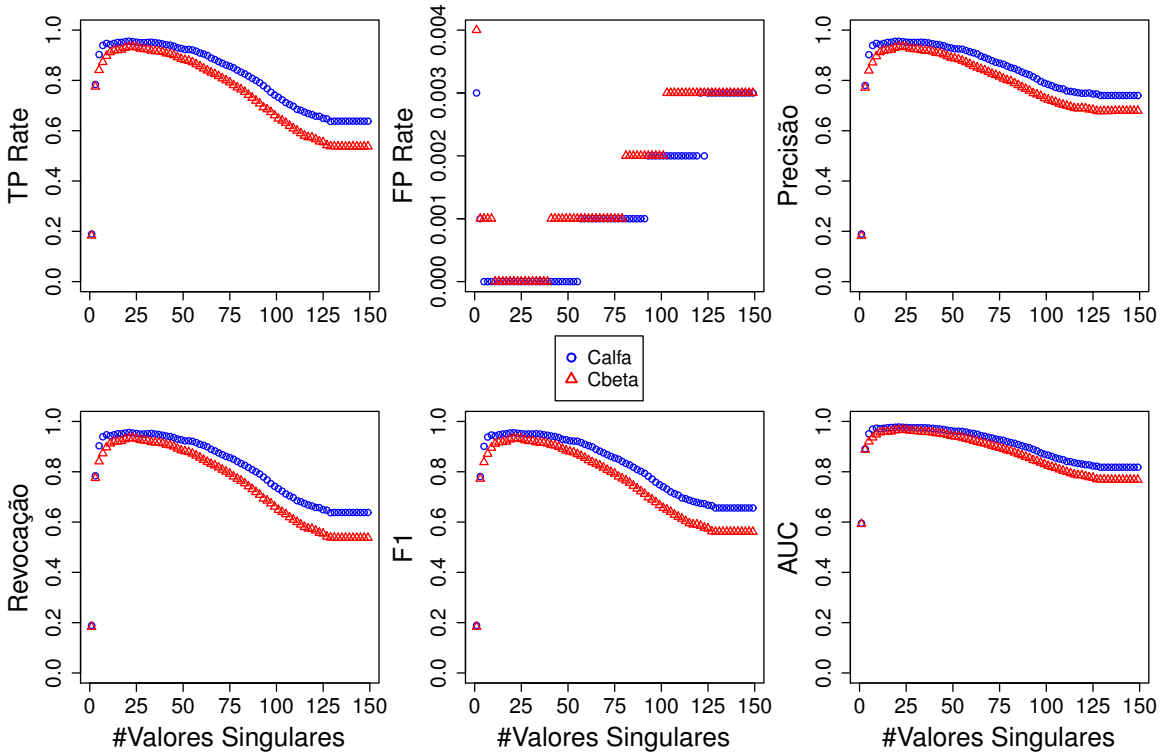

Figura 4.10. Comparação de desempenho entre os centróides C_α e C_β na geração da matriz CSM, para a base de números EC: em todos os experimentos, C_α apresentou melhores indicadores em termos das métricas apresentadas na figura.

disponibilidade de dados, novos e eficientes métodos para anotação automática e semi-supervisionada são necessários.

Como mecanismo para aproveitar a íntima relação entre estrutura e função proteica, desenvolvemos uma assinatura estrutural baseada no conceito CSM que tem por fundamento padrões de distância inter-resíduo e a validamos em tarefas de classificação estrutural e predição de função. A motivação para essa abordagem advém da hipótese de que proteínas com diferentes estruturas teriam diferentes padrões de distâncias entre seus resíduos, e de forma complementar, similaridade estrutural seria refletida nessas distâncias. A interpretação e entendimento dos padrões de distância intrínsecos gerados pelo CSM demandam investigações complementares.

Como requisito e demanda para sua aplicação em grandes volumes de dados em contínuo crescimento, mostramos que nossa metodologia é escalável para cenários reais, como a classificação completa do SCOP, como mostrado em capítulos anteriores, além de ter demonstrado uma eficácia compatível ou superior ao estado-da-arte. Gostaríamos de enfatizar que nosso método é provavelmente o primeiro a apresentar

um experimento de classificação completa do SCOP em tempo e desempenho aceitáveis (alguns minutos em uma máquina com quatro núcleos - *quad-core*).

O trabalho descrito nesse Capítulo foi documentado e publicado na forma de artigo científico. Mais informações podem ser encontradas no Apêndice A.

Capítulo 5

Predição de Ligantes

O reconhecimento molecular desempenha um papel fundamental em muitos processos celulares. As condições responsáveis pela ligação e interação entre duas ou mais moléculas são uma combinação de complementaridade conformacional e físico-química [Kahraman et al., 2007]. A descoberta e compreensão das características dos bolsões de ligação (ou *pockets*) do receptor enquanto pré-requisitos para o processo de reconhecimento são passos fundamentais para tarefas como a predição de ligantes para proteínas, a identificação de possíveis alvos terapêuticos, a descoberta de compostos líderes e o projeto de novos fármacos.

Assume-se que ligantes semelhantes possuem sítios de ligação similares em termos tanto do formato geométrico/espacial quanto de suas propriedades físico-químicas. Vários métodos foram propostos na literatura para descrever, comparar e prever ligantes para sítios ativos. No entanto, a despeito de importantes contribuições da maioria dos trabalhos, métodos que dependem de alinhamentos múltiplos de estruturas e comparações par-a-par de *pockets* podem ser proibitivamente caros para bases de dados de grande escala. À medida que a disponibilidade de dados biológicos vem crescendo de uma forma exponencial nos últimos anos, escalabilidade tornou-se uma característica crucial para a execução de tais tarefas em cenários reais.

Para superar estes desafios, foi proposta uma nova metodologia para a predição de ligantes baseada no receptor proteico, que é suportada por uma assinatura estrutural e fisico-química para *pockets*. Padrões de distâncias presentes nos *pockets* proteicos são extraídos a partir de uma modelagem computacional dos bolsões como grafos atômicos, a partir do conceito CSM. É também realizada uma etapa de pré-processamento, filtragem e melhoria dos dados pela redução de ruído e dimensionalidade, o que atribuiu não apenas uma melhoria significativa na eficácia da metodologia de predição em relação a métodos competidores, mas também garantiu escalabilidade à metodologia.

Padrões de distâncias atômicas capturam os arranjos estruturais de uma proteína e, por conseguinte, refletem sua função. Nesse sentido, utilizar essa fonte de informação para descrever sítios de ligação é uma estratégia apropriada, dada a estreita relação entre estrutura e função proteica bem como a importância da complementaridade geométrica no processo de reconhecimento molecular. Adicionalmente, considerar propriedades físico-químicas nesses padrões, outro importante requisito para o reconhecimento, também parece ser uma boa abordagem, tendo em mente a necessidade de uma fonte de informações eficazes para que seja possível descrever, comparar e prever interações proteína-ligante com sucesso.

Bolsões em receptores proteicos podem ser vistos como grafos onde os nós representam os átomos da proteína e as arestas as interações químicas estabelecidas entre esses. Propriedades topológicas e químicas podem ser extraídas a partir desses grafos e sumarizadas em uma assinatura de reconhecimento molecular. Essas assinaturas compactas podem, então, ser utilizadas em experimentos em larga escala de predição de ligantes. No presente trabalho, derivamos uma nova assinatura para *pockets* proteicos baseada no conceito CSM e em uma modelagem de grafos atômicos. Propomos uma versão atômica, rotulada da assinatura (doravante chamada *aCSM*), que é independente de orientação molecular e não requer nenhum tipo de informação do ligante para seu cálculo.

Dada a complexidade do processo de reconhecimento molecular, essas assinaturas devem ser robustas para que sejam efetivas na predição de ligantes. Uma das dificuldades, e importante fonte de introdução de ruído nos dados, é a flexibilidade do ligante, o que leva a uma grande diversidade conformacional. Por exemplo, a Figura 5.1 (a) ilustra cinco representantes de Nicotinamida-Adenina Dinucleotídeo (NAD), um ligante altamente flexível, obtidos do repositório *online Protein Data Bank* (PDB). Eles foram alinhados e seus bolsões calculados por um critério de distância. Podemos ver que a variabilidade das conformações impacta diretamente na definição e forma do sítio de ligação. Além disso, os mecanismos de ajuste do *pocket* por indução do ligante [Koshland Jr, 1958] bem como a regulação alostérica [Monod et al., 1963] podem promover mudanças conformacionais expressivas na proteína alvo.

Outro fator desafiador refere-se às várias poses adotadas por um ligante em diferentes *pockets*, bem como sua acessibilidade ao solvente quando ligado à proteína receptora. A Figura 5.2 apresenta um exemplo onde o ligante Flavina-Adenina Dinucleotídeo (FAD) está ligado a três *pockets* com graus muito diferentes de acessibilidade ao solvente. Podemos notar claramente que esses fatores podem afetar dramaticamente a forma e tamanho do *pocket* de um mesmo ligante, o que pode impor limitações severas a métodos que se baseiam exclusivamente em alinhamentos

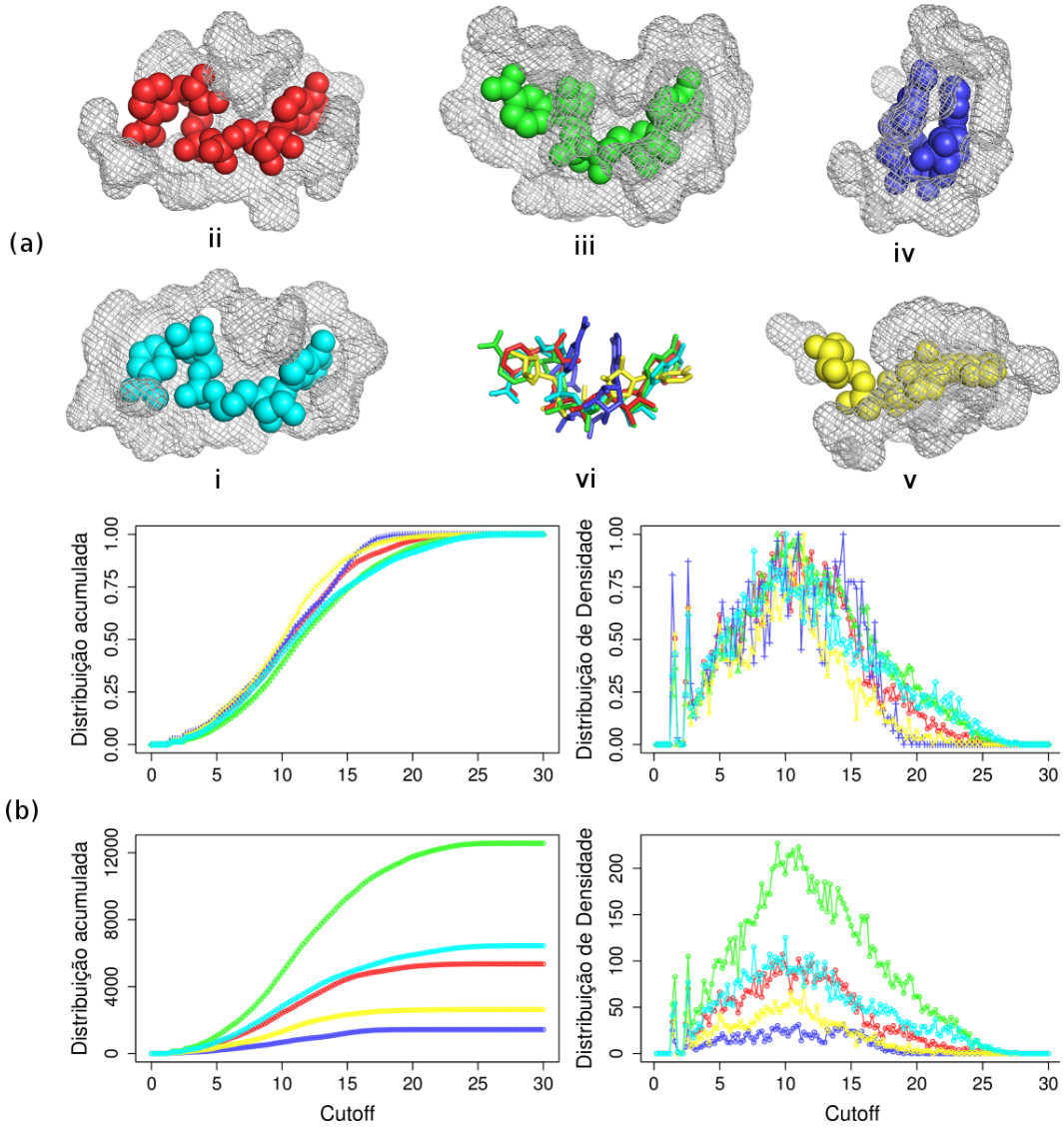

Figura 5.1. Diversidade conformacional de um ligante. (a) Mostra moléculas de NAD apresentando diferentes conformações e o impacto em seus respectivos *pockets* (calculados utilizando uma distância limite de 5Å). Os identificadores PDB considerados foram (a.i) 3KSD:Q (ligante em ciano), (a.ii) 1A5Z:A (ligante em vermelho), (a.iii) 1NAH:A (ligante em verde), (a.iv) 1ZRQ:B (ligante em azul), (a.v) 2OOR:B (ligante em amarelo). (a.vi) Mostra os ligantes alinhados pelo programa LigAlign [Abraham & Lilien, 2010]. A distribuição acumulada e de densidade das assinaturas aCSM geradas, explicadas em detalhe em Seções posteriores, são apresentadas nas mesmas cores em (b), considerando valores normalizados (acima) e absolutos (em baixo).

estruturais. No nosso caso, essa característica também é uma importante fonte de ruído para as assinaturas propostas, que são baseados em padrões distância inter-atômicas.

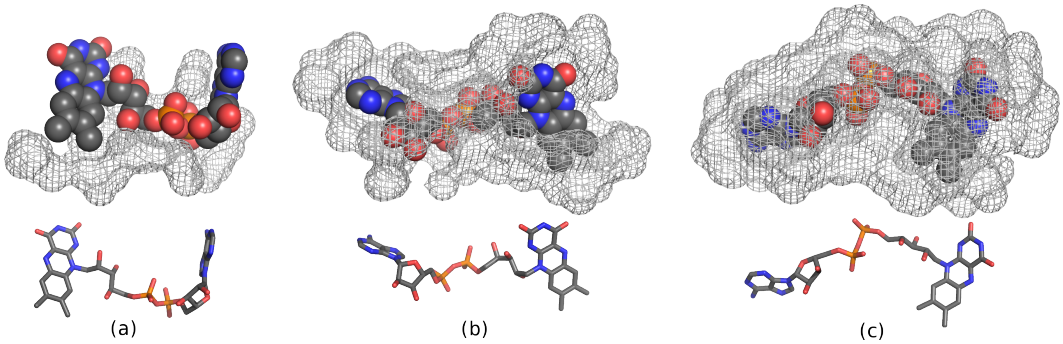

Figura 5.2. Diversidade da acessibilidade ao solvente do ligante. A Figura mostra moléculas do ligante FAD com três diferentes graus de acessibilidade ao solvente e seu impacto da definição de seu respectivo *pocket* (calculado utilizando uma distância limite de 5Å). Os identificadores PDB utilizados foram (a) 1O26:A, (b) 1AHV:A and (c) 1H83:C. Esquema de cores CPK do Pymol: carbonos em cinza, nitrogênios em azul, oxigênios em vermelho e enxofres e amarelo.

Para lidar com esses desafios e eliminar o ruído inerente às assinaturas, aplicamos uma estratégia de redução de dimensionalidade de ruído baseada na Decomposição em Valores Singulares (SVD, do inglês *Singular Value Decomposition*). A Figura 5.1 (b) apresenta as assinaturas propostas para os sítio de ligação de moléculas de NAD antes da etapa de redução de ruído. Podemos notar que, apesar da semelhança no perfil das curvas, ainda vemos uma considerável variabilidade entre elas o que é reduzido pela normalização de dados conferida pelo pré-processamento das assinaturas com a SVD. Esse passo de pré-processamento é essencial para extrair das assinaturas originais, os componentes mais importantes para descrever os *pockets*, descartando redundância e dimensões não conservadas.

Nas Seções a seguir fazemos a formulação do problema, sua definição e descrição da modelagem computacional empregada. Também descrevemos os métodos e bases de dados utilizadas. Mostramos a partir dos experimentos realizados que as assinaturas propostas conseguem lidar com os aspectos desafiadores da predição de ligantes em larga escala, alcançando uma Área Abaixo da Curva ROC (AUC) de 0,92 para um conjunto de dados composto por mais de 35,000 *pockets* de enzimas. Não é de nosso conhecimento nenhum outro relato na literatura que tenha testado um conjunto de dados de volume comparável para tal tarefa. Apesar da variabilidade conformacional proeminente do ligante NAD, nossa metodologia foi capaz de recuperar seus *pockets* com uma AUC de 0,96. Da mesma forma, fomos capazes de recuperar sítios de ligação de FAD, apresentando moléculas com diferentes acessibilidades ao solvente, com

uma AUC de 0,99. Quando comparado a métodos considerados estado-da-arte, nossa abordagem alcança resultados comparáveis ou até melhores. Por fim, apresentamos um estudo de caso, onde prevemos novos ligantes para proteínas de *Trypanosoma cruzi*, o parasita responsável pela doença de Chagas e efetuamos uma validação dos mesmos *in silico* através de um protocolo de *docking*, mostrando o aplicabilidade do método proposto em um cenário de aplicação real.

5.1 Modelagem Computacional

Na presente Seção descrevemos a modelagem computacional realizada que corresponde à representação de *pockets* proteicos como grafos de interações atômicas.

- **Grafo de contatos inter-atômicos:** grafo simples, completo (clique), não-direcionado, ponderado, rotulado;
- **Nós:** correspondem a todos os átomos pesados presentes no *pocket*;
- **Arestas:** todos os nós são conectados, formando um grafo completo ou clique;
- **Peso das arestas:** corresponde à distância Euclidiana entre os átomos (nós);
- **Rótulo dos nós:** nós são rotulados de acordo com suas características fisico-químicas em três níveis de especificidade.

Em suma, as estruturas dos *pockets* proteicos foram modeladas como cliques onde, os nós representam os átomos e o peso das arestas a distância entre esses. A rotulação dos nós por características fisico-químicas foi realizada em três níveis de especificidade, dando origem a três diferentes assinaturas aCSM:

- **aCSM**: assinatura sem rotulação, que por sua vez gera um valor por *cutoff*, correspondendo ao número de átomos em contato (número de arestas) de acordo com uma dada distância de corte.
- **aCSM-HP**: assinatura categoriza átomos em hidrofóbico ou polar, gerando três valores por *cutoff*, *i.e.*, a frequência de contatos hidrofóbico-hidrofóbico, hidrofóbico-polar e polar-polar.
- **aCSM-ALL**: assinatura categoriza os átomos em oito classes: hidrofóbico, positivo, negativo, aceptor, doador, aromático, enxofre e neutro. A combinação desses rótulos gera 36 valores por *cutoff*. A classificação dos átomos foi obtida a

partir do programa PMapper [ChemAxon, 2012] em pH 7. O PMapper identifica propriedades farmacofóricas de átomos a partir de suas estruturas moleculares.

No Apêndice C, Tabela C.1, é exibida a classificação dos átomos utilizada para o cálculo das assinaturas aCSM.

5.1.1 Métodos

Nesta seção descrevemos o fluxo metodológico empregado, bem como o procedimento para geração das assinaturas, redução de ruído e dimensionalidade das mesmas e, por fim, explicamos como o método foi avaliado e validado.

A Figura 5.3 mostra o *workflow* de predição de ligantes baseada em assinaturas aCSM. Este é dividido nas seguintes etapas principais: coleta e modelagem de dados, geração das assinaturas e redução de dimensionalidade/ruído, aprendizado supervisionado, predição e validação de ligantes.

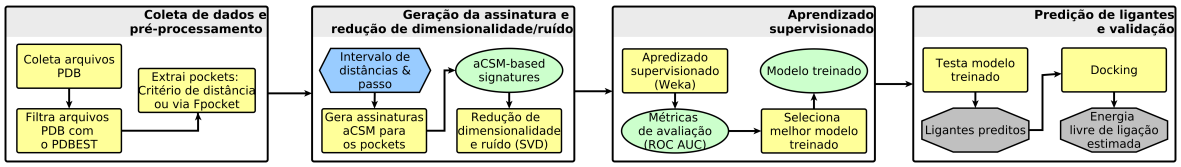

Figura 5.3. *Workflow* de predição de ligantes baseado em assinaturas aCSM. O *workflow* é dividido em quatro etapas principais: coleta e modelagem de dados, geração de assinaturas e redução de ruído/dimensionalidade, aprendizado supervisionado, predição de ligantes e validação. Caixas hexagonais azuis denotam arquivos ou parâmetros de entrada, caixas elipsoidais verdes são arquivos intermediários gerados, caixas retangulares amarelas denotam passos intermediários, e caixas octogonais cinzas as saídas, *i.e.* os ligantes preditos e a energia livre de ligação estimada para esses.

No presente Capítulo, estendemos a assinatura CSM de grafos inter-resíduo descrita no Capítulo 4 para seu nível atômico (CSM atômica, ou simplesmente aCSM). As assinaturas aCSM são geradas da seguinte forma: para cada grafo proteico criamos um vetor de atributos. Em primeiro lugar, calcula-se a distância euclidiana entre todos os pares de átomos, de modo a ponderar as arestas do grafo. Definimos também um intervalo de distâncias a ser considerado bem como um passo de distância. As distâncias desse intervalo são percorridas, calculando-se a frequência de pares de átomos que estão próximos de acordo com o limiar de distância atual, isto é, os átomos em contato (alternativamente, o número de arestas do grafo cujo peso é menor que a distância corrente). Como resultado, temos uma distribuição acumulada do número de arestas

no grafo para um dado intervalo de distâncias. A contagem da frequência de arestas por distância é discretizada pelos tipos de rótulos dos nós. O Algoritmo 3 exibe a função que calcula a aCSM.

Algorithm 3 Cálculo da assinatura aCSM

```
1: function aCSM(ConjuntoPockets, ClassesAtomos, D_MIN, D_MAX, D_PASSO)
2:     for all pocket i ∈ (ConjuntoPockets) do
3:         j = 0
4:         matrizDist ← calculaDistanciaParAPar(pocket)
5:         for distancia ← D_MIN; até D_MAX; passo D_PASSO do
6:             for all classe ∈ (ClassesAtomos) do
7:                 aCSM[i][j] ← obtemFrequencia(matrizDist, distancia, classe)
8:                 j++
9:     retorna aCSM
```

A geração de assinaturas aCSM é executada em tempo quadrático, ou seja, tem complexidade de tempo $O(n^2)$, onde n é o número de átomos do *pocket*, devido ao cálculo de distância par-a-par entre os átomos (nós). É importante salientar que o método é facilmente paralelizável, uma importante e desejável característica para sua utilização eficiente em arquiteturas *multi-core*.

Nos experimentos apresentados neste trabalho, nós consideramos o intervalo de distâncias de 0-30Å, com um passo de 0,2Å, o que gerou vetores de atributos de 151, 453 e 5436 dimensões, para cada *pocket*, para as assinaturas aCSM, aCSM-HP e aCSM-ALL, respectivamente. Na Figura 5.4, pode-se ver, através da distribuição de diâmetro dos *pockets* da base de dados de larga escala, que intervalo de distâncias de 0-30Å utilizado na geração das assinaturas engloba cerca de 95% dos diâmetros dos *pockets*.

É importante destacar a generalidade método ao passo que este pode ser aplicado para predizer tanto ligantes proteicos quanto não-proteicos. A Decomposição em Valores Singulares (SVD) é uma técnica de álgebra linear amplamente utilizado em tarefas de redução de dimensionalidade. No presente trabalhar usamos a SVD também para reduzir ou eliminar o ruído inerente das assinaturas geradas e, por consequência, melhorar a eficácia e reduzir o tempo de processamento dos algoritmos de classificação, em termos de tempo de execução e os requisitos de memória.

5.1.2 Tarefas de Classificação

Um conjunto de algoritmos de classificação foi avaliado para os experimentos comparativos. Resultados detalhados podem ser encontrados na Figura 5.5. Dois desses algoritmos foram escolhidos tendo em mente um compromisso entre eficácia e

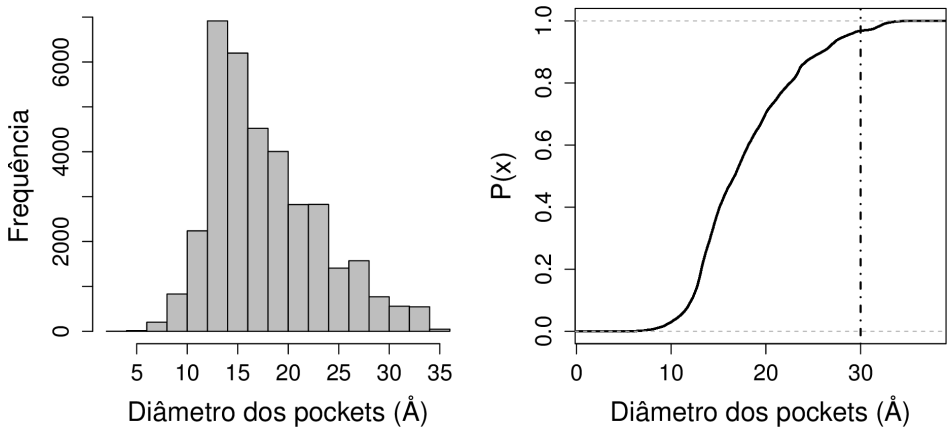

Figura 5.4. Distribuição do diâmetro dos *pockets* da base de dados de larga escala composta por enzimas. A porção esquerda da figura exibe um histograma e a porção direita uma Função de Distribuição Acumulada (CDF, do inglês, *Cumulative Distribution Function*) dos diâmetros dos *pockets* considerados. O intervalo de distâncias de 0-30Å utilizado na geração das assinaturas engloba cerca de 95% dos diâmetros dos *pockets*.

eficiência: KNN foi escolhido para experimentos de larga escala e a Regressão Logística Multinomial para as bases de dados comparativas, onde o objetivo era maximizar a taxa de sucesso. Para todas as tarefas de classificação, a ferramenta Weka Hall et al. [2009], *developer version 3.6.2*, foi utilizada.

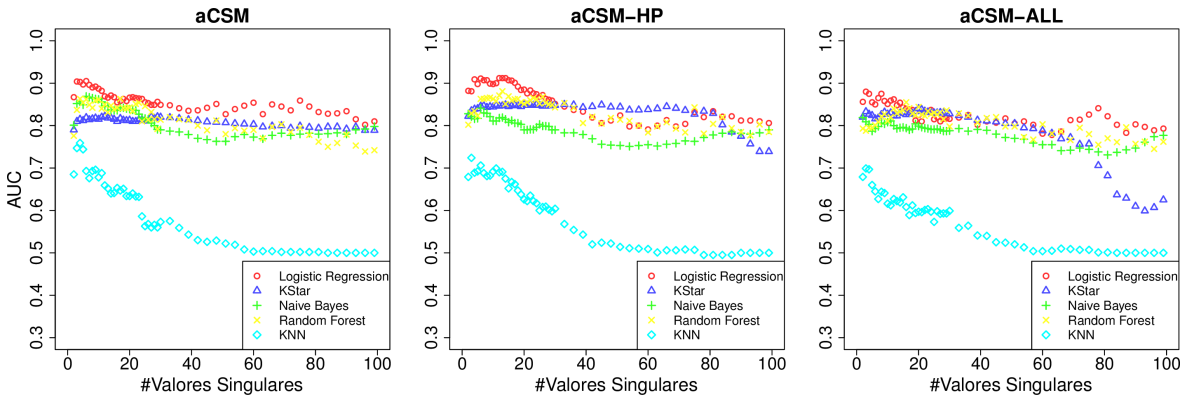

Figura 5.5. Comparativo do desempenho de algoritmos de classificação para a base de dados Kahraman. Treinamos diversos classificadores fornecendo os diferentes tipos de assinaturas, para diferentes números de valores singulares utilizados na etapa de redução de dimensionalidade. Os algoritmos escolhidos para serem avaliados foram: Regressão Logística Multinomial [Landwehr et al., 2005], K* [Cleary & Trigg, 1995], Naive Bayes [Lewis, 1998], Random Forest [Breiman, 2001] e KNN [Cover & Hart, 1967]. Para cada um dos três tipos de assinaturas propostas (aCSM, aCSM-HP e aCSM-ALL) o algoritmo com melhor desempenho foi sistematicamente a Regressão Logística.

5.1.2.1 Metodologia de Avaliação

Avaliamos e comparamos o método proposto com o estado-da-arte em predição de ligantes utilizando como métrica a Área Sob a Curva ROC (*Area under ROC curve* - AUC), mesma métrica utilizada por métodos competidores.

No presente trabalho foi também utilizada a validação cruzada em 10 partições para todos os experimentos, exceto para os comparativos, onde a mesma metodologia descrita pelos métodos concorrentes foi empregada. Para os experimentos comparativos, utilizamos a validação cruzada do tipo *leave-one-out*, onde o número de partições é o mesmo número de instâncias, ou seja, o conjunto de teste tem sempre tamanho 1.

5.1.3 Bases de Dados

A fim de obter uma ampla variedade de experimentos para validar o método proposto, sua generalidade e aplicabilidade em cenários reais, utilizamos quatro bancos de dados com finalidades diferentes.

a. **Base de enzimas de larga escala**: Em um trabalho anterior [Pires et al., 2011] propusemos uma base de dados com proteínas dos 950 números EC mais populosos em termos do número de estruturas de proteínas depositadas, com pelo menos 9 representantes por classe, totalizando 55.474 cadeias. Tal conjunto de dados consiste de enzimas com anotação UniProt revisada, ou seja, anotações validadas para essa base. Apenas ligantes com pelo menos 7 átomos pesados foram considerados e apenas *pockets* com pelo menos 10 átomos pesados. No total, 36.480 *pockets* foram extraídos para 604 ligantes distintos, com pelo menos 10 representantes por ligante. Essa base é utilizada de modo a demostrar-se a aplicabilidade do método para uma base de dados de grande porte, para enzimas bastante distintas, o que aproxima o experimentos de possíveis cenários reais de aplicação.

b. **Base Kahraman**: proposta por [Kahraman et al., 2007], essa base é composta de 100 sítios de ligação. Dentre as principais características da base podemos destacar:

   - Sítios não são evolutivamente relacionados;
   - Estruturas proteicas resolvidas por cristalografia com difração de raio-X;
   - Complexos proteína-ligante com 10 ligantes distintos com tamanho e flexibilidade variados
   - Os códigos dos ligantes considerados, conforme referenciados pelo PDB, foram: AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD.

Essa base é utilizada em experimentos comparativos entre o métodos proposto e seus competidores.

c. **Base Hoffmann HD**: proposta por [Hoffmann et al., 2010], esta base é formada de 100 *pockets* proteicos formando complexos com 10 ligantes distintos. Esses ligantes possuem tamanhos similares e foram selecionados pelos autores de modo a complementar a base Kahraman, uma vez que esse último contempla ligantes de volumes bastante variados. Os códigos dos ligantes considerados, conforme referenciados pelo PDB, foram: PMP, SUC, LLP, LDA, BOG, PLM, SAM, U5P, GSH, 1PE. Essa base também é utilizada em experimentos comparativos entre o métodos proposto e seus competidores.

d. **Base de *Trypanosoma cruzi***: composto por proteínas de *Trypanosoma cruzi*. Os critérios adotados para seleção das proteínas e resultados foram:

   - Estruturas proteicas resolvidas por cristalografia com difração de raio-X;
   - Com resolução abaixo de 2.5Å;
   - Resultando em 104 estruturas PDB, divididas em 200 cadeias;

   Com auxílio do programa FPocket [Le Guilloux et al., 2009], 1.846 *pockets* foram extraídos. Adicionalmente, a partir do critério de distância (5.0Å), foram extraídos 225 *pockets* dos ligantes presentes nas estruturas originais, eliminando-se ligantes pequenos e artefatos cristalográficos. Essa base foi utilizada de modo a levantar-se ligantes candidatos para proteínas cuja interação é desconhecida.

5.2 Trabalhos Relacionados

A fim de descrever *pockets* de ligação em proteína de modo a compará-los ou predizer ligantes para esses, vários métodos têm sido propostos na literatura. Alguns deles são baseadas em um paradigma de métricas de similaridade *pockets*. Em [Davies et al., 2007], os autores introduziram uma pontuação de similaridade de sítios com base em um modelo probabilístico e comparou sua métrica com o índice de Tanimoto. Bolsões de ligação em proteínas foram comparados por decomposição em harmônicos esféricos [Morris et al., 2005], técnica também usada em um estudo de sua variação conformacional [Kahraman et al., 2007]. Mais recentemente, em [Hoffmann et al., 2010] os autores propuseram um método para quantificar a similaridade de *pockets*, representando-os como nuvens de átomos, e comparando alinhamentos resultantes com um *kernel* de convolução.

Uma medida de similaridade também foi derivada em um trabalho recente [Ueno et al., 2012] a partir de Funções de Distribuição Radial (RDFs) das propriedades físico-químicas de sítios catalíticos, informações que foram então usadas para agrupar enzimas por função. Em [Gonçalves-Almeida et al., 2012], *pockets* são comparados usando regiões hidrofóbicas (ou *patches*), representadas por centroides geométricos, e a sua conservação é detectada a despeito de dissimilaridade de sequência e estrutura.

Outro conjunto de métodos tenta comparar sítios de ligação com base em alinhamentos múltiplos. Métricas de similaridade foram derivados a partir dos alinhamentos de sítios de ligação ou *fingerprints* de cavidades representadas por suas propriedades físico-químicas ou topológicas em [Shulman-Peleg et al., 2008; Schalon et al., 2008], enquanto o autor de [Spitzer et al., 2011] propôs uma abordagem baseada na caracterização da superfície proteica. Há também esforços que usam alinhamentos múltiplos de grafos e a algoritmos de casamento de cliques [Weskamp et al., 2007; Najmanovich et al., 2008] para capturar interações receptor-ligante. Outras abordagens alternativas para o estudo do mecanismo de ligação incluem o uso de técnicas de *docking* e *Quantitative Structure-Activity Relationships techniques* (QSARs) [Sippl, 2000].

5.3 Resultados

A fim de testar e validar a capacidade de nossa assinatura de descrever sítios de ligação para suportar e auxilar em tarefas de predição de interações proteína-ligante nós projetamos um extenso conjunto de experimentos. Em primeiro lugar, mostramos que o nosso método pode ser utilizado em experimentos de predição de larga escala e avaliamos sua precisão no desempenho dessa tarefa. Em seguida, comparamos as três versões propostas de assinaturas aCSM e avaliamos qual delas apresenta o melhor poder descritivo para predição de ligantes. Depois disso, apresentamos os resultados comparativos como métodos estado-da-arte descritos na literatura e utilizando suas respectivas bases de dados. Finalmente, aplicamos nossa metodologia para predizer ligantes para *pockets* de proteínas de *Trypanosoma cruzi*, comparando a energia livre de ligação estimada por procedimentos de *docking* entre receptores e ligantes preditos pelas assinaturas aCSM, como complexos reais disponíveis no PDB através de um protocolo de *redocking* e com a atribuição aleatória de ligantes (modelo nulo).

5.3.1 Experimentos em Larga Escala

A Figura 5.6 apresenta os valores de AUC para nosso método para os dados da base de grande escala que contempla enzimas revisadas do UniProt a partir das quais, mais

de 35.000 *pockets* foram extraídos. Podemos notar que o método foi capaz de predizer ligantes com sucesso, em cada tipo de experimento, AUC variando de 0,6 a 0,92. Nas próximas seções, vamos explicar as variações do método que geraram os resultados mostrados na Figura.

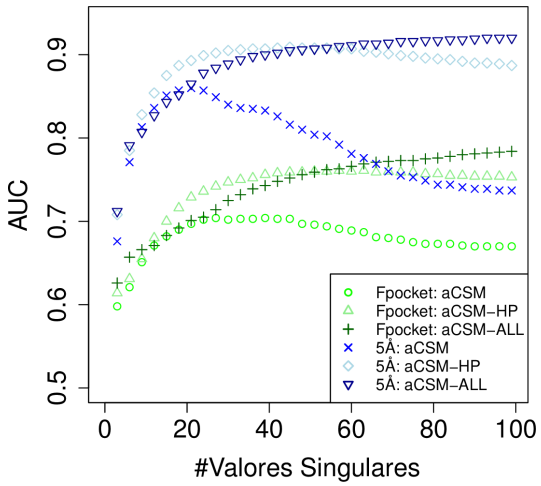

Figura 5.6. Comparativo de desempenho da tarefa de predição, em termos da métrica AUC, entre dois métodos de definição de *pockets*: FPocket (três curvas inferiores em tons de verde) e a definição a partir de uma distância de corte (três curvas superiores em tons de azul). Para cada método, o desempenho dos três tipos de assinaturas propostas são comparados. A base de dados de larga escala, que contempla *pockets* de enzimas, foi empregada nesse experimento, bem como o algoritmos de classificação KNN.

5.3.1.1 Avaliação dos Tipos de Assinaturas

Na Figura 5.6, comparamos as assinaturas aCSM, aCSM-HP e aCSM-ALL em termos dos valores de AUC obtidos em tarefas de predição de ligantes para a base de enzimas de larga escala, considerando um número diferente de valores singulares usados para aproximação da original. Por um lado, podemos ver que, quanto mais específica a assinatura é em termos das propriedades físico-químicas dos átomos mais precisa esta é na predição de ligantes. Com 100 valores singulares, para *pockets* extraídos através de um critério de distância (três curvas superiores em tons de azul), a assinatura aCSM-ALL atinge seu máximo em termos de AUC em 0,92, enquanto a assinatura básica (e puramente topológica) aCSM apresenta uma AUC de 0,75. Por outro lado, com menos de 20 valores singulares, a diferença entre as diferentes assinaturas é quase nula, alcançando valores relativamente altos de ACU, por volta de 0,85.

É interessante notar que aCSM-ALL é o única que parece ter se beneficiado a adição de um número grande de valores singulares. O primeiro valor singular

responde pela variabilidade mais alta dos dados e são os mais informativos. Enquanto adicionamos valores singulares para a assinatura, também é introduzido ruído aos dados, bem como é exigido mais tempo computacional para seu processamento. Esses resultados mostram que a aCSM apresenta uma limitação intrínseca quando 20 valores singulares são considerados, onde seu pico de desempenho é atingido (uma AUC de cerca de 0,86 é obtida). A aCSM-HP chega a mais de 0,90 com cerca de 40 valores singulares. Na aCSM-ALL, a AUC é melhorada quando adicionamos sucessivos valores singulares e não converge até atingir 100 valores singulares. Isso pode indicar que o ruído é menor quando rotulamos os nós dos grafos atômicos de uma maneira mais precisa e específica.

Na Tabela 5.1, comparamos as três assinaturas propostas, aCSM, aCSM-HP e aCSM-ALL a partir de diversas métricas de qualidade bem como em termos do tempo médio de execução da etapa de classificação. Para obtenção do tempo médio de execução, foi executado o algoritmo de classificação 5 vezes, e apresentamos os tempos de usuários médios. Note que os melhores resultados foram alcançados pelo aCSM-HP e aCSM-ALL após o pré-processamento com SVD. O aCSM-ALL obteve um desempenho ligeiramente superior, muito embora leve o dobro do tempo para executar em comparação com o aCSM-HP.

No entanto, uma análise estatística feita através de um teste de proporção indica que esta diferença na precisão (0,818 contra 0,842, para aCSM-HP e aCSM-ALL, respectivamente, considerando-se um grande conjunto de dados com mais de 35.000 *pockets*) é altamente significativa (*p-value*: $2,2e^{-16}$). De fato, em números absolutos, classificamos corretamente 837 ligantes de *pockets* a mais usando a assinatura aCSM-ALL do que com a aCSM-HP. A fim de ser insignificante (num intervalo de 95% de confiança), como um artefato de variação resultante do processo de amostragem aleatória, a diferença entre as precisões deveria ser inferior a 0,004 (cerca de 142 *pockets*). Em suma, acreditamos que o aCSM-ALL é a melhor escolha dentre as assinatura propostas uma vez que seu tempo de execução não é proibitivo. É importante salientar que esses modelos foram construídas com o algoritmo KNN dado o grande custo de execução das assinaturas sem redução de dimensionalidade.

5.3.1.2 Influência do Método de Definição de *Pockets*

A Figura 5.6 também mostra a comparação das assinaturas aCSM computados para *pockets* delimitados por critérios de distância geométricos (três curvas de baixo, em tons de verde). Podemos ver que, com o uso do método geométrico, os resultados são sistematicamente cerca de 15% piores do que utilizando simplesmente os critérios

Tabela 5.1. Comparativo de desempenho e tempo de execução entre as três assinaturas propostas, com e sem a utilização da etapa de redução de ruído e dimensionalidade com o auxílio da SVD. Para esses experimentos, utilizamos o algoritmos KNN sobre a base de enzimas de larga escala. São exibidos além do tempo médio de execução (média de 5 execuções), métricas como precisão e revocação, além do número de dimensões consideradas para cada assinatura.

SVD	Assinatura	#Dim.	Precisão	Revoc.	AUC	Tempo(s)
Sem SVD	aCSM	151	0,602	0,607	0,803	699,8
Com SVD	aCSM	21	0,722	0,720	0,860	528,8
Sem SVD	aCSM-HP	453	0,706	0,709	0,855	1.735,4
Com SVD	**aCSM-HP**	**45**	**0,822**	**0,818**	**0,909**	**827,6**
Sem SVD	aCSM-ALL	5436	0,804	0,805	0,902	29.832,0
Com SVD	**aCSM-ALL**	**99**	**0,846**	**0,842**	**0,920**	**1.427,0**

de distância. Isto se deve provavelmente à perda de informação molecular importante quando se utiliza o algoritmo Fpocket.

Mesmo se adicionarmos mais átomos àqueles pockets definidos pelo critério de 5Å, nosso método ainda se comporta de forma robusta, sendo capaz de descartar informações desnecessárias ou irrelevantes. A Figura 5.7 mostra que para distâncias de corte superior a 5Å o desempenho na predição para nosso método pode até ser melhorado.

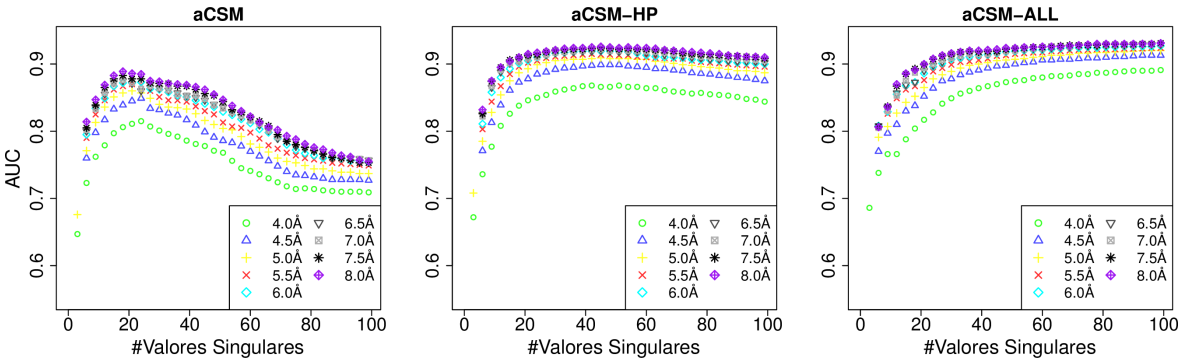

Figura 5.7. Desempenho comparativo das assinaturas de acordo com diferentes critérios de distância para definição dos *pockets*. Os gráficos mostram as diferenças de desempenho considerando a métrica AUC, para as três assinaturas propostas (aCSM, aCSM-HP e aCSM-ALL), para a base de dados de larga escala composta por enzimas.

Utilizamos 5Å como critério de corte, pois foi a mesma distância adotada pelos trabalhos concorrentes. No entanto, este valor parece ser definido arbitrariamente e não reflete necessariamente o melhor possível ponto de corte para cada método. Para

avaliar esta hipótese comparamos o desempenho de nossas assinaturas, para a maior base de dados considerada, variando a distância de corte para definição dos *pockets*.

De fato, na Figura 5.8, mostramos que o melhor critério de distância para a assinatura aCSM, usando o classificador KNN, era na verdade 6Å. Este valor está de acordo com outros autores que também têm investigado sobre o melhor ponto de corte para geração de redes de contatos atômicos entre átomos pesados usando um critério de proximidade [Zhang et al., 1997; Kamagata & Kuwajima, 2006]. É importante salientar que, com este corte, obtemos o mínimo de ruído nas nossas assinaturas, ao passo que o corte da SVD que maximiza a AUC é escolhido.

Acreditamos que a diferença no desempenho observado na Figura 5.6 é devida à limitação de o método geométrico em encontrar o conjunto preciso de átomos de proteínas que se encontram em contato com o ligante.

Como podemos ver na Figura 5.9, há uma grande discrepância entre os pockets encontrados pelo FPocket e aqueles definidos por um critério de distância (usamos 5,0Å), o que significa que o FPocket pode encontrar vários bolsões na estrutura, mas tem sérias limitações em encontrar sítios de ligação reais.

A Figura 5.10 mostra o percentual dos átomos definidos por um critério de distância, que também são encontrados pelo Fpocket. Neste caso, utilizou-se os ligantes e *pockets* do estudo de caso, que corresponde a proteínas de *T. cruzi*. Podemos notar que, em média, a intersecção entre a definição dos *pockets* não passa de 50%.

O Fpocket cobre, em geral, apenas um pequena porção dos átomo selecionados pelos critérios de distância e adicionalmente pode incluir átomos que estão muito distantes do ligante (como o exemplo da Figura 5.9). Existem ainda casos de *pockets* na superfície da proteína onde o método geométrico falha totalmente em definir sequer um único átomo em contato com o ligante, como exibido na Figura 5.11.

Acreditamos que atributos geométricos não são suficientes, em muitos casos, para delimitar um sítio de ligação. Na verdade, outras características como propriedades físico-químicas e conservação de resíduos pode melhorar métodos puramente geométricos, como o FPocket.

5.3.2 Análise comparativa

Os experimentos cujos resultados são descritos abaixo foram realizados sobre dois conjuntos de dados (*Hoffmann HD* e *Kahraman*) já utilizado por vários estudos relacionados, a fim de compará-los com a nossa assinatura de *pockets* baseada em grafos atômicos *aCSM*. Validação cruzada na forma *leave-one-out* foi utilizado em todos os experimentos relativos a esses dois conjuntos de dados, a mesma metodologia utilizada

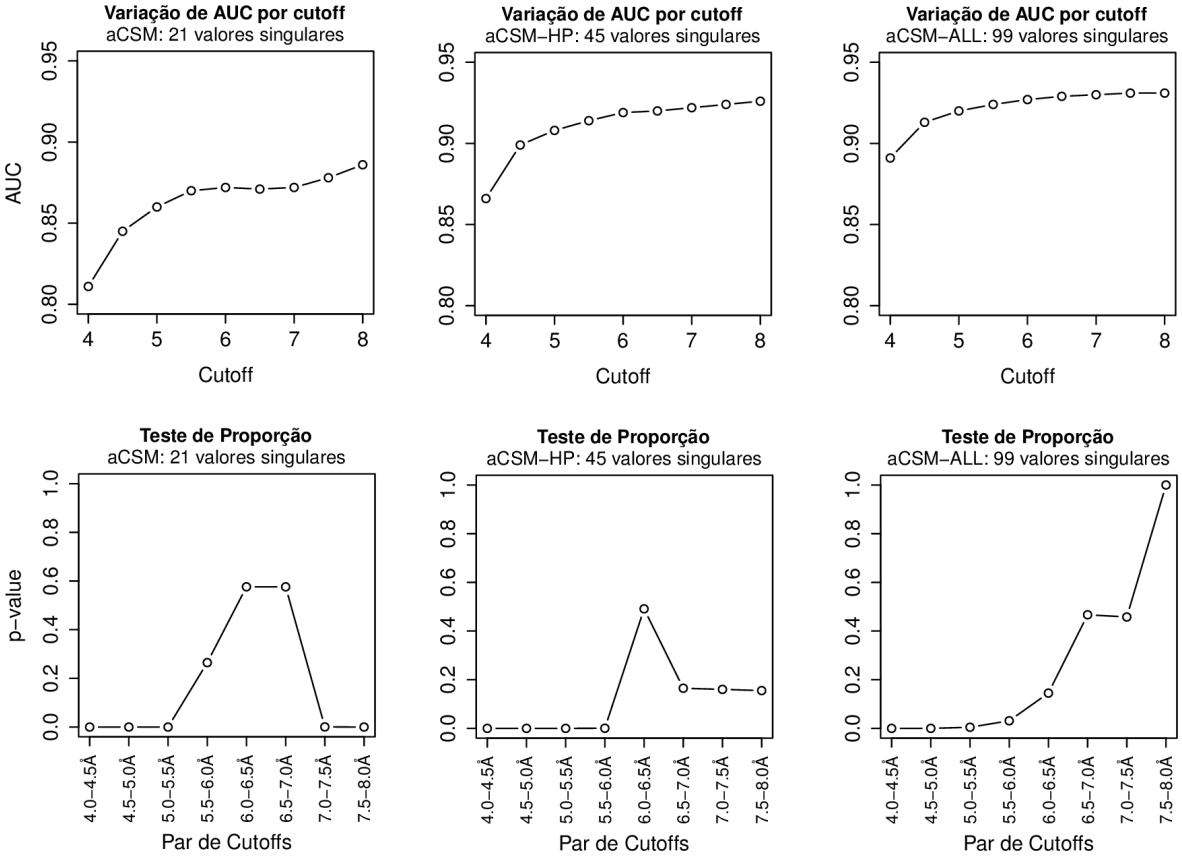

Figura 5.8. Análise estatística da métrica AUC em função da distância de corte utilizada para definição dos *pockets*, para as assinaturas propostas, considerando diferentes números de valores singulares para redução de dimensionalidade. Os números de valores singulares selecionados correspondem aos valores máximos de AUC obtidos para cada tipo de assinatura. Para mensurar a significância estatística dessas séries de valores de AUC, realizamos testes de proporção de duas caudas contra a hipótese nula de similaridade do valor de AUC, considerando um conjunto universo de 35.000 instâncias (*i.e.*, *pockets*). Esse experimento foi realizado a partir de um *script* implementado na linguagem de programação R, versão 2.12.1. Consideramos que um *p-value* mais que 0.05 indica que as diferenças nas proporções não são significativas e podem ocorrer devido à variações na amostra. Podemos notar valores altos de *p-value* para as distâncias entre 6.0Å to 7.0Å para todos os tipos de assinaturas, o que significa que nesse intervalo de distâncias não há ganho significativo de informação quando mais átomos são adicionados ao *pocket*. Nesse sentido, podemos concluir que 6.0Å é a melhor distância de corte para definição de *pockets* para nosso sistema de classificação.

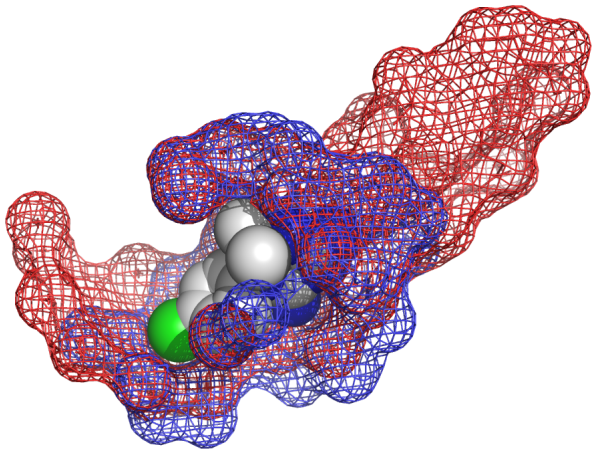

Figura 5.9. Dois métodos para definição de *pockets*. Neste exemplo, para a estrutura de identificador PDB 3IRM:C, mostramos dois resultados bastante distintos para definição do *pocket* do ligante Cicloguanil (1CY). Em azul, uma distância máxima de 5Å é utilizada e em vermelho é exibido o *pocket* mais próximo do ligante encontrado pelo FPocket, que corresponde à uma abordagem geométrica baseada na teoria de *alpha-shapes*.

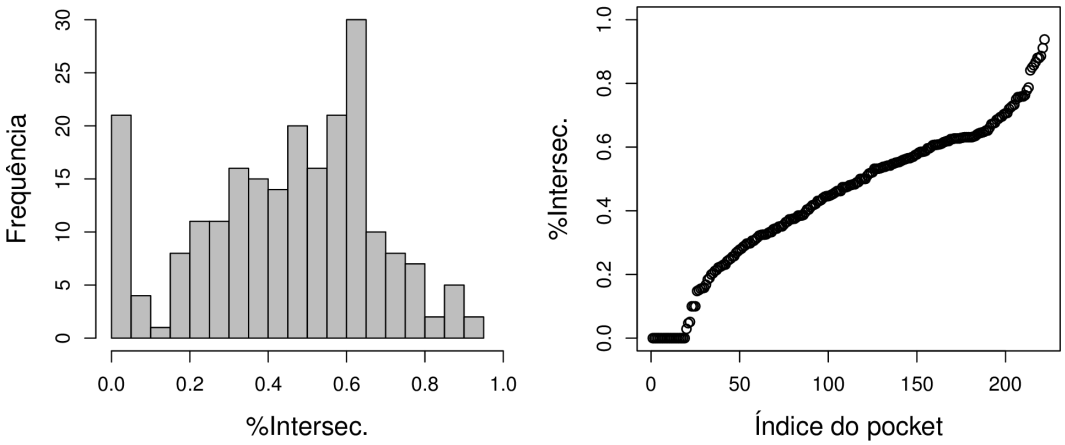

Figura 5.10. Intersecção entre metodologias de definição de *pockets*. No gráfico da esquerda, é exibido um histograma da percentagem dos átomos pertencentes a *pockets* definidos via 5Å que também foram encontrados pelo método geométrico (FPocket). No gráfico da direita, é exibida a distribuição do percentual de intersecção de átomos por *pocket* (em ordem crescente). Os *pockets* utilizados nesse caso foram obtidos da base de dados do estudo de caso de proteínas de *T. cruzi*. Note que apenas uma pequena parte dos átomos em contato com o ligante são, de fato, são incluídos pelo pocket mais próximo retornado pelo FPocket.

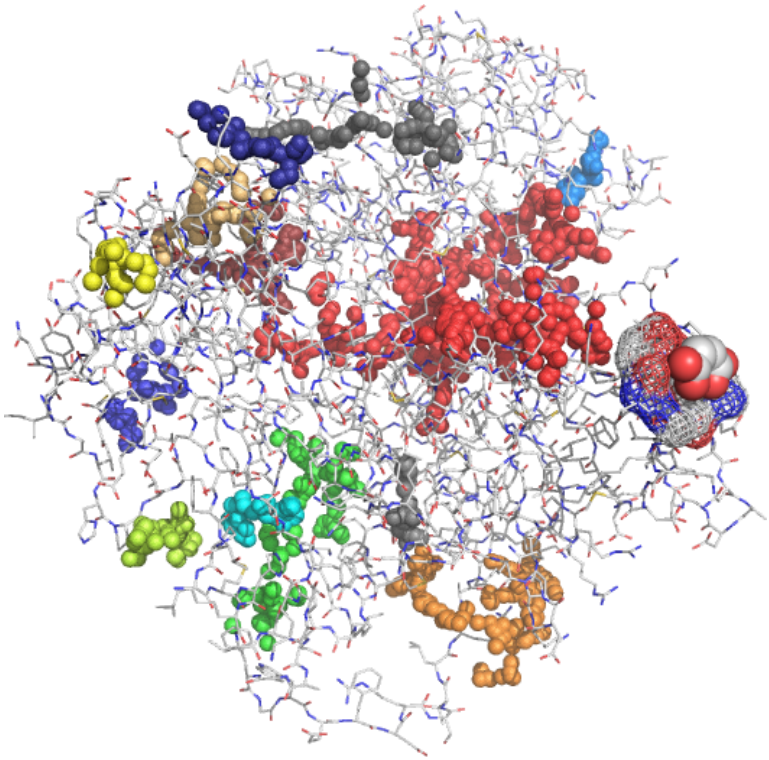

Figura 5.11. Exemplo de falha do FPocket. Na Figura são exibidos os pockets resultantes da execução do FPocket em *spacefill* (cada *pocket* de uma cor) para o PDB id 1AOGA:A. O ligante é mostrado também em *spacefill* e seu *pocket* real (delimitado via 5Å) é exibido em forma de representação em malha. Nesse caso, nenhum dos vários *pockets* retornados pelo FPocket teve um átomo sequer a uma distância inferior a 5Å do ligante.

nos trabalhos competidores.

A Tabela 5.2 resume os resultados obtidos. A assinatura aCSM obteve resultados compatíveis ou melhores, considerando valores de AUC, em comparação com outros métodos, também apresentando um desvio padrão mais baixo. É importante ressaltar que a validação cruzada *leave-one-out* é computacionalmente cara e não é adequada para experimentos em larga escala, que seriam mais aderentes a um cenário real. Além disso, os dois conjuntos de dados acima mencionados são pequenos (apenas 100 *pockets* cada) e, no caso da base Karahman, divididos em classes muito desbalanceadas, o que torna o processo de aprendizagem dos classificadores muito difícil.

5.3.3 Estudo de Caso: Predição de Ligantes para Proteínas de *T. cruzi*

A doença de Chagas é uma infecção tropical causada pelo protozoário *T. cruzi* que afeta por volta de 8 milhões de pessoas na América Latina [Rassi Jr. et al., 2010] e é

Tabela 5.2. Resultados comparativos avaliados pela média e desvio padrão da AUC. A assinatura aCSM-ALL obteve o melhor desempenho para esses experimentos. Os valores de AUC foram obtidos diretamente de [Hoffmann et al., 2010; Spitzer et al., 2011] e os resultados para a assinatura aCSM foram obtidos utilizando a Regressão Logística Multinomial.

Método	**Base de dados**	**AUC**
Sequence	Kahraman	0.550 ± 0.08
MultiBind	Kahraman	0.715 ± 0.17
SHD	Kahraman	0.770[1]
PSIM	Kahraman	0.790 ± 0.19
sup-PI	Kahraman	0.815 ± 0.13
sup-CK_L	Kahraman	0.861 ± 0.13
aCSM	**Kahraman**	$\mathbf{0.901 \pm 0.07}$
Sequence	Hoffmann HD	0.577 ± 0.09
MultiBind	Hoffmann HD	0.690 ± 0.14
sup-PI	Hoffmann HD	0.702 ± 0.19
sup-CK_L	Hoffmann HD	0.752 ± 0.16
PSIM	Hoffmann HD	0.760 ± 0.15
aCSM	**Hoffmann HD**	$\mathbf{0.804 \pm 0.13}$

a etiologia líder para doença cardíaca não-isquêmica em todo o mundo. A despeito da gravidade e impacto da doença na sociedade, apenas dois medicamentos (Nifurtimox e Benznidazol) estão disponíveis para tratamento da doença e ambos possuem efeitos colaterais tóxicos e eficácia variável [Canavaci et al., 2010]. As limitações do tratamento disponível atualmente tem motivado diversos esforços para desenvolvimento de novas drogas ou mesmo vacinas para o combate do *T. cruzi*. Além disso, um estudo recente [Lee et al., 2010a] propôs um modelo de Markov de decisão que mostrou que tal vacina traria um benefício econômico substancial às regiões que mais sofrem com a doença. Algumas abordagens recentes descritas na literatura par esse problema incluem esforços de *screening* de novos inibidores para alvos conhecidos de *T. cruzi* e o desenvolvimento de técnicas de validação em larga escala de compostos anti-*T. cruzi* [Canavaci et al., 2010].

Nesta seção, utilizamos os modelos de classificação treinados para predizer novos ligantes potenciais para proteínas de *T. cruzi* cujas estruturas estão disponíveis no PDB.

Após uma análise extensiva das assinaturas propostas selecionados os modelos com melhor desempenho treinados na maior base de dados considerada (que agrega mais de 35.000 *pockets*). As assinaturas dos *pockets* obtidos de proteínas de *T. cruzi*

[1]Desvio padrão não informado pelos autores.

foram fornecidas aos modelos treinados e um único ligante foi atribuído a cada *pocket*.

De modo a validar as predições realizadas foram projetados experimentos de *docking* dos ligantes nos pockets de *T. cruzi*, com o auxílio do programa AUTODOCK [Goodsell et al., 1996]. Em seguida, contrastamos as energias de ligação estimadas calculadas para os ligantes preditos pelas assinaturas aCSM com aquelas de complexos reais, informação calculada por um procedimento de *redocking*. Para avaliar a significância estatística do método, comparamos nossos resultados com um modelo nulo. Selecionamos, para cada *pocket*, três ligantes ao acaso dentre aqueles considerados na etapa de treino do classificador, ao acaso e de maneira independente. O *workflow* adotado no presente trabalho para execução do *docking* é exibido na Figura 5.12.

Na Figura 5.13, podemos ver que a distribuição de energia obtida para predições realizadas pelas assinaturas aCSM são mais semelhantes ao perfil daquelas obtidas nos experimentos de *redoking*, em relação aos dados obtidos pelos modelos nulos. De fato, testes-t pareados revelam um alto valor de *p-value* (0,26) entre as médias de energia da predição aCSM e do *redocking*, mas um baixo valor de *p-value* $(1,2e^{-9})$ entre as médias das predições via aCSM e os modelos nulos. Isso sugere fortemente que os ligante encontrados pelas assinaturas aCSM podem ter o mesmo perfil energético dos ligantes de *redocking*, mas esse perfil pode ser significativamente diferente daqueles obtidos pelos modelos nulos. Em resumo, mostramos que a energia livre de ligação para ligantes preditos pelo aCSM são melhores (mais baixas) em comparação com aquelas obtidas via modelos nulos. Além disso, as energias obtidas pelo protocolo de redocking são indistinguíveis daquelas obtidos para pela predição aCSM.

5.4 Conclusões

No presente trabalho, propusemos uma nova e escalável assinatura para *pockets* proteicos baseada em grafos atômicos denominada aCSM. Tal assinatura agrega padrões de distâncias em grafos de átomos proteicos que compõe *pockets* de ligação, gerando um vetor que representa uma distribuição acumulada do número de arestas de grafos de contatos definidos para diferentes distâncias de corte. Essa informação é, então, utilizada como evidência para algoritmos de aprendizado supervisionado. A Decomposição em Valores Singulares é também utilizada como um passo de pré-processamento para reduzir a dimensionalidade da assinatura, diminuir os custos computacionais e garantir escalabilidade à metodologia, bem como reduzir o ruído inerente aos dados, o que como consequência aumentou a taxa de sucesso das predições realizadas.

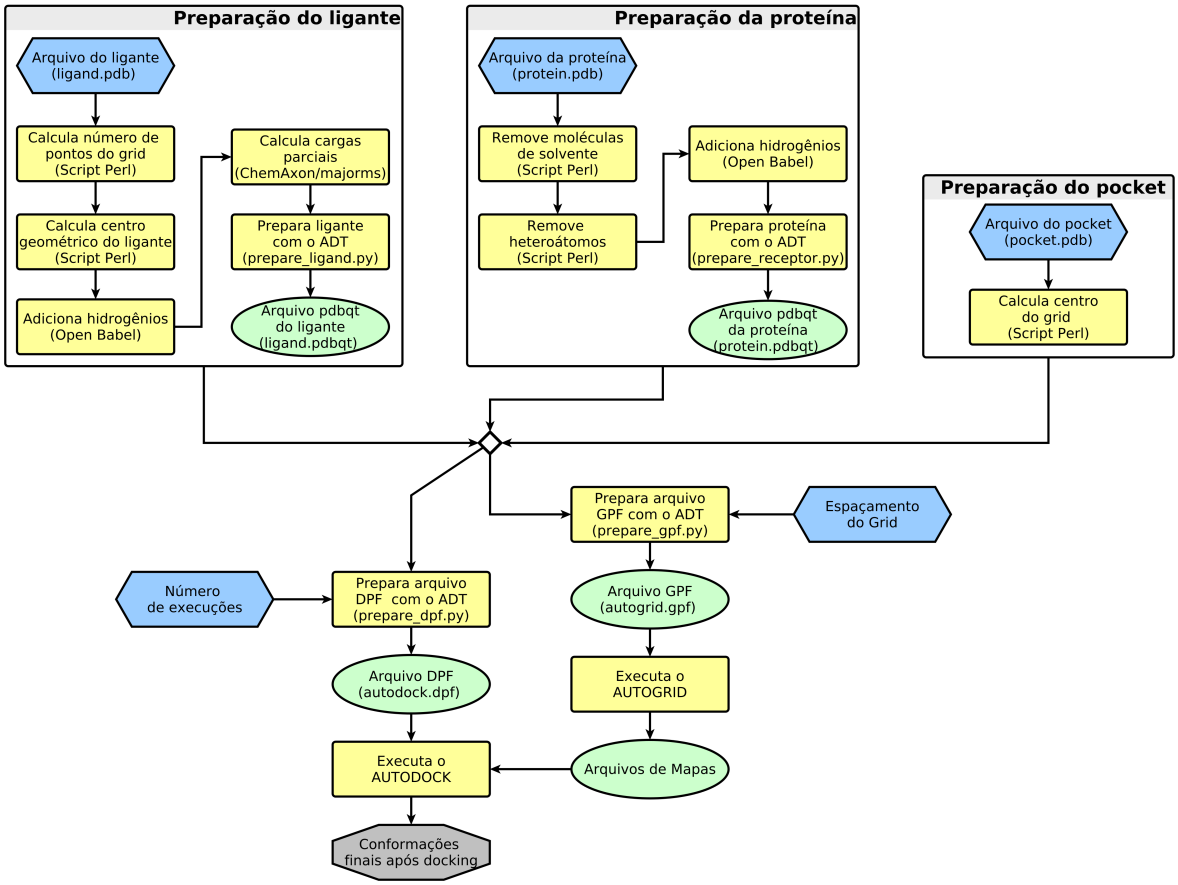

Figura 5.12. Protocolo de *docking*. Os procedimentos de *docking* foram realizados pelo programa AUTODOCK [Goodsell et al., 1996] e a partir de programas auxiliares como o Open Babel [O'Boyle et al., 2011], ChemAxon [ChemAxon, 2012], ADT [AutoDock, 2012], e um conjunto de programas implementados na linguagem de programação *scripting* Perl. Caixas azuis denotam arquivos ou parâmetros de entrada, caixas verdes são arquivos intermediários gerados no processo de preparação do *docking*, caixas amarelas denotam etapas de preparação, bem como os programas utilizados, e caixas cinzas a saída, *i.e.* as conformações obtidas pelo docking, bem como sua energia livre de ligação estimada.

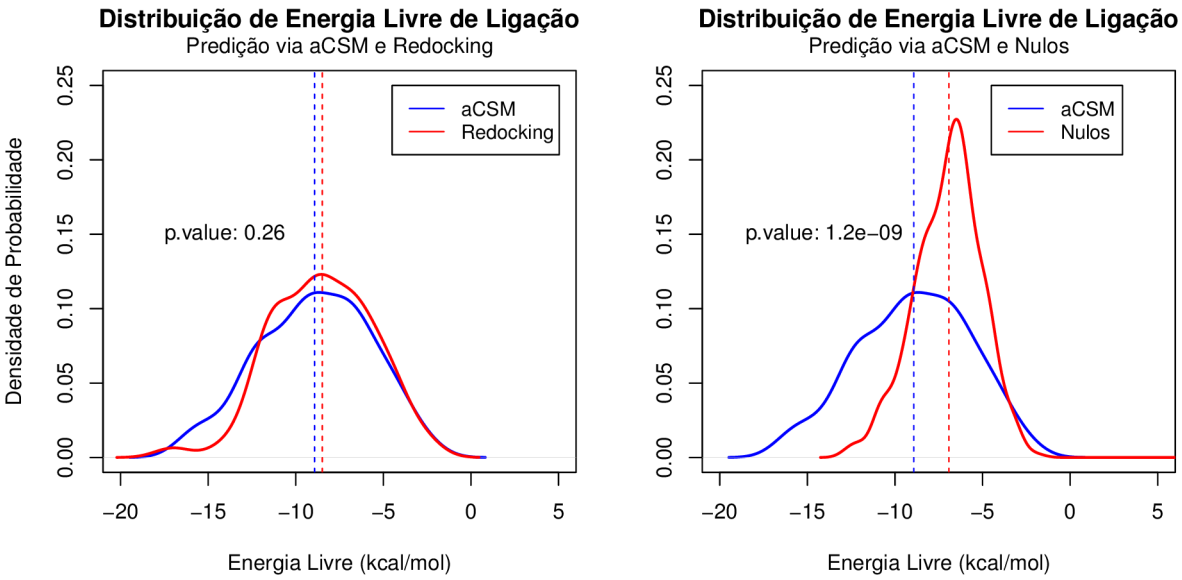

Figura 5.13. Análise comparativa das distribuições de energia livre de interação estimadas para complexos preditos pelas assinaturas aCSM, por meio de um procedimento de *redocking* e via modelos nulos. Linhas tracejadas indicam os valores médios. Os valores de *p-value* para testes *t-student* para a significância das médias também são apresentados. A energia livre de interação para ligantes preditos pelo aCSM são menores (melhores) em comparação com aqueles preditos pelos modelos nulos e são indistinguíveis daquelas energias obtidas através de um protocolo de *redocking*.

Algumas das vantagens mais notáveis das assinaturas aCSM é que elas não requerem qualquer informação do ligante para o seu cálculo, e também são independentes de orientação molecular. Além disso, nosso algoritmo apresenta uma notável generalidade ao passo que esse pode ser aplicado tanto na predição de ligantes proteicos quanto de não-proteico, para qualquer tipo de alvo biomolecular (não limitando-se apenas a proteína).

As assinaturas aCSM foram bem sucedidas quando aplicadas a tarefas de predição de ligantes, apresentando uma eficácia compatível ou superior em comparação com os métodos concorrentes estado-da-arte. Adicionalmente, como uma exigência e demanda para a sua aplicação em bancos de dados que estão em contínuo crescimento, a metodologia revelou-se escalável para cenários em larga escala, e foi capaz de ter uma boa eficácia para um conjunto de dados composto por mais de 35.000 *pockets*. Por fim, aplicamos a metodologia a fim de predizermos novos inibidores potenciais para proteínas de *T. cruzi*. A validação desse passo através da execução de *dockings*, confirmou que os inibidores preditos representam bons candidatos para futura investigação e validação experimental.

O trabalho descrito nesse Capítulo foi documentado na forma de artigo científico e encontra-se em fase de revisão. Mais informações podem ser encontradas no Apêndice B.

Capítulo 6

Predição de Toxicidade, Mutagênese e Atividade Anti-câncer

Com o grande aumento na disponibilidade de dados biológicos, uma demanda cada vez maior por paradigmas, modelos e algoritmos eficientes e escaláveis para suporte à descoberta de novos compostos candidatos (*lead compounds*) a fármacos tem surgido nesse contexto.

O uso de métodos computacionais nos dias atuais permeia todos os aspectos da descoberta de novas drogas, propiciando a entrega de novos compostos candidatos de maneira mais rápida e com menor custo em comparação com abordagens tradicionais [Jorgensen, 2004] e ainda em estágios iniciais de validação.

Adicionalmente, repositórios *online* como o DrugBank [Wishart et al., 2006] e o PubChem [Wang et al., 2009] dão suporte à estratégias *in silico* e compõem um recurso de bioinformática e quimioinformática que agrega dados detalhados e abrangentes de compostos e alvos terapêuticos.

Apesar dos grandes avanços e contribuição de áreas como a bioinformática na descoberta de fármacos, a indústria farmacêutica vem sofrendo uma crise de produtividade em termos do lançamento de novos fármacos a despeito do aumento de investimentos no setor [Pammolli et al., 2011]. Nesse sentido, existe uma demanda crescente por modelos computacionais mais adequados e eficazes que possam auxiliar nas diversas tarefas relacionadas à descoberta de novos fármacos.

Uma representação matemática/computacional intuitiva de moléculas químicas pode ser obtida a partir de grafos. Em um grafo molecular, os átomos podem ser representados como vértices e as ligações químicas entre seus átomos como as arestas. Adicionalmente, pode ser criada uma função de rotulação l que mapeia cada vértice v a um rótulo $l(v)$ que pode denotar alguma propriedade relevante de cada átomo.

A partir de tais modelos deseja-se identificar padrões topológicos dos grafos que caracterizem as pequenas moléculas quanto à toxicidade, atividade biológica e mutagênese. Para tal, propriedades topológicas e químicas podem ser extraídas a partir desses grafos e sumarizadas em uma assinatura molecular. Tais assinaturas podem, então, ser utilizadas na proposta de um modelo preditivo de caracterização desses compostos químicos.

No presente Capítulo descrevemos uma assinatura para grafos moleculares (doravante chamada gCSM) baseada no conceito CSM que é aplicada em tarefas preditivas que tentam avaliar a toxicidade de compostos químicos, seu potencial mutagênico e sua atividade quanto inibidor de crescimento carcinomas em diferentes tecidos humanos.

6.1 Modelagem Computacional

Na presente Seção descrevemos a modelagem computacional realizada que corresponde à representação das pequenas moléculas químicas como grafos atômicos.

- **Grafo molecular:** grafo simples, não-direcionado, não-ponderado, rotulado;
- **Nós:** correspondem a todos os átomos da pequena molécula;
- **Arestas:** representam as ligações químicas (covalentes) entre as moléculas;
- **Rótulo dos nós:** nós são rotulados de acordo com suas características fisico-químicas em dois níveis de especificidade.

Em suma, as pequenas moléculas químicas foram modeladas como grafos onde, os nós representam os átomos e as arestas as ligações covalentes entre esses. A rotulação dos nós por características fisico-químicas foi realizada em dois níveis de especificidade, dando origem a duas diferentes assinaturas gCSM:

- **gCSM**: assinatura sem rotulação, que por sua vez gera um valor por *cutoff*, correspondendo ao número de átomos a uma dada distância de corte, considerando uma métrica de distância.
- **gCSM-ALL**: assinatura categoriza os átomos em dez classes: positivo, negativo, aromático, hidrofóbico, aceptor, doador, aceptor/doador, positivo/doador, negativo/aceptor e neutro. A combinação desses rótulos gera 55 valores por *cutoff*. A classificação dos átomos foi obtida a partir do programa PMapper [ChemAxon,

2012] em pH 7. O PMapper identifica propriedades farmacofóricas de átomos a partir de suas estruturas moleculares.

A métrica de distância entre os nós dos grafos utilizadas foi o caminho mais curto. Para calcular os caminhos mais curtos entre os pares de vértices dos grafos, gerando um matriz de distâncias, foi utilizado o Algoritmo de Johnson [Johnson, 1977].

6.1.1 Métodos

Nesta seção descrevemos o fluxo metodológico empregado, bem como o procedimento para geração das assinaturas, redução de ruído e dimensionalidade das mesmas e, por fim, explicamos como o método foi avaliado e validado.

No presente Capítulo, utilizamos o conceito CSM na definição de uma assinatura para grafos que representam pequenas moléculas. As assinaturas gCSM são geradas da seguinte forma: para cada grafo molecular criamos um vetor de atributos. Em primeiro lugar, calcula-se a distância entre todos os pares de átomos, de modo a gerar uma matriz de distâncias. Utilizamos como métrica de distância o comprimento do caminho mínimo entre dois nós. Definimos também um intervalo de distâncias a ser considerado bem como um passo de distância. As distâncias desse intervalo são percorridas, calculando-se a frequência de arestas do grafo cujo peso é menor que a distância corrente. Como resultado, temos uma distribuição acumulada do número de arestas no grafo para um dado intervalo de distâncias. A contagem da frequência de arestas por distância é discretizada pelos tipos de rótulos dos nós. O Algoritmo 4 exibe a função que calcula a gCSM.

Algorithm 4 Cálculo da assinatura gCSM

```
1: function gCSM(ConjuntoCompostos, ClassesAtomos, D_MIN, D_MAX, D_PASSO)
2:     for all molecula i ∈ (ConjuntoCompostos) do
3:         j = 0
4:         matrizDist ← calculaCaminhosMinimosJohnson(molecula)
5:         for distancia ← D_MIN; até D_MAX; passo D_PASSO do
6:             for all classe ∈ (ClassesAtomos) do
7:                 gCSM[i][j] ← obtemFrequencia(matrizDist, distancia, classe)
8:                 j++
9:     retorna gCSM
```

A geração de assinaturas gCSM é executada com complexidade de tempo de $O(n^2 log(n))$, onde n é o número de átomos da molécula (ou nós do grafo), e corresponde ao tempo de cálculo das distâncias par-a-par entre os nós do grafo, ou seja, o tempo necessário para que todos os caminhos mínimos seja computados.

Nos experimentos apresentados neste capítulo, consideramos o intervalo de distâncias de 0-50 arestas, com um passo unitário, o que gerou vetores de atributos de 51 e 2805 dimensões, para cada molécula, para as assinaturas gCSM e gCSM-ALL, respectivamente.

A Decomposição em Valores Singulares (SVD) foi utilizada para reduzir ou eliminar o ruído inerente das assinaturas geradas e, por consequência, melhorar a eficácia e reduzir o tempo de processamento dos algoritmos de classificação, em termos de tempo de execução e os requisitos de memória.

6.1.2 Tarefas de Classificação

Para as tarefas preditivas executadas foram treinados classificadores com o algoritmo Random Forest [Breiman, 2001], um classificador *ensemble* que consiste de várias árvores de decisão que gera como saída a classe mais frequente dentre aquelas obtidas pelas árvores de decisão individuais. Esse algoritmo de aprendizado é reconhecidamente eficaz, bem como eficiente para bases de dados de larga escala.

Os tipos de validação utilizada nos experimentos comparativos foram a *leave-one-out* e validação cruzada em 5 partições, mesmos procedimentos empregados por métodos competidores. As métricas de avaliação AUC e Acurácia também foram utilizadas pelo mesmo motivo.

Para todas as tarefas de classificação, o Weka Toolkit [Hall et al., 2009] *developer version 3.6.2* foi utilizado.

6.1.3 Bases de Dados

Com o intuito de obter uma ampla variedade de experimentos para validar a assinatura proposta, bem como sua generalidade e aplicabilidade em diversos cenários reais, utilizamos três bases de dados com finalidades diferentes.

a. **Atividade anticâncer**: utilizada por um método concorrente [Yan et al., 2008], corresponde a 11 conjuntos de dados de atividade anticâncer de pequenas moléculas disponíveis no PubChem [Wang et al., 2009]. O PubChem fornece informações sobre ensaios bioquímicos de atividade biológica de moléculas pequenas, que contêm um grande número de registos de testes para atividade anticâncer de pequenas moléculas para diferentes linhagens celulares de câncer, em diversos tecidos humanos. Esses conjuntos são descritos na Tabela 6.1. Cada conjunto de dados pertence a um determinado tipo de câncer e as moléculas são rotuladas como sendo *ativas* ou *inativas*. Para todos os conjuntos de dado relativos ensaios

experimentais, foi aleatoriamente selecionado 5% dos compostos ativos e uma quantidade comparável de inativos de cada conjunto de dados, de modo a ter-se uma amostra compacta e balanceada. Também é derivada uma base estendida com um número maior de compostos ativos (6x mais compostos). O número de vértices na maioria destes compostos varia de 10 a 200.

b. **Mutagênese**: utilizada por diversos trabalhos descritos na literatura, a base MUTAG [Debnath et al., 1991] é formada por 188 compostos, cada um possuindo um valor indicando a existência ou não de mutagenicidade no organismo *Salmonella typhimurium*, sendo 125 positivos e 63 negativos.

c. **Toxicidade**: a base de dados do *Predictive Toxicology Challenge* (PTC [Helma et al., 2001]) foi utilizada para prever a capacidade das assinaturas gCSM na predição de toxicidade de pequenos compostos. O conjunto de dados PTC é classificado por carcinogenicidade e é dividida em 4 grupos de acordo com o tipo de organismo considerado no ensaio biológico: rato macho (MR - *male rat*), camundongo macho (MM - *male mouse*), rato fêmea (FR - *female rat*) e camundongo fêmea (FM - *female mouse*).

Tabela 6.1. Bases de ensaios bioquímicos quanto a atividade anticâncer obtidos a partir do PubChem [Wang et al., 2009]. Cada conjunto de dados pertence a um determinado tipo de câncer e as moléculas são rotuladas como sendo *ativas* ou *inativas*.

Nome da Base	ID	#Compostos	#Ativos	Descrição do Tumor
MCF-7	83	28.420	2.357	Câncer de Mama
MOLT-4	123	40.614	3.200	Leucemia
NCI-H23	1	40.989	2.104	Câncer Pulmonar
OVCAR-8	109	41.159	2.128	Câncer Ovariano
P388	330	44.492	2.349	Leucemia
PC-3	41	28.012	1.623	Câncer de Próstata
SF-295	47	40.911	2.081	Sistema Nervoso Central
SN12C	145	40.630	2.007	Câncer Renal
SW-620	81	41.172	2.464	Câncer de Cólon
UACC257	33	40.678	1.690	Melanoma
Yeast	167	80.492	9.615	Atividade em Leveduras

6.2 Trabalhos Relacionados

A classificação de pequenas moléculas tem sido abordadas na literatura sob diferentes perspectivas. Sub-estruturas moleculares podem ser mineradas por algoritmos tradicionais de busca de subgrafos frequentes e utilizadas como evidências nas tarefas de classificação. Estratégias mais eficientes, que reduzem a complexidade dessa tarefa de mineração tem sido reportadas, como o caso do método LEAP [Yan et al., 2008], aqui utilizado como comparativo em diversos experimentos. Também é válido destacar os esforços baseados na seleção de subgrafos relevantes [Kong & Yu, 2010; Kong et al., 2011].

Outras abordagens possíveis incluem a proposta de *kernels* de grafos que tentam quantificar a similaridade dos compostos químicos, a um custo computacional menor, em relação à mineração de subgrafos frequentes. Essas abordagens tem sido utilizadas com relativo sucesso em tarefas de predição de toxicidade, mutagênese e atividade anticâncer [Fröhlich et al., 2005; Swamidass et al., 2005; Kashima et al., 2003]. O método *2D Tanimoto* descrito em [Swamidass et al., 2005] e o método *PD* descrito em [Kashima et al., 2003] também são utilizados nos experimentos comparativos.

Uma revisão sobre diferentes *Kernels* de grafos para similaridade de moléculas é descrita em [Rupp & Schneider, 2010]. Estudos sobre métodos de predição de toxicidade *in silico* aplicada ao desenvolvimento de fármacos são reportados em [Muster et al., 2008; Valerio, 2009].

6.3 Resultados

De modo a testar a capacidade de nossa metodologia predizer toxicidade, mutagênese e atividade anticâncer de pequenas moléculas, executamos três conjuntos de experimentos, projetados para essas três diferentes tarefas.

6.3.1 Predição de Toxicidade e Mutagênese

A Tabela 6.2 exibe os resultados comparativos para as tarefas de predição de toxicidade e mutagênese. Os resultados foram obtidos com validação do tipo *leave-one-out* e os valores médios de Acurácia são exibidos para as duas variações de assinatura gCSM propostas e outros dois métodos concorrentes, descritos em [Swamidass et al., 2005]. Podemos verificar que nossa abordagem apresenta resultados equiparáveis ou até superiores aos métodos concorrentes tanto para a base MUTAG, quanto para as quatro bases do *Predictive Toxicology Challenge*. Para esses experimentos não observou-se um

ganho tão significativo ao considerarmos a rotulação dos nós dos grafos pelo PMapper. Isso pode indicar que uma rotulação mais genérica, considerando um número menor de classes seja mais apropriada nesse caso, fato que pretendemos investigar em trabalhos futuros.

Tabela 6.2. Resultados comparativos avaliados em relação à métrica de Acurácia para as tarefas de predição de toxicidade e mutagênese induzida por pequenas moléculas. Valores de Acurácia para os trabalhos concorrentes foram obtidos diretamente de [Swamidass et al., 2005]. Nestes experimentos foi empregada a validação do tipo *leave-one-out*. É exibido o resultado para o melhor corte obtido pelo SVD.

Base de dados	gCSM-ALL	gCSM	PD	2D Tanimoto
MUTAG	**0.904**	0.894	0.891	0.878
PTC-MM	0.652	0.652	0.610	**0.664**
PTC-FM	**0.648**	0.642	0.610	0.642
PTC-MR	0.622	0.616	0.628	**0.637**
PTC-FR	0.664	0.661	**0.667**	**0.667**

6.3.2 Predição de Atividade Anti-câncer

A Tabela 6.3 exibe os resultados para predição de atividade anticâncer de pequenas moléculas, tendo como métrica de sucesso os valores de AUC. Os resultados foram obtidos com validação cruzada em 5 partições, mesmo procedimento adotado pelos métodos concorrentes. Nota-se que a gCSM foi superior em 10 das 11 bases consideradas por uma margem significativa. A base Yeast será averiguada e melhor caracterizada em trabalhos futuros de modo a verificar-se o que levou a um desempenho inferior de nossa metodologia em relação às demais bases.

Na Tabela 6.4 constam os resultados para as bases estendidas. Nestes fica ainda mais clara a superioridade da assinatura gCSM frente às abordagens concorrentes, bem como o expressivo ganho na taxa de sucesso ao considerarmos os rótulos farmacofóricos dos átomos na geração da assinatura.

6.4 Conclusões

No presente Capítulo, propusemos uma nova e escalável assinatura para grafos moleculares denominada gCSM, que é baseada na extração de padrões de distâncias em grafos que representam pequenos compostos químicos. Essas assinaturas foram, então, utilizadas como evidência para algoritmos de aprendizado supervisionado em tarefas

Tabela 6.3. Resultados comparativos avaliados em relação à métrica AUC para a tarefa de predição de atividade anticâncer de pequenas moléculas. Valores de AUC para os trabalhos concorrentes foram obtidos diretamente de [Yan et al., 2008]. Nestes experimentos foi empregada a validação cruzada em 5 partições. É exibido o resultado para o melhor corte obtido pelo SVD.

Base de dados	gCSM-ALL	gCSM	OA	LEAP
MCF-7	**0.749**	0.702	0.68	0.67
MOLT-4	**0.758**	0.686	0.65	0.66
NCI-H23	**0.814**	0.751	0.79	0.76
OVCAR-8	**0.799**	0.738	0.67	0.72
P388	**0.835**	0.762	0.79	0.82
PC-3	**0.796**	0.768	0.66	0.69
SF-295	**0.790**	0.745	0.75	0.72
SN12C	**0.834**	0.783	0.75	0.75
SW-620	**0.805**	0.762	0.70	0.74
UACC257	**0.804**	0.764	0.65	0.64
Yeast	0.672	0.629	**0.71**	0.64
Média	**0.787**	0.735	0.70	0.72

Tabela 6.4. Resultados comparativos avaliados em relação à métrica AUC para a tarefa de predição de atividade anticâncer de pequenas moléculas (considerando a base estendida). Valores de AUC para os trabalhos concorrentes foram obtidos diretamente de [Yan et al., 2008]. Nestes experimentos foi empregada a validação cruzada em 5 partições. É exibido o resultado para o melhor corte obtido pelo SVD.

Base de dados	gCSM-ALL(6x)	gCSM(6x)	OA(6x)	LEAP(6x)
MCF-7	**0.834**	0.769	0.75	0.76
MOLT-4	**0.825**	0.765	0.69	0.72
NCI-H23	**0.880**	0.816	0.77	0.79
OVCAR-8	**0.871**	0.809	0.79	0.78
P388	**0.879**	0.805	0.81	0.84
PC-3	**0.866**	0.811	0.79	0.76
SF-295	**0.859**	0.814	0.79	0.77
SN12C	**0.876**	0.823	0.76	0.80
SW-620	**0.863**	0.801	0.76	0.76
UACC257	**0.875**	0.807	0.71	0.75
Yeast	**0.728**	0.643	0.64	0.71
Média	**0.851**	0.788	0.75	0.77

de classificação. A Decomposição em Valores Singulares é também utilizada como um passo de pré-processamento para reduzir a dimensionalidade da assinatura, diminuir os custos computacionais e garantir escalabilidade à metodologia, bem como reduzir o ruído inerente aos dados.

As assinaturas gCSM obtiveram desempenho equiparável ou superior aos trabalhos concorrentes para as três tarefas em questão. Esse resultados podem suportar no futuro a definição de uma métrica de similaridade de compostos químicos e possível desenvolvimento de um sistema *web* de recuperação de moléculas similares.

O comportamento observado em tarefas de predição de toxidade, que diz respeito à ausência de ganho efetivo de precisão quando os rótulos dos nós são considerados na assinatura aponta para a necessidade de melhorias na modelagem dos grafos ou até mesmo na geração de um conjunto menor e mais genérico de rótulos, o que será estudado em trabalhos futuros.

Capítulo 7

Conclusões

A comunidade científica tem presenciado nos últimos anos uma mudança no paradigma que rege o desenvolvimento e o avanço científico. Os desafios e demandas criadas pelo grande ritmo de geração e disponibilização de dados em todas áreas do conhecimento tem moldado a maneira como a pesquisa tecnológica precisa ser conduzida.

O grande volume e dinamicidade de dados biológicos tem gerado desafios computacionais principalmente no que diz respeito ao processamento, e extração de conhecimento a partir dessas fontes de informações. A modelagem de dados biológicos na forma de grafos ou redes compõe estratégias promissoras e eficazes na modelagem de fenômenos, sistemas e processos naturais, dado o crescente número maior de sistemas reais podem ser modelados computacionalmente como grafos.

Nesse sentido, novos modelos, algoritmos e paradigmas necessários para que redes biológicas em larga escala sejam devidamente analisadas e os fenômenos que as governam, compreendidos.

No presente trabalho, foi apresentada uma nova metodologia para análise de grafos biológicos baseada na geração de assinaturas denominada Cutoff Scanning Matrix (CSM). O CSM gera vetores de atributos que representam padrões de distâncias entre nós de um grafo, que são então usados como evidência em tarefas de classificação. Adicionalmente, a Decomposição em Valores Singulares (SVD) é empregada como um passo de pré-processamento para reduzir a dimensionalidade e o ruído inerente aos dados. A metodologia proposta é instanciada com sucesso em redes biológicas bastante distintas, que se diferenciam tanto em termos da modelagem computacional empregada quanto do tipo de assinaturas geradas, aplicadas em tarefas como a anotação automática proteica, predição de ligantes e de atividade de pequenas moléculas.

Os resultados obtidos nas diferentes tarefas mostram que os padrões de distâncias em grafos correspondem a um componente robusto e conservado, sendo uma importante

fonte de informação topológica. Adicionalmente, as assinaturas derivadas do conceito CSM mostraram-se eficazes e eficientes na resolução das diversas tarefas propostas sendo comparáveis ou superiores aos principais trabalhos concorrentes. Por fim, a aplicabilidade do conceito CSM em diferentes grafos biológicos mostra, além de sua generalidade, um grande potencial ainda a ser explorado.

7.1 Perspectivas e Trabalhos Futuros

7.1.1 Anotação Automática

Dentre os principais desdobramentos e possíveis trabalhos futuros para a assinatura estrutural proteica CSM encontram-se:

Aplicação em outras tarefas de anotação: Como parte de estudos futuros, pretendemos explorar a generalidade do CSM em outros aspectos de função proteica, como predição de localização subcelular e predição de termos do GO (Gene Onthology), bem como sob outras diferentes bases de classificação estrutural, como o CATH [Orengo et al., 1997]. A eficácia assinatura estrutural baseada em CSM em tarefas correlatas mas diferentes como predição de classe estrutural e função é um indício de que a metodologia pode ser generalizável e desejamos testá-la em cenários mais complexos. Dentre outras possíveis aplicações para o CSM pretendemos avaliar uma proposta de anotação para domínios de função desconhecida, como os DUFs presentes no PFam, bem como avaliar quais os piores casos de classificação para o CSM.

Avaliação dos parâmetros do CSM: O ganho significativo em poder de predição provido pela redução de dimensionalidade via SVD pode implicar que ainda existe espaço para melhora no processo de geração do CSM, indicando que outros intervalos de *cutoff* e outras granularidades devem ser testadas.

Análise da semântica dos padrões encontrados: O fato do CSM ter apresentado eficácias comparáveis nos quatro níveis do SCOP pode indicar que a assinatura estrutural proposta captura características específicas a cada nível, desde mais gerais (nível de classe, identificação do tipo de estruturas secundárias que compõe as proteínas) até mais específicas (nível de família, identificação de similaridades estruturais específicas do empacotamento da proteína). A elucidação dessas características demanda investigação mais aprofundada. Nesse sentido, em etapas futuras do trabalho pretendemos averiguar a hipótese de que as características específicas de cada nível estejam relacionadas com faixas de valores de *cutoff*.

O sucesso do CSM no nível de família do SCOP indica que similaridade estrutural é bem computada pelo CSM. Uma vez que em casos de convergência evolutiva

enzimática estamos lidando com estruturas, de modo geral, bastante dissimilares mas que apresentam uma conservação do sítio catalítico e provavelmente de sua vizinhança, a aplicação do CSM em regiões delimitadas da proteína (como em sítios e *pockets*) passa a ser uma estratégia mais adequada para esses casos (ao invés de utilizar toda a proteína). Desejamos investigar mais detalhadamente essa possibilidade.

Técnicas de Mineração de Dados: Desejamos também realizar a classificação binária das bases. É importante ressaltar que os resultados aqui apresentados compreendem o uso de classificadores treinados com múltiplas classes. No pior caso foram consideradas mais 4000 classes para um único classificador (certamente esse não é o melhor cenário para a tarefa de mineração de dados projetada), o que demostra a riqueza de informação presente na assinatura CSM.

Análise semântica da SVD: Em uma etapa futura desejamos ainda, atribuir semântica aos valores singulares considerados avaliando a contribuição de cada atributo do CSM (*cutoff*) na composição desses valores singulares. Esperamos com essa análise avaliar quais valores (ou faixas de valores) agregam mais informação e quais apresentam mais ruído. Também planejamos contrastar a SVD com seleção de atributos como métodos de descoberta de informações discriminantes em assinaturas CSM.

Evolução convergente e divergente: Acreditamos que casos de evolução convergente (proteínas com mesma função mas estruturas bastante dissimilares) e divergente (proteínas muito similares mas com função diferente) sejam os grandes limitantes no desempenho da metodologia. Nesse sentido, realizaremos uma caracterização dos dois casos em função da classificação SCOP e dos números EC. Verificaremos para cada número EC a diversidade de classificações encontradas (nos quatro níveis) e, para cada família, a diversidade de números EC anotados. Uma possível solução para esse cenário, como já dito, seria a aplicação do CSM em porções da proteína (no sítio ativo e em sua vizinhança, caso esse seja conhecido), regiões onde espera-se uma maior conservação.

7.1.2 Predição de Ligantes

No que tange as tarefas de predição de ligantes, planejamos predizer inibidores de proteínas de organismos patogênicos de interesse. Pretendemos, com isso, expandir a utilização da assinatura para cenários como a descoberta de compostos líderes baseada nos ligantes.

A descoberta de compostos candidatos (também conhecidos como compostos líderes ou protótipos) é atualmente um fator limitante significativo no processo de descoberta de novas drogas para doenças tropicais como a malária, tuberculose,

leishmaniose e doença de Chagas [Nwaka & Hudson, 2006]. Existem motivações sociais e econômicas importantes para estudo e combate de doenças tropicais [Lee et al., 2010b], muitas delas negligenciadas.

É importante destacar a aplicabilidade do trabalho aqui proposto em cenários reais e de extrema importância para a indústria, não somente farmacêutica mas também agropecuária e agroquímica, setores de extrema relevância para o Brasil. A metodologia proposta pode ser aplicada de forma transparente para quaisquer alvos de interesse, partindo da busca por fármacos contra doenças em seres humanos e até mesmo para controle de doenças em rebanhos ou mesmo na busca por defensivos agrícolas. Assim, a proposta é oportuna ao passo que aborda questões atuais e relevantes para o desenvolvimento da Bioinformática Estrutural, em um contexto de alta disponibilidade e volume de informação a ser tratada.

Acreditamos também ser possível caracterizar alterações conformacionais devido à *induced fitting* e regulações alostéricas a partir dos padrões de distâncias extraídos pelas assinaturas aCSM.

7.1.3 Predição de Toxicidade, Mutagênese e Atividade Anti-câncer

Em relação à assinatura gCSM, proposta para grafos que representam pequenas moléculas, deseja-se desenvolver uma métrica de similaridade de moléculas com base nas assinaturas de modo a ser possível o desenvolvimento de sistema escalável, eficiente e eficaz de recuperação de moléculas similares.

Em última instância, deseja-se também avaliar o potencial da gCSM na predição de características denominadas **ADME-Tox** (**A**bsorção, **D**istribuição, **M**etabolismo, **E**xcreção e **Tox**icidade), que descrevem a adequação do composto enquanto potencial droga.

7.1.4 Caracterização das Redes para Aplicação do CSM

Foram relatados no presente trabalho, cenários onde a aplicação do modelo proposto é pertinente e bem sucedida. Desejamos, em um segundo momento, estabelecer os limites de aplicação do CSM a partir da caracterização dos grafos em termos de sua topologia, dentre outros possíveis atributos. Com isso esperamos delimitar um conjunto de requisitos topológicos dos grafos que garantam a eficácia das assinaturas, e permitam extrapolar o conceito CSM para outros domínios.

7.1.5 Grafos não-biológicos

Por fim, desejamos em última instância extrapolar o conceito por trás do funcionamento do CSM para outros contextos, incluindo redes não-biológicas. Acreditamos que ele possa ser utilizado sobre quaisquer conjuntos de dados cujas entidades possam ser modeladas como grafos, como em tarefas de classificação de documentos e imagens, abrangendo com isso uma ampla gama de aplicações possíveis.

Referências Bibliográficas

Abraham, H. & Lilien, R. H. (2010). LigAlign: flexible ligand-based active site alignment and analysis. *Journal of Molecular Graphics and Modelling*, 29(1):93–101.

Al Hasan, M.; Chaoji, V.; Salem, S.; Besson, J. & Zaki, M. (2007). Origami: Mining representative orthogonal graph patterns. Em *Proceedings of the 7th IEEE International Conference on Data Mining*, ICDM '07, pp. 153–162.

Al Hasan, M. & Zaki, M. (2009). Musk: Uniform sampling of k maximal patterns. Em *Proceedings of the SIAM International Conference on Data Mining*, pp. 650–661.

Al Hasan, M. & Zaki, M. J. (2009). Output space sampling for graph patterns. *Proceedings of the VLDB Endowment*, 2(1):730–741.

Altigan, A. R.; Akan, P. & Baysal, C. (2004). Small-World communication of residues and significance for protein dynamics. *Journal of Molecular Biology*, pp. 85–91.

Alvarez, M. A. & Yan, C. (2010). Exploring structural modeling of proteins for kernel-based enzyme discrimination. Em *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–5.

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. the gene ontology consortium. *Nature Genetics*, 25(1):25–29.

Ausiello, G.; Gherardini, P. F.; Marcatili, P.; Tramontano, A.; Via, A. & Helmer-Citterich, M. (2008). FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics*, 9 Suppl 2:S2.

AutoDock (2012). AutoDock Tools
http://autodock.scripps.edu/resources/adt.

Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V. & Edelman, M. (2008). Prediction of transition metal-binding sites from apo protein structures. *Proteins: Structure, Function and Bioinformatics*, 70(1):208–17.

Barker, J. A. & Thornton, J. M. (2003). An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649.

Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D. & Zardecki, C. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(Pt 6 No 1):899–907.

Berry, M. W.; Dumais, S. T. & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

Borgwardt, K. M.; Ong, C. S.; Schönauer, S.; Vishwanathan, S. V. N.; Smola, A. J. & Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl 1):i47–i56.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brenner, S. E.; Koehl, P. & Levitt, M. (2000). The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Research*, 28(1):254–256.

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

Canavaci, A. M. C.; Bustamante, J. M.; Padilla, A. M.; Perez-Brandan, C. M.; Simpson, L. J.; Xu, D.; Boehlke, C. L. & Tarleton, R. L. (2010). In vitro and in vivo high-throughput assays for the testing of anti-trypanosoma cruzi compounds. *PLoS Neglected Tropical Diseases*, 4(7):e740.

Chakrabarti, D. & Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1).

Chandonia, J. M. & Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science*, 311(5759):347–351.

ChemAxon (2012). PMapper
http://www.chemaxon.com/jchem/doc/user/pmapper.html.

Cheng, J. & Baldi, P. (2006). A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22(12):1456–1463.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4):823–826.

Cleary, J. G. & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. Em *Proceedings of the 12th International Conference on Machine Learning*, pp. 108–114.

Consortium, T. U. (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, 38(Database issue):D142–D148.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

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

Davies, J. R.; Jackson, R. M.; Mardia, K. V. & Taylor, C. C. (2007). The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, 23(22):3001–3008.

Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J. & Hansch, C. (1991). Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797.

Deerwester, S. C.; Dumais, S. T.; Furnas, G. W.; Harshman, R. A.; Landauer, T. K.; Lochbaum, K. E. & Streeter, L. A. (1989). Computer information retrieval using latent semantic structure.

del Castillo-Negrete, D.; Hirshman, S. P.; Spong, D. A. & D'Azevedo, E. F. (2007). Compression of magnetohydrodynamic simulation data using singular value decomposition. *Journal of Computational Physics*, 222(1):265--286.

Delaunay, B. (1934). Sur la sphère vide. *Izv. Akad. Nauk SSSR Math*, 7:793–800.

Ding, C. H. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358.

Dobson, P. D. & Doig, A. J. (2005). Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, 345:187–199.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327–384.

Eldén, L. (2007). *Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms)*. Society for Industrial and Applied Mathematics.

Finn, R. D.; Mistry, J.; Coggill, P.; Heger, A.; Pollington, J.; Gavin, O. L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sohhhammer, E. L. L.; Eddy, S. R. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–222.

Fröhlich, H.; Wegner, J. K.; Sieker, F. & Zell, A. (2005). Optimal assignment kernels for attributed molecular graphs. Em *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 225–232.

Gonçalves-Almeida, V.; Pires, D.; de Melo-Minardi, R.; da Silveira, C.; Meira, W. & Santoro, M. (2012). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349.

Goodsell, D.; Morris, G. & Olson, A. (1996). Automated docking of flexible ligands: applications of autodock. *Journal of Molecular Recognition*, 9(1):1–5.

Goyal, K.; Mohanty, D. & Mande, S. C. (2007). PAR-3D: a server to predict protein active site residues. *Nucleic Acids Research*, 35(Web Server issue):W503–505.

Gross, J. L. & Yellen, J. (2004). *Handbook of Graph Theory*. CRC Press, first edição.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18.

Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J. & Green, D. V. S. (2004). The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *Journal of Chemical Information and Computer Sciences*, 44(6):2145–2156.

Helma, C.; King, R. D.; Kramer, S. & Srinivasan, A. (2001). The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17(1):107–108.

Hoffmann, B.; Zaslavskiy, M.; Vert, J. P. & Stoven, V. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics*, 11:99.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–38.

Hu, J.; Shen, X.; Shao, Y.; Bystroff, C. & Zaki, M. J. (2002). Mining protein contact maps. Em *2nd BIOKDD Workshop on Data Mining in Bioinformatics*.

Huan, J.; Wang, W.; Bandyopadhyay, D.; Snoeyink, J.; Prins, J. & Tropsha, A. (2004). Mining protein family specific residue packing patterns from protein structure graphs. Em *RECOMB*, pp. 308–315. ACM.

Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF - a program to identify and analyze structural motifs in proteins. *Protein Science*, 5(2):212–220.

Jain, P. & Hirst, J. D. (2010). Automatic structure classification of small proteins using random forest. *BMC Bioinformatics*, 11(364):1–14.

Johnson, D. B. (1977). Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13.

Jorgensen, W. (2004). The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818.

Kahraman, A.; Morris, R. J.; Laskowski, R. A. & Thornton, J. M. (2007). Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*, 368(1):283–301.

Kamagata, K. & Kuwajima, K. (2006). Surprisingly high correlation between early and late stages in non-two-state protein folding. *Journal of Molecular Biology*, 357(5):1647–1654.

Kashima, H.; Tsuda, K. & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. Em *Proceedings of the 20th International Conference on Machine Learning*, number 1, pp. 321–328.

Kersting, K.; Raiko, T.; Kramer, S. & De Raedt, L. (2003). Towards discovering structural signatures of protein folds based on logical hidden markov models. Em *Proceedings of the Pacific Symposium on Biocomputing*, pp. 192–203.

Koehl, P. (2001). Protein structure similarities. *Current Opinion in Structural Biology*, 11(3):348–353.

Kolodny, R.; Koehl, P. & Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of Molecular Biology*, 346(4):1173–1188.

Kong, X.; Fan, W. & Yu, P. (2011). Dual active feature and sample selection for graph classification. Em *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 654–662.

Kong, X. & Yu, P. S. (2010). Semi-supervised feature selection for graph classification. Em *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pp. 793–802.

Koshland Jr, D. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98.

Landwehr, N.; Hall, M. & Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.

Laskowski, R. A.; Watson, J. D. & Thornton, J. M. (2005a). Profunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research*, 33(Web Server issue):W89–93.

Laskowski, R. A.; Watson, J. D. & Thornton, J. M. (2005b). Protein function prediction using local 3D templates. *Journal of Molecular Biology*, 351(3):614–626.

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

Lee, B. Y.; Bacon, K. M.; Connor, D. L.; Willig, A. M. & Bailey, R. R. (2010a). The potential economic value of a trypanosoma cruzi (chagas disease) vaccine in latin america. *PLoS Neglected Tropical Diseases*, 4(12):e916.

Lee, B. Y.; Bacon, K. M.; Connor, D. L.; Willig, A. M. & Bailey, R. R. (2010b). The potential economic value of a Trypanosoma cruzi (Chagas disease) vaccine in Latin America. *PLoS Neglected Tropical Diseases*, 4(12):e916.

Lee, D.; Redfen, O. & C, O. (2007). Predicting protein function from sequence and structure. *Nature Reviews: Molecular Cell Biology*, 8(12):995–1005.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. *Machine Learning*, pp. 4–15.

Li, G.; Semerci, M.; Yener, B. & Zaki, M. J. (2011). Graph classification via topological and label attributes. Em *9th Workshop on Mining and Learning with Graphs (with SIGKDD)*.

Mahé, P.; Ueda, N.; Akutsu, T.; Perret, J.-L. & Vert, J.-P. (2004). Extensions of marginalized graph kernels. Em *Proceedings of the 21st International Conference on Machine Learning*, ICML '04, pp. 70–77.

Melo-Minardi, R. C. (2008). *Classificação Estrutural de Famílias de Proteínas com Base em Mapas de Contatos*. Doutorado em Bioinformática, Depto. de Bioquímica e Imunologia do Instituto de Ciências Biológicas – Universidade Federal de Minas Gerais.

Melo-Minardi, R. C.; Lopes, C. E.; Fernandes Jr, F. A.; da Silveira, C. H.; Santoro, M. M.; Carceroni, R. L.; Meira Junior, W. & Araújo, A. A. (2006). A contact map matching approach to protein structure similarity analysis. *Genetics and Molecular Research*, 5(2):284–308.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A*, 209:415–446.

Monod, J.; Changeux, J. & Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6(4):306–329.

Morris, R. J.; Najmanovich, R. J.; Kahraman, A. & Thornton, J. M. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355.

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

Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L. & Pähler, A. (2008). Computational toxicology in drug development. *Drug discovery today*, 13(7-8):303–310.

Najmanovich, R.; Kurbatova, N. & Thornton, J. M. (2008). Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24(16):i105–111.

Nwaka, S. & Hudson, A. (2006). Innovative lead discovery strategies for tropical diseases. *Nature Reviews Drug Discovery*, 5(11):941–955.

O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T. & Hutchison, G. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33.

Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B. & Thornton, J. M. (1997). Cath - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–108.

Pammolli, F.; Magazzini, L. & Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*, 10(6):428–38.

Pires, D. E. V.; da Silveira, C. H.; Santoro, M. M. & Meira Jr, W. (2007). PDBEST: PDB Enhanced Structures Toolkit. Em *Proceedings of the 3rd International Conference of Brazilian Association for Bioinformatics and Computational Biology*.

Pires, D. E. V.; Melo-Minardi, R. C.; ; Santos, M. A.; H, d. C.; Santoro, M. M. & W, M. (2011). Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12 Suppl 4:S12.

Provost, F. & Kohavi, R. (1998). On applied research in machine learning. *Machine Learning*, 30:127–132.

PSI (2011). Protein Structure Initiative. http://www.structuralgenomics.org/.

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 Computational Biology*, 4(10):e1000160.

Rassi Jr., A.; Rassi, A. & Marin-Neto, J. A. (2010). Chagas disease. *Lancet*, 375(9723):1388–1402.

Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, 3(2):71–84.

Rupp, M. & Schneider, G. (2010). Graph kernels for molecular similarity. *Molecular Informatics*, 29(4):266–273.

Schalon, C.; Surgand, J. S.; Kellenberger, E. & Rognan, D. (2008). A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Structure, Function and Bioinformatics*, 71(4):1755–1778.

Shazman, S.; Celniker, G.; Haber, O.; Glaser, F. & Mandel-Gutfreund, Y. (2007). Patch finder plus (pfplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Research*, 35(Web Server issue):W526–30.

Shen, H. B. & Chou, K. C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22(14):1717–1722.

Shulman-Peleg, A.; Shatsky, M.; Nussinov, R. & Wolfson, H. J. (2008). MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Research*, 36(Web Server issue):W260–264.

Sippl, W. (2000). Receptor-based 3D QSAR analysis of estrogen receptor ligands - merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *Journal of Computer-Aided Molecular Design*, 14:559–572.

Soundararajan, V.; Raman, R.; Raguram, S.; Sasisekharan, V. & Sasisekharan, R. (2010). Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, 5(2):e9391.

Spitzer, R.; Cleves, A. E. & Jain, A. N. (2011). Surface-based protein binding pocket similarity. *Proteins: Structure, Function and Bioinformatics*, 79(9):2746–2763.

Stark, A. & Russell, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in non-homologous tertiary structures. *Nucleic Acids Research*, 31(13):3341–3344.

Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L. & Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21 Suppl 1:i359–368.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.

Turcotte, M.; Muggleton, S. H. & Sternberg, M. J. E. (2001). Automated discovery of structural signatures of protein fold and function. *Journal of Molecular Biology*, 306(3):591–605.

Ueno, K.; Mineta, K.; Ito, K. & Endo, T. (2012). Exploring functionally related enzymes using radially distributed properties of active sites around the reacting points of bound ligands. *BMC Structural Biology*, 12(1):5.

Valerio, L. (2009). < i>in silico</i> toxicology for the pharmaceutical sciences. *Toxicology and applied pharmacology*, 241(3):356–370.

Voronoi, G. M. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die reine und angewandte Mathematik*, 134:198–287.

Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J. & Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(Web Server issue):W623–33.

Watson, J. D.; Roman, A. L. & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15(3):275–284.

Watson, J. D.; Sanderson, S.; Ezersky, A.; Savchenko, A.; Edwards, A.; Orengo, C.; Joachimiak, A.; Laskowski, R. A. & Thornton, J. M. (2007). Towards fully automated structure-based function prediction in structural genomics: a case study. *Journal of Molecular Biology*, 367(5):1511–1522.

Weskamp, N.; Hullermeier, E.; Kuhn, D. & Klebe, G. (2007). Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2):310–320.

Willett, P. (2006). Similarity-based virtual screening using 2d fingerprints. *Drug Discovery Today*, 11(23–24):1046–1053.

Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(Database issue):D668–672.

Yan, X.; Cheng, H.; Han, J. & Yu, P. S. (2008). Mining significant graph patterns by leap search. Em *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pp. 433–444. ACM.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374.

Zhang, C.; Vasmatzis, G.; Cornette, J. L. & DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology*, 267(3):707–726.

Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710.

Apêndice A

Artigo 1: BMC Genomics

- Título: *Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns*
- Autores: **Pires, DEV**; Melo-Minardi, RC; Santos, MA; da Silveira, CH; Santoro, MM; Meira Junior, W
- Periódico: BMC Genomics (*Impact Factor*: 4.07)
- Volume: 12 Suppl
- Ano: 2011
- URL: http://www.biomedcentral.com/1471-2164/12/S4/S12/

Pires et al. BMC Genomics 2011, 12(Suppl 4):S12
http://www.biomedcentral.com/1471-2164/12/S4/S12

BMC Genomics

PROCEEDINGS Open Access

Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns

Douglas EV Pires[1,2*], Raquel C de Melo-Minardi[2], Marcos A dos Santos[2], Carlos H da Silveira[3], Marcelo M Santoro[1], Wagner Meira Jr.[2]

From 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010)
Ouro Preto, Brazil. 15-18 November 2010

Abstract

Background: The unforgiving pace of growth of available biological data has increased the demand for efficient and scalable paradigms, models and methodologies for automatic annotation. In this paper, we present a novel structure-based protein function prediction and structural classification method: Cutoff Scanning Matrix (CSM). CSM generates feature vectors that represent distance patterns between protein residues. These feature vectors are then used as evidence for classification. Singular value decomposition is used as a preprocessing step to reduce dimensionality and noise. The aspect of protein function considered in the present work is enzyme activity. A series of experiments was performed on datasets based on Enzyme Commission (EC) numbers and mechanistically different enzyme superfamilies as well as other datasets derived from SCOP release 1.75.

Results: CSM was able to achieve a precision of up to 99% after SVD preprocessing for a database derived from manually curated protein superfamilies and up to 95% for a dataset of the 950 most-populated EC numbers. Moreover, we conducted experiments to verify our ability to assign SCOP class, superfamily, family and fold to protein domains. An experiment using the whole set of domains found in last SCOP version yielded high levels of precision and recall (up to 95%). Finally, we compared our structural classification results with those in the literature to place this work into context. Our method was capable of significantly improving the recall of a previous study while preserving a compatible precision level.

Conclusions: We showed that the patterns derived from CSMs could effectively be used to predict protein function and thus help with automatic function annotation. We also demonstrated that our method is effective in structural classification tasks. These facts reinforce the idea that the pattern of inter-residue distances is an important component of family structural signatures. Furthermore, singular value decomposition provided a consistent increase in precision and recall, which makes it an important preprocessing step when dealing with noisy data.

Background

With the increasing number of genome and metagenome projects, sequence databases have grown exponentially. On the one hand, the August 2010 release of the UniprotKB/TrEMBL database [1] contains about 12,000,000 protein sequences. In the last month, more than 300,000 new sequences have been added to that repository, and about 6,000,000 entry annotations have been revised. On the other hand, the Pfam database of protein families [2] represents about 12,000 families, and about 20% of these are domains of unknown function (DUFs), revealing that state-of-the-art sequence similarity-based and even profile-based annotation

* Correspondence: dpires@dcc.ufmg.br
[1]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil
Full list of author information is available at the end of the article

methods have had limited success in assigning functions to novel proteins.

Protein structural classification databases, such as SCOP [3], also present difficulties in keeping up with the increasing number of protein structures solved and deposited in public repositories. Approximately 53% of the Protein Data Bank (PDB) [4] entries are classified by the current release of SCOP (1.75) as of April 2011, and after removing redundancy (sequence similarity at 90%), the coverage drops to about 41%. As international structural genomics initiatives have produced a huge number of structures of unknown function, attempting to automatically assign functions to these proteins is becoming even more necessary, and significant efforts have been devoted to this task [5-8].

In this context, novel paradigms, models and methodologies for automatic annotation must be investigated. Because protein structure and function are more conserved than protein sequence [9], the identification of similarities between novel sequences and known structures would greatly improve the characterization of these sequences. Fold recognition refers to identifying main structural features by the connections and positions of secondary structure elements. Conversely, according to Murzin et al. [3], structural classification is conducted at hierarchical levels (class, fold, superfamily and family) that embody evolutionary and structural relationships. In this work, we focused on structural classification, which encompasses the problem of fold recognition. Both fold recognition and structural classification are important steps toward function prediction.

Over the years, protein fold recognition has been addressed through different approaches. The authors of [10] extracted a series of features from protein sequences and used support vector machines and neural network learning methods as the base classifiers in a dataset composed of SCOP folds. Later, ensemble classifiers [11] were applied to these same feature vectors, improving the success rate. The use of a combination of sequence and structure information brought an improvement to fold recognition, as mentioned in the information retrieval approach introduced in [12].

Likewise, several efforts toward structure-based protein function prediction have been made. We can quote, for instance, the search for structural motifs [13-15] and functional residues (such as DNA [16] and metal [17] binding sites), the use of 3D templates [5] and the comparison of protein folds by structure alignments [18,19]. There have also been attempts to infer function from structure without the use of alignment algorithms, such as in enzyme classification [20,21]. Similarly, in the present work, we do not use alignment techniques or any sequence information in our method, relying only on structural grounds. A primary problem faced when dealing with protein function, as pointed out in [22], s defining the scope and function. Protein function prediction may be understood from different perspectives. It could mean the prediction of the cellular process in which a protein is involved, its enzymatic activity or even its physiological role. For instance, a protein's enzymatic activity could be described by EC numbers, while its physiological role might be related to its subcellular localization. In this work, the aspect of protein function considered is enzyme activity. However, the study might be extended, without loss of generality, to other functional features, like the terms of the Gene Ontology (GO) [23] annotation.

Even though function cannot be directly implied from the specific fold adopted by a certain protein, structural data can be used to detect proteins with similar functions whose sequences have diverged during evolution [24]. In this context, one possible strategy is the definition of structural signatures, which are sets of features that are able to unequivocally identify a protein fold and the nature of interactions it can establish with other proteins and ligands. These feature sets are concise representations of protein structures, and we believe that their discovery and comprehension will be an important milestone in the protein function prediction field, being a step beyond sequence homology-based methods.

In this paper, we investigate a special type of feature that might be part of structural signatures: the patterns in inter-residue distances (or contacts). Proteins with different folds and functions present significant differences in the distribution of distances among residues as a consequence of the underlying interaction and packing of the atomic network, which is fundamental for defining protein folding [25]. In [26], we have used these distribution distances to compare and correlate different methodologies of protein inter-residue contacts. We found, surprisingly, that the traditional cutoff-dependent approach was a simpler, more complete and more reliable technique for contact definition than other cutoff-independent methods, such as Delaunay tessellation [27], especially when the target is the discrimination of first-order contacts. In this work, we propose using inter-residue distance patterns for protein classification.

The structural data we used are the cumulative contact distributions based on the Euclidean distances among alpha carbons, the Cutoff Scanning Matrix (CSM). The motivation for the use of this kind of information lies in the fact that proteins with different folds and functions have significantly different distributions of distances between their residues, and protein similarity is reflected in these distance distributions, information that is captured in the CSM. After generating this structural data, we apply singular value decomposition (SVD)

to reduce dimensionality and noise. The processed matrix is finally submitted to different, previously described classification algorithms. Therefore, the main innovation of this work relies more on the powerful combination of the new structural feature of inter-residue contacts used as a discriminator and principal components selection by SVD rather than in the creation of a new classification method per se. Indeed, we showed our methodology to be, in general, independent of the classifiers utilized, giving even results for different classification heuristics.

Having in mind these considerations, we showed that the patterns derived from CSMs might effectively be used in automatic protein function prediction and structural classification. At first glance, in the case of enzyme function prediction, the proposed method achieved (over the superfamilies) an average precision of 98.2% (sd = 1.6) and average recall of 97.9% (sd = 2.0), using a gold-standard dataset of enzymes [28]. Using a much larger set of enzymes with their respective EC numbers (the 950 most-populated EC numbers in terms of available structures), CSM was able to achieve up to 95.1% precision and recall results. For the recall results, considering the levels of hierarchical structure of SCOP [3], we were able to accomplish an average precision of 93.5% (sd = 1.4) and average recall of 93.6% (sd = 1.4). In comparison to the state-of-the-art methods used in this context, such as that given by Jain and Hirst [29], using very similar database input (SCOP release 1.75), our methodology presented more robust and homogeneous results, with an average precision a bit below that of those authors: 90.7% versus 93.6%, but with less dispersion (sd of 3.0 versus 6.4). We had remarkably better recall results: an average of 90.7% versus 77.0%, with significantly lower dispersion (sd of 2.9 versus 18.4). Further details are discussed in the next section.

Results and discussion

To test the ability of our method to successfully predict functions and recognize folds, we performed two sets of experiments with datasets designed for these different tasks.

For function prediction, as mentioned in the Methods section, we built one database based on manually curated protein superfamilies and another based on EC numbers to test if the present structure-based method could help in protein function annotation.

For structural classification, we performed experiments to verify our ability to assign SCOP class, superfamily, family and fold to protein domains. Furthermore, to place this work into the context of the literature, we also tested a superset of the dataset used by Jain and Hirst in [29]. As far as we know, their work presents the highest precision in protein fold recognition published thus far.

Finally, we relate some experiments that aimed to evaluate an SVD-based noise reduction strategy.

Function prediction

In the function prediction experiments, our goal was to assess how well three different classification algorithms predict protein function according to protein EC numbers and a mechanistically diverse gold-standard database of functional family classes [28]. We used 10-fold cross validation for all the experiments.

For the dataset of the top 950 most-populated EC numbers, CSM was able to achieve 95.1% precision and recall after SVD processing using the KNN (K-Nearest Neighbors) algorithm. The four levels of the EC number were used together as the classes to train and test the classifier. Additional file 1, Figure S1 shows the variation in the performance metrics for each EC number class considered. Even though the number of proteins assigned to each EC number is very unbalanced, the majority of classes were classified properly, with high quality according to the metrics extracted.

Considering the gold-standard dataset, without SVD and using KNN, our method achieved an average precision of 94.2% (sd = 5.5) and a recall of 94.5% (sd = 5.5) (Table 1). For naive Bayes, it achieved 82.3% (sd = 13.8) precision and 79.2% (sd = 15.4) recall (Additional file 1, Table S1), and for random forest, it achieved 92.0% (sd = 6.9) precision and 91.6% (sd = 7.2) recall (Additional file 1, Table S2). We also showed that by using SVD, we may significantly improve these results, and in the worst case, we had 94.6% precision and 93.1% recall for the enolase superfamily using naive Bayes. The KNN and random forest methods were able to detect isoprenoid synthase type I with 100% precision and recall. Additionally, we performed experiments using all six superfamilies to train a single classifier. In this scenario, even

Table 1 Function prediction performance using KNN for the gold-standard dataset

Superfamily	Before SVD		After SVD		ΔPrec.	ΔRec.
	Precision	*Recall*	*Precision*	*Recall*		
Amidohydrolase	0.983	0.983	1.000	1.000	+1.7%	+1.7%
Crotonase	0.955	0.953	0.979	0.977	+2.4%	+2.4%
Enolase	0.876	0.853	0.971	0.967	+9.5%	+11.4%
Haloacid Dehalogenase	0.881	0.925	0.984	0.981	+10.3%	+5.6%
Isoprenoid Synthase Type I	1.000	1.000	1.000	1.000	+0.0%	+0.0%
Vicinal Oxygen Chelate	1.000	1.000	1.000	1.000	+0.0%	+0.0%
All	0.901	0.903	0.991	0.989	+9.0%	+8.6%

Prediction performance for the gold-standard dataset using KNN. The experiment was performed in an intra-superfamily fashion, and the *classes* for prediction represent the enzyme's families. The precision and recall metrics are weighted averages. Ten-fold cross validation was employed.

with a greater number of families in the training and testing phases, we were still able to achieve up to 99.0% precision with KNN and random forest after SVD preprocessing.

Protein structural classification

To the best of our knowledge, no test of the structural classification of very large databases, such as the entire SCOP containing about 110,000 domains, has been published. Due to SVD dimensionality reduction ability and the possibility of representing protein instances by a few significant attributes, we present a method that can efficiently handle such volume of data.

We may recognize protein folds at a 92.2% precision and 92.3% recall using KNN (Table 2). Even broad proteins categories, such as the SCOP class level, can be separated using CSM with very significant precision and recall (95.4% for both). The proposed method was able to classify proteins in the four levels of SCOP hierarchy with very high precision and recall, showing that CSM is a suitable method for fold recognition and also that CSMs are a very promising component of protein structural signatures. Additionally, we verified the impact of imposing a minimum number of entities per node of the SCOP hierarchy on the precision of the prediction. Additional file 1, Figure S2 shows an approximately linear correlation between these variables for the fold, superfamily and family levels with and without the SVD processing. This correlation was not analyzed for the class level because all of the classes have more than 100 entities.

Performance comparison

In [29], the authors presented a random forest-based method to predict the SCOP class, fold, superfamily and family levels based on secondary structure element descriptors that achieved precisions of up to 99.0%. Using a similar dataset, we tried to compare our results to theirs. As far as we are concerned, this was the state-of-art method for automatic structural classification. They used a subset of SCOP database as they aimed to recognize protein folds. In our comparison of results, we were able to achieve similar precision levels but with higher recall (overcoming in up to 50.0%) in most of the cases. In only 3 of the 16 experiments, we obtained a lower recall value with our method and our F1 scores were also superior. The complete set of information regarding this experiment is available in Table 3. Figure 1 shows the performance comparison for each experiment in terms of precision and recall. CSM significantly overcomes the recall of the aforementioned study while preserving a compatible precision level. We stress that our method is not limited to small proteins. These results show that our method is not only comparable to [29] but also presents a considerable gain in terms of recall.

Table 2 Structural classification performance using KNN for the Full-SCOP dataset

SCOP Level	Before SVD		After SVD		ΔPrec.	ΔRec.
	Precision	*Recall*	*Precision*	*Recall*		
Class	0.927	0.926	0.954	0.954	+2.7%	+2.8%
Fold	0.868	0.869	0.922	0.923	+5.4%	+5.4%
Superfamily	0.871	0.872	0.926	0.927	+5.5%	+5.5%
Family	0.888	0.889	0.938	0.938	+5.0%	+4.9%

Prediction performance for the full-SCOP dataset using KNN. The experiment was performed for each classification level of SCOP. The precision and recall metrics are weighted averages. A 10-fold cross validation was employed.

Noise reduction strategy

As we mentioned, SVD-based noise reduction was able to improve the precision and recall levels. We obtained a gain of up to 10.3% with the KNN classifier, 35.0% with naive Bayes and 16.2% with random forest. Interestingly, we verified that the different classifiers achieved comparable results after the use of SVD for dimensionality reduction (all levels remained above 90%). Dimension reduction ability is important for scalability in this scenario because many protein domains are experiencing exponential growth. There are about 110,000 domains, i.e., instances to classify, in the SCOP database. Each of these instances can be represented by 151 attributes (dimensions) in the case of the CSM with a cut-off of up to 30Å.

To find the point that maximizes the noise reduction, we studied the singular value distribution obtained for the gold-standard dataset. Figure 2 shows the elbow of the curve of the contribution of each singular value to represent the original information. Using about 9 dimensions we can represent the same information (reducing the noise) and obtain very high precision in classification with a considerably smaller dataset. As shown in Figure 3, maximum precision can be achieved with about 9 singular values for all experiments.

Conclusions

Function and fold prediction, while means of understanding the composition, operation, interaction and evolution of proteins, are still great challenges in the face of the explosive growth of protein data generation and storage in public databases. To keep up with the frenetic pace imposed by this increasing data availability, novel, efficient methods for automatic and semi-supervised annotation are needed. As a mechanism to exploit the close relationship between protein structure and function, we developed a structure-based method for function prediction and fold recognition based on protein inter-residue distance patterns. The motivation for

Table 3 Comparison of prediction performance

Dataset	SCOP level	CSM+SVD			Jain et al.			ΔPrec.	ΔRec.
		Prec .	*Recall*	*F1*	*Prec .*	*Recall*	*F1*		
3SSE	Class	0.991	0.991	0.991	0.890	0.840	0.864	+10.1%	+15.1%
	Fold	0.956	0.957	0.956	0.860	0.450	0.591	+9.6%	+50.7%
	Superfamily	0.956	0.957	0.956	0.800	0.550	0.652	+15.6%	+40.7%
	Family	0.935	0.935	0.935	0.820	0.870	0.844	+11.5%	+6.5%
4SSE	Class	0.961	0.962	0.961	0.990	0.990	0.990	-2.9%	-2.8%
	Fold	0.939	0.939	0.938	0.960	0.830	0.890	-2.1%	+10.9%
	Superfamily	0.938	0.937	0.937	0.880	0.690	0.774	+5.8%	+24.7%
	Family	0.935	0.934	0.933	0.980	0.920	0.949	-4.5%	+1.4%
5SSE	Class	0.985	0.985	0.985	0.980	1.000	0.990	+0.5%	-1.5%
	Fold	0.969	0.969	0.969	1.000	0.690	0.817	-3.1%	+27.9%
	Superfamily	0.970	0.969	0.969	0.980	0.650	0.782	-1.0%	+31.9%
	Family	0.967	0.965	0.965	0.980	0.920	0.949	-1.3%	+4.5%
6SSE	Class	0.966	0.965	0.965	0.970	1.000	0.985	-0.4%	-3.5%
	Fold	0.943	0.943	0.942	0.950	0.510	0.664	-0.7%	+43.3%
	Superfamily	0.937	0.939	0.937	0.950	0.570	0.713	-1.3%	+36.9%
	Family	0.932	0.932	0.930	0.980	0.840	0.905	-4.8%	+9.2%

A comparison of prediction performance between the current study and the method introduced by [29]. The precision and recall metrics are weighted averages. This result comprises a 10-fold cross validation in KNN.

this approach arose from the hypothesis that proteins with different structures would show different inter-residue distance patterns, and structural similarity would be reflected in these distances.

One of the most remarkable advantages of the CSM-based structural signature is its generality, as we successfully instantiated it in different problem domains, such as function and fold prediction. Also, as a requirement and demand for its application to databases that are continuously growing, it is scalable for real-world scenarios, such as whole-SCOP classification tasks, as shown in previous sections, and it shows an efficacy comparable or superior to state-of-the-art protein folding and function predictors. We would like to stress that our method is probably the first to present a full-SCOP automatic classification in acceptable time (a few hours in a quad-core machine).

The interpretation and understanding of the intrinsic distance patterns generated by CSM demand further investigation. As part of future studies, we intend to explore the generality of CSMs in other aspects of protein function, such as subcellular localization prediction and prediction of GO terms, as well as under different structural classification databases, such as CATH [30]. We also plan to contrast SVD with feature selection as methods for discriminant information discovery in CSMs.

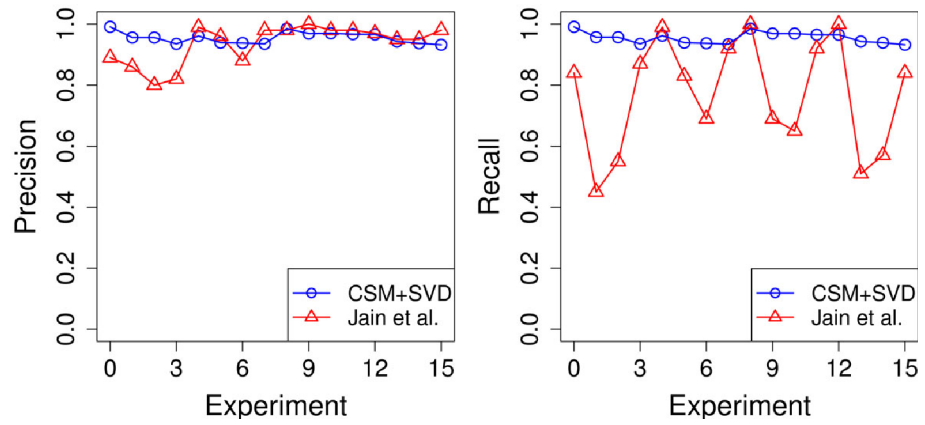

Figure 1 Comparison of precision and recall. A comparison of the prediction performance of the CSM+SVD approach and the work of Jain and colleagues in terms of precision and recall. CSM, while achieving a compatible level of precision, presents a significant improvement in recall.

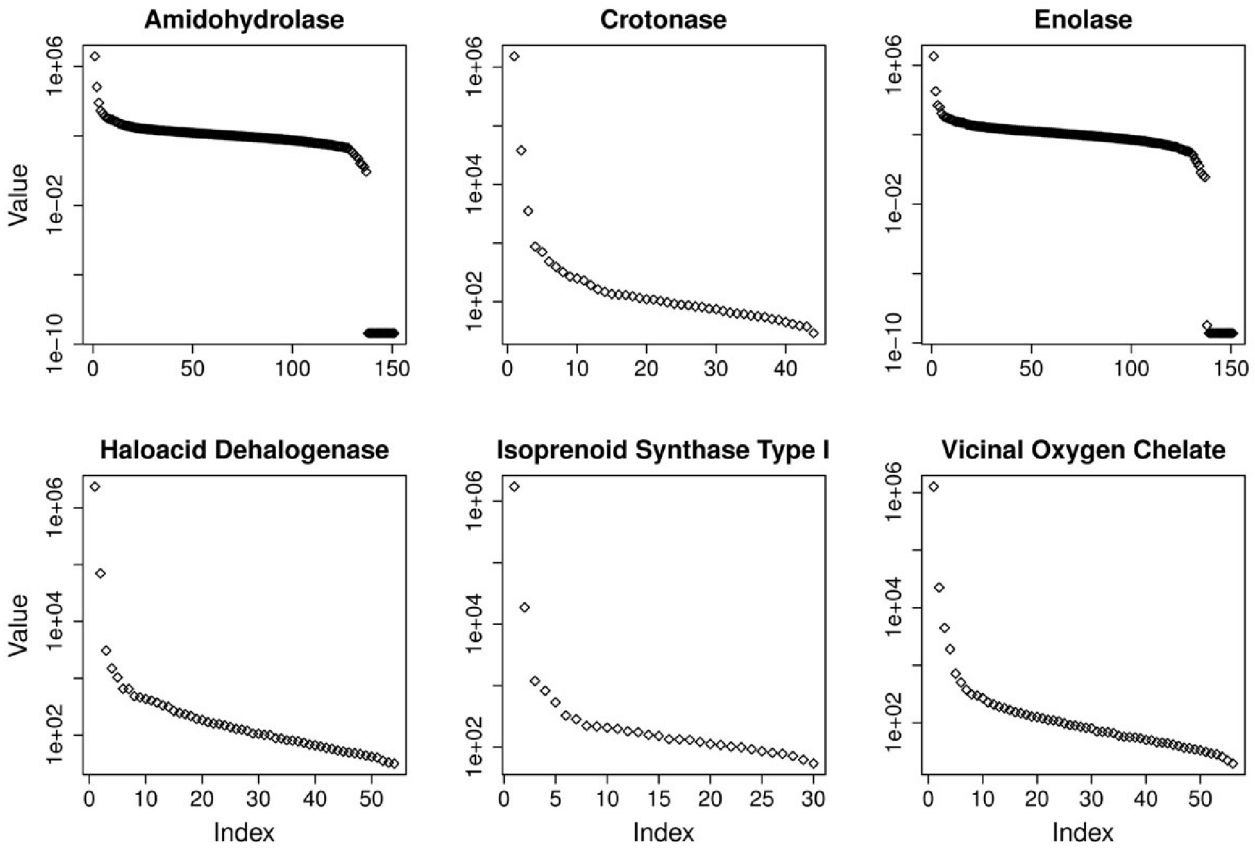

Figure 2 Singular value distribution. Singular value distribution obtained after the execution of the SVD routine for each superfamily considered in the gold-standard dataset. A sudden drop in the singular values denotes the cutoff point for dimensionality reduction. The Y-axes have a logarithmic scale.

Furthermore, the significant gain in prediction power provided by SVD processing might imply that there is room to improve in terms of the data input, indicating that other cutoff ranges and granularities should also be tested, which is a study already in progress in our group.

Methods

CSM-based approach

Figure 4 gives a schematic view of the CSM-based approach for protein function prediction and fold recognition employed in this work, which can be divided into data preprocessing, CSM generation, SVD-based dimensionality reduction and classification steps.

After the data acquisition and filtering steps for a certain dataset (designed either for function prediction or fold recognition purposes), the CSMs are generated (the details of the procedure are explained later in this section). The CSM defines a feature vector that is then processed with SVD. To define a threshold value for dimensionality reduction, the singular values distribution is analyzed. The elbow of this distribution is used as a threshold for data approximation and recomposition (the explanation of the SVD procedure is detailed in the next subsections) and indicates that the contribution of the other singular values to describing the matrix is insignificant, and thus they might be seen as noise.

These singular values are then discarded. Finally, the processed CSM is submitted for classification tasks under different algorithms. Metrics such as precision and recall are calculated to assess the prediction power of the classifiers.

Cutoff scanning matrices

In a previous work [26], we conductedd a comparative analysis between two classical methodologies to prospect residue contacts in proteins, one based on geometric aspects, and the other based on a distance threshold or cutoff, by varying (scanning) this distance to find a robust and reliable way to define these contacts. In the present work, we used the cutoff scanning approach for classification purposes, which is the basis of the CSMs. The motivation for the use of this kind of information relies on the fact that proteins with different folds and functions present significant differences in the

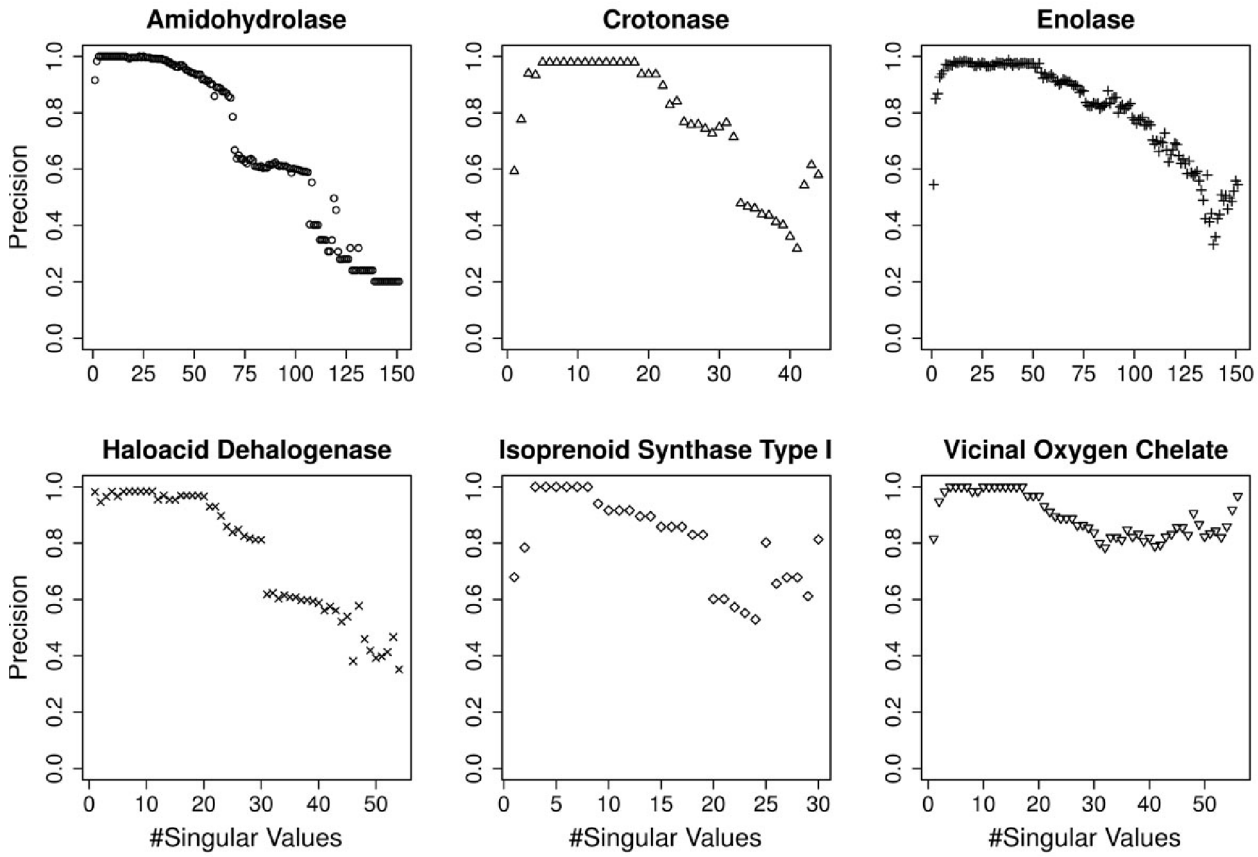

Figure 3 Influence of the number of singular values chosen in precision. Influence of the cutoff point for dimensionality reduction in the average weighted precision for the superfamilies considered in the gold-standard dataset. A drop in the precision after a certain number of singular values might indicate the point where noisy components start to appear.

distribution of distances between their residues. On the other hand, one can expect that proteins with similar structures would also have similar distance distributions between their residues, information that is captured in a CSM.

The CSMs were generated as follows: for each protein of the datasets, we generated a feature vector. First, we calculated the Euclidean distance between all pairs of C_α and defined a range of distances (cutoffs) to be considered and a distance step. We scanned through these distances, computing the frequency of pairs of residues, each represented by its C_α, that are close according to this distance threshold. Algorithm 1 shows the function that calculates the CSM.

Algorithm 1 Cutoff Scanning Matrix calculation

function GENERATECSM(*ProteinSet*, *CSM*, $Distance_{MIN}$, $Distance_{MAX}$, $Distance_{STEP}$)
  for all protein $i \in$ (*ProteinSet*) **do**
    $j = 0$
    Calculate the distances between all pairs of C_α
    for $dist \leftarrow Distance_{MIN}$; **to** $Distance_{MAX}$; **step** $Distance_{STEP}$ **do**
      $CSM[i][j] \leftarrow$ Get frequency of pairs of C_α within a distance $dist$
      $j++$
  return *CSM*

In this work, we vary the distance threshold from 0.0 Å to 30.0 Å, with a 0.2-Å step, which generates a vector

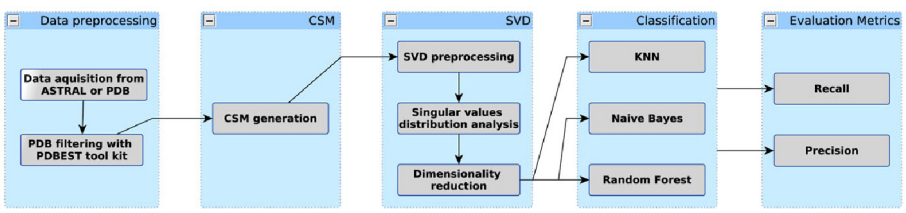

Figure 4 CSM-based function prediction and fold recognition workflow. A schematic view of the CSM-based function prediction and fold recognition approach is shown. The workflow is divided into steps of data preprocessing, CSM generation, SVD dimensionality reduction, classification and evaluation.

of 151 entries for each protein. Together, these vectors compose the CSM. In short, each line of the matrix represents one protein, and each column represents the frequency of residue pairs within a certain distance. Alternatively, this frequency might be seen as the number of contacts in the protein for a certain cutoff distance or the edge count of the contact graph defined using that distance threshold. This step was implemented in the Perl programming language.

It is important to mentioned that other centroids could be chosen instead of the C_α, such as the C_β or the last heavy atom (LHA) of the side chain. Additional file 1, Figure S3 shows the performance comparison between the C_α and C_β for the EC number dataset. The C_α performed better in all experiments, a fact that demands further investigation.

The motivation for using CSMs comes from the differences in the contact distributions for proteins of different structural classes, as can be seen in Additional file 1, Figure S4, which shows the normalized edge count density distribution per cutoff for proteins from different SCOP classes, namely: *all alpha*, *all beta*, *alpha+beta* and *alpha/beta*. It is possible to see that the differences between the distributions emerged at different cutoff ranges. For example, the first peaks for the alpha proteins indicate first-order contacts of their helices and the differences at higher cutoffs might happen due to the diameter and density of the proteins. We stress that these variations in the edge count are not only a phenomenon of the secondary structure composition of the proteins but a phenomenon of the protein packing itself. It is important to explain the cutoff variation. The cutoff variation (scanning) aggregates important information related to the packing of the protein and captures, implicitly, the protein shape. We believe that pockets on the surface and even core cavities are well accounted for by this novel type of structure data we proposed. Another example of contact distributions is shown in Figure 5. Three proteins with very different shapes were selected (a globin, PDB:1A6M; a porin, PDB:2ZFG; and a collagen, PDB:1BKV), and the topology of the contact graph obtained with different cutoffs is shown (6.0 Å, 9.0 Å and 12.0 Å). The cumulative and normalized density distributions for the CSM feature vectors for these representatives are also plotted. We can see from these examples that an expressive difference in shape is accounted for in the CSM. In the contact profile, the peaks indicate high frequency of recurrent distance patterns present in proteins structures. A higher peak under 3.8-4.0 Å provides evidence for the distances given by consecutive C_αs. These distances will tend to be independent of the protein structural class in face of the planar property that characterize the peptide link intermediating two contiguous C_αs in the chain. In addition to this pattern, in proteins rich in helices, we will find new suggestive peaks between 5.0 Å and 7.0 Å, representing mainly the recurrent distances between the local (in sequence) C_αs positions (i, i + 2), (i, i + 3) and (i, i + 4) that compose turns of a helix, and also some nonlocal contacts. Conversely, in proteins rich in beta strands, important peaks will be noted around 6.0 Å and 5.0 Å, referring not only the distances in local C_α positions (i, i + 2) but also nonlocal C_α contacts (i, i + k) present in companion strands. This implies that CSM is manipulating two essential structural information levels: local and nonlocal relevant contacts. We also can see that the shapes of the proteins directly interfere in the underlying contact network, which is reflected in the protein folding, as pointed by [25]. These properties make the CSM a rich and important source of information when dealing with problems like protein function prediction and structural classification.

Noise reduction with SVD

To reduce the inherent noise in the generated data and also reduce the cost of the classification algorithms in terms of execution time and memory requirements, we used an SVD-based dimensionality reduction. SVD establishes non-obvious, relevant relationships among clustered elements [31-33]. The rationale behind SVD is that a matrix A, composed of m rows by n columns, can be represented by a set of derived matrices [33] that allows for a numerically different representation of data without loss in semantic meaning. That is:

$$A = TSD^T$$

Where T is an orthonormal matrix of dimensions m x m, S is a diagonal matrix of dimensions m x n and D is an orthonormal matrix with dimensions n x n. The diagonal values of S are the singular values of A, and they are ordered from the most to the least significant values.

When considering only a subset of singular values of size $k < p$, where p is the rank of A, we can achieve A_k, an approximate matrix of the original matrix A:

$$A \approx A_k = T_k S_k D_k^T$$

Thus, data approximation depends on how many singular values are used [34]. In this case, the k number of singular values is also the rank of the matrix A_k. The possibility of extraction of information with less data is part of this technique's success, as it can permit data compression/decompression within a non-exponential execution time, making analysis viable [34]. A dataset represented by a smaller number of singular values than the full-size original dataset has a tendency to group together certain data items that would not be grouped if we used the original dataset [33]. This grouping could

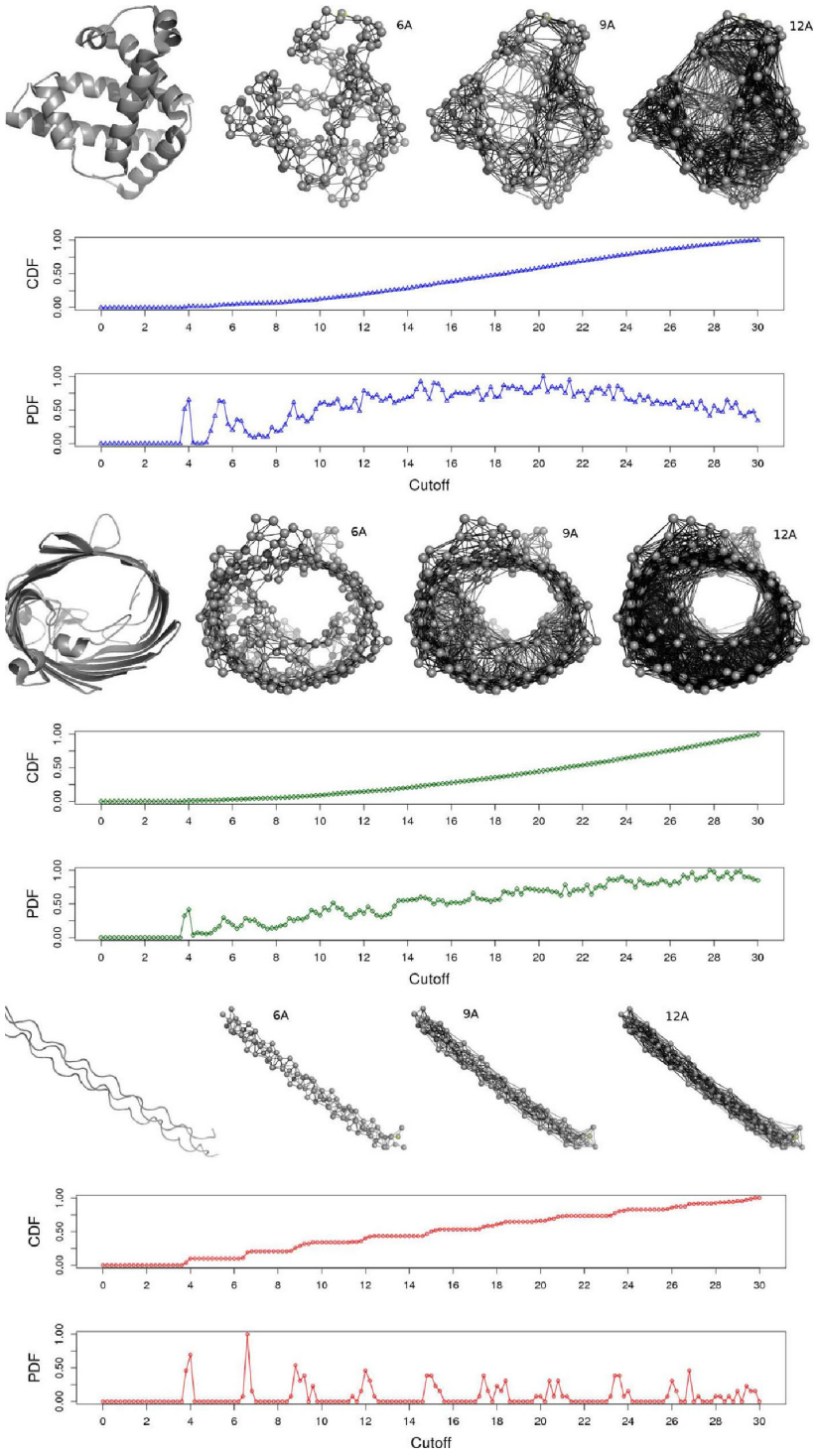

Figure 5 Contact graphs topology per cutoff for proteins with different folds. The topologies of the contact graphs of three distinct structures (from top to bottom: globin, porin and collagen) at different cutoff values: 6.0Å, 9.0Å and 12.0Å are shown. The edge count for each graph represents an entry in the cutoff scanning feature vector. The normalized cumulative distribution and density distribution of the cutoff scanning profile of these proteins are also shown.

explain why clusters derived from SVD can expose non-trivial relationships between the original dataset items [35]. In this paper, we use A_k, the product's factorization by SVD, to rank k, but with only two arrays of SVD, the matrix V_k[32] can be represented in the context of the matrix:

$$A_k = T_k S_k D_k^T = T_k(S_k D_k^T) = T_k V_k$$

The justification for using only V_k is that the relationships among the columns of A_k are preserved in V_k because T_k is a base for the columns of A_k.

We evaluated the singular values distribution in an effort to find a good threshold to reduce the number of dimensions without losing information. This step, as well as the generation of all graphics, was performed via R programming language scripts.

Evaluation methodology

An extensive series of experiments was designed to evaluate the efficacy of CSMs as a source of information for protein fold recognition and function prediction.

In the classification tasks, the Weka Toolkit [36], developer version 3.7.2 was used. For the gold-standard dataset, three classification algorithms were used, and their performances were compared: KNN, random forest and naive Bayes. For the other datasets, KNN was used. The algorithms' parameters, when applicable, were varied and the best result computed. In all scenarios, 10-fold cross validation was applied. The classification performance was evaluated using metrics such as *precision* ($Precision = TP/(TP + FP)$), *recall* ($Recall = TP/(TP + FN)$), *F1 score* (the harmonic mean between precision and recall: $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$) and the Area Under the ROC Curve (AUC). The variation in precision was used to measure the gain obtained with SVD processing, and the recall variation was evaluated to compare the results with those for the dataset derived from [29].

We also correlated the precision obtained by the classifiers and the number of singular values considered and compared it with the results using the whole CSM.

Datasets

Our datasets consisted of proteins structures available in the Protein Data Bank [4]. The domains covered by SCOP release 1.75 were obtained through the ASTRAL compendium [37]. The protein structures were grouped according to the purpose of the experiment, namely, function prediction or fold recognition. For structures solved by NMR, we only considered the first model. The chains were split into separate files and the C_α co-ordinates extracted using PDBEST toolkit.

The first dataset concerns a gold-standard of mechanistically diverse enzyme superfamilies [28]. We consider *six superfamilies* (amidohydrolase, crotonase, haloacid dehalogenase, isoprenoid synthase type I and vicinal oxygen chelate), comprising 47 families distributed among 566 different *chains*. The list of PDB IDs as well as the family and superfamily assignments are available in Additional file 2.

The second dataset contains enzymes with EC numbers. We considered the top 950 most-populated EC numbers in terms of available structures, with at least 9 representatives per class, in a total of 55,474 chains, which covered 95% of the reviewed enzymes from Uniprot [1], i.e., the experimentally validated annotations from that database.

The third dataset originated from SCOP version 1.75 for fold recognition tasks. We selected all PDB IDs covered by SCOP with at least 10 residues and 10 representatives per node in the SCOP classification hierarchy. These IDs represented a total of 110,799, 108,332, 106,657 and 102,100 domains at the class, fold, superfamily and family levels, respectively. We would like to emphasize that this is a very large dataset and that we found no other paper relating the use of such a complete dataset in strutcural classification tasks. The last dataset was derived from [29] for comparison in fold recognition tasks. We selected all domains described in its additional files with a minimum of 10 representatives per node in the SCOP classification hierarchy. It was not possible to identify exactly the domains they used from the additional files and only those pairs of domains with a sequence identify below 35% were retained. It is important to stress that the work of Jain and colleagues only contemplate structures with 3, 4, 5 or 6 secondary structure elements.

Additional material

Additional file 1: Additional figures and tables. Figure S1 - Performance metrics across EC classes. Figure S2 - Correlation between precision and minimum number of representatives. Figure S3 - The influence of C_α and C_β distances in the performance. Figure S4 - Feature vector density distribution for proteins of different SCOP classes. Table S1 - Function prediction performance using naive Bayes for gold-standard dataset. Table S2 - Function prediction performance using random forest for the gold-standard dataset.

Additional 2: Enzyme gold-standard dataset. List of PDB identifiers that compose the enzyme gold-standard dataset and its family and superfamily assignments.

List of abbreviations used

EC: Enzyme Commission; CSM: Cutoff Scanning Matrix; DUF: Domain of Unknown Function; SVD: Singular Value Decomposition; PDB: Protein Data Bank; SCOP: Structural Classification of Proteins; GO: Gene Ontology; LHA: Last Heavy Atom; KNN: K-Nearest Neighbors; AUC: Area Under the ROC Curve.

Acknowledgements

This work was supported by the Brazilian agencies: CAPES, CNPq, FAPEMIG and FINEP. The EC number dataset was kindly provided by Elisa Lima.

This article has been published as part of *BMC Genomics* Volume 12 Supplement 4, 2011: Proceedings of the 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S4

Author details
[1]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil. [2]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil. [3]Advanced Campus at Itabira, Universidade Federal de Itajubá, Itabira, 37500-903, Brazil.

Authors' contributions
DEVP conceived of the study, developed the algorithms, performed the experiments and drafted the manuscript. RCMM participated in the design of the study, helped with presenting and analyzing the results and drafted the manuscript. MAS participated in the design of the study, provided advice on the SVD analysis and helped draft the manuscript. CHS helped with presenting the results, provided advice on its analysis and helped draft the manuscript. MMS helped draft the manuscript and provided advice on analyzing the results. WM participated in the coordination of the study and helped draft the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 22 December 2011

References

1. Consortium TU: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Research* 2010, **38(Database issue)**:D142-D148.
2. Finn RD, Mistry J, Coggill P, Heger A, Pollington J, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sohhhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Research* 2010, **38(Database issue)**:D211-D222.
3. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *Journal of Molecular Biology* 1995, **247(4)**:536-40.
4. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **58(Pt 6 No 1)**:899-907.
5. Laskowski RA, Watson JD, Thornton JM: **Protein function prediction using local 3D templates.** *Journal of Molecular Biology* 2005, **351(3)**:614-626.
6. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Research* 2005, **33(Web Server issue)**:W89-93.
7. Watson JD, Roman AL, Thornton JM: **Predicting protein function from sequence and structural data.** *Current Opinion in Structural Biology* 2005, **15(3)**:275-284.
8. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: **Towards fully automated structure-based function prediction in structural genomics: a case study.** *Journal of Molecular Biology* 2007, **367(5)**:1511-1522.
9. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5(4)**:823-6.
10. Ding CH, Dubchak I: **Multi-class protein fold recognition using support vector machines and neural networks.** *Bioinformatics* 2001, **17(4)**:349-58.
11. Shen HB, Chou KC: **Ensemble classifier for protein fold pattern recognition.** *Bioinformatics* 2006, **22(14)**:1717-22.
12. Cheng J, Baldi P: **A machine learning information retrieval approach to protein fold recognition.** *Bioinformatics* 2006, **22(12)**:1456-63.
13. Barker JA, Thornton JM: **An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis.** *Bioinformatics* 2003, **19(13)**:1644-9.
14. Goyal K, Mohanty D, Mande SC: **PAR-3D: a server to predict protein active site residues.** *Nucleic Acids Research* 2007, **35(Web Server issue)**:W503-5.
15. Stark A, Russell RB: **Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.** *Nucleic Acids Research* 2003, **31(13)**:3341-4.
16. Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y: **Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces.** *Nucleic Acids Research* 2007, **35(Web Server issue)**:W526-30.
17. Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M: **Prediction of transition metal-binding sites from apo protein structures.** *Proteins* 2008, **70**:208-217.
18. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *Journal of Molecular Biology* 1993, **233**:123-38.
19. Kolodny R, Koehl P, Levitt M: **Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures.** *Journal of Molecular Biology* 2005, **346(4)**:1173-88.
20. Dobson PD, Doig AJ: **Predicting enzyme class from protein structure without alignments.** *Journal of Molecular Biology* 2005, **345**:187-199.
21. Alvarez MA, Yan C: **Exploring structural modeling of proteins for kernel-based enzyme discrimination.** *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* 2010, 1-5.
22. Punta M, Ofran Y: **The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function.** *PLoS Computational Biology* 2008, **4(10)**:e1000160.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature Genetics* 2000, **25**:25-29.
24. Lee D, Redfen O, C O: **Predicting protein function from sequence and structure.** *Nature Reviews: Molecular Cell Biology* 2007, **8(12)**:995-1005.
25. Soundararajan V, Raman R, Raguram S, Sasisekharan V, Sasisekharan R: **Atomic interaction networks in the core of protein domains and their native folds.** *PLoS One* 2010, **5(2)**:e9391.
26. da Silveira CH, Pires DE, Minardi RC, Ribeiro C, Veloso CJ, Lopes JC, Meira W Jr, Neshich G, Ramos CH, Habesch R, Santoro MM: **Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins.** *Proteins* 2009, **74(3)**:727-743.
27. Delaunay B: **Sur la sphere vide. A la memoire de Georges Voronoi.** *Izv Akad Nauk SSSR* 1934, **7**:793-800.
28. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC: **A gold standard set of mechanistically diverse enzyme superfamilies.** *Genome Biology* 2006, **7**:R8.
29. Jain P, Hirst JD: **Automatic structure classification of small proteins using random forest.** *BMC Bioinformatics* 2010, **11(364)**:1-14.
30. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH - a hierarchic classification of protein domain structures.** *Structure* 1997, **5(8)**:1093-108.
31. Eldén L: **Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms).** *Society for Industrial and Applied Mathematics* 2007.
32. Eldén L: **Numerical linear algebra in data mining.** *Acta Numerica* 2006, **15**:327-384.
33. Berry MW, Dumais ST, O'Brien GW: **Using linear algebra for intelligent information retrieval.** *SIAM review* 1995, **37(4)**:573-595.
34. del Castillo-Negrete D, Hirshman SP, Spong DA, D'Azevedo EF: **Compression of magnetohydrodynamic simulation data using singular value decomposition.** *Journal of Computational Physics* 2007, **222**:265-286.
35. Deerwester SC, Dumais ST, Furnas GW, Harshman RA, Landauer TK, Lochbaum KE, Streeter LA: **Computer information retrieval using latent semantic structure.** 1989.
36. Witten IH, Frank E: **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann;, second 2005.
37. Brenner SE, Koehl P, Levitt M: **The ASTRAL compendium for sequence and structure analysis.** *Nucleic Acids Research* 2000, **28**:254-256.

doi:10.1186/1471-2164-12-S4-S12
Cite this article as: Pires *et al.*: Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* 2011 **12**(Suppl 4):S12.

Apêndice B

Artigo 2: Bioinformatics

- Título: *aCSM: Noise-free graph-based signatures to large-scale receptor-based ligand prediction*
- Autores: **Pires, DEV**; Melo-Minardi, RC; da Silveira, CH; Meira Junior, W
- Periódico: Bioinformatics (*Impact Factor*: 5.46)
- Ano: 2012
- Em processo de revisão

BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–7

aCSM: Noise-free graph-based signatures to large-scale receptor-based ligand prediction

Douglas E. V. Pires[1,2*], Raquel C. de Melo-Minardi[1*],
Carlos H. da Silveira[3], Wagner Meira Jr.[1]

[1]Department of Computer Science, Universidade Federal de Minas Gerais, Brazil
[2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Brazil
[3]Advanced Campus at Itabira, Universidade Federal de Itajubá, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Receptor-ligand interactions are a central phenomenon in most biological systems. They are characterized by molecular recognition, a complex process mainly driven by physicochemical and structural properties of both receptor and ligand. Understanding and predicting these interactions are major steps towards protein ligand prediction, target identification, lead discovery and drug design.

Results: We propose a novel graph-based binding pocket signature called aCSM, that proved to be efficient and effective in handling large-scale protein ligand prediction tasks. We compare our results with those described in the literature and demonstrate that our algorithm overcomes the competitors techniques. Finally, we predict novel ligands for proteins from *Trypanosoma cruzi*, the parasite responsible for Chagas Disease, and validate them in silico via a docking protocol, showing the applicability of the method in suggesting ligands for pockets in a real-world scenario.

Availability and Implementation: Data sets and the source code are available at http://www.dcc.ufmg.br/∼dpires/acsm

Contact: dpires@dcc.ufmg.br and raquelcm@dcc.ufmg.br

1 INTRODUCTION

1.1 Background

Molecular recognition plays a fundamental role in most cellular processes. The conditions responsible for the binding and interaction of two or more molecules is a combination of conformational and physicochemical complementarity (Kahraman *et al.*, 2007). Understanding the receptor binding pocket requirements for this recognition process is a major step towards protein ligand prediction, target identification, lead discovery and drug design.

It is assumed that similar ligands have similar binding sites in terms of shape and physicochemical properties. Several methods were proposed to describe, compare and predict ligands to binding pockets. However, despite the relevant contributions of the majority of the works, methods that rely on multiple structure alignments and pairwise pocket comparisons might be prohibitively expensive for large-scale experiments. As the availability of biological data have been growing in a exponential fashion in the past years, scalability has become a crucial characteristic for the execution of such tasks in real-world scenarios.

To overcome these challenges we proposed a novel methodology for receptor-based protein ligand prediction, which is supported by a graph-based pocket signature. We extract distance patterns from protein pockets modeling them as atomic graphs and performing a noise and dimensionality reduction preprocessing step, which granted not only an improvement in efficacy in comparison to competitors works but also scalability to the methodology.

Atomic distance patterns perceive the structure arrangements of the protein and therefore, reflect its function. This way, using this information to describe ligand binding pockets is an appropriate strategy, given the close relationship between protein structure and function as well as the importance of shape complementarity in the molecular recognition process. Furthermore, considering the physicochemical properties in these patterns, also an important requirement for recognition, gives the description power needed to successfully describe, compare and predict protein-ligand interactions.

Receptor binding pockets can be seen as graphs where nodes are the protein atoms and the edges are the chemical interactions established among them. Topological and chemical properties can be extracted from these graphs and summarized in a molecular recognition signature. These compact signatures can then be used in large-scale ligand prediction tasks. In this work we derive a novel pocket signature from the Cutoff Scanning Matrix (CSM (Pires *et al.*, 2011)) which is essentially a graph-based signature successfully used for structural classification and function prediction tasks. We propose an atomic labeled version of the signature (henceforth called aCSM) that is independent of molecular orientation and does not require any ligand information in its calculation.

Given the complexity of the recognition process, these signatures are expected to be robust for ligand prediction. One of these difficulties, and an important source of noise in data, is ligand flexibility which leads to a great conformational diversity. For instance, Figure 1(a) illustrates five representatives of Nicotinamide Adenine Dinucleotide (NAD), a highly flexible ligand, obtained from the Protein Data Bank (PDB). They were aligned and their pockets calculated by a distance criterion. We can see that the

*To whom correspondence should be addressed

variability of conformations directly impacts on the binding pocket size and shape. Besides that, the induced fit mechanisms (Koshland Jr, 1958) as well as allosteric regulations (Monod *et al.*, 1963) may promote expressive conformational changes in the protein target.

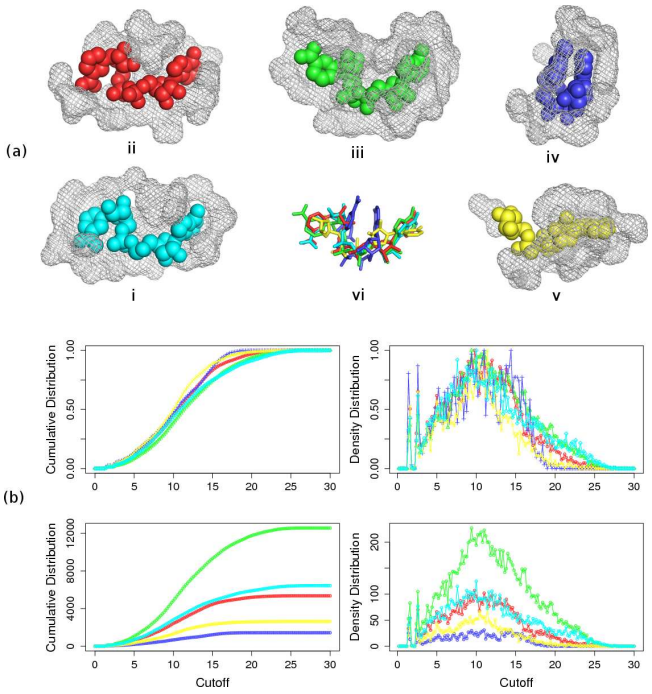

Fig. 1. Ligand conformational diversity. (a) Shows NAD molecules presenting different conformations and the impact in its respective pockets (calculated using a distance threshold of 5Å). The PDB ids considered were (a.i) 3KSD:Q (ligand in cyan), (a.ii) 1A5Z:A (ligand in red), (a.iii) 1NAH:A (ligand in green), (a.iv) 1ZRQ:B (ligand in blue), (a.v) 2OOR:B (ligand in yellow). (a.vi) Show the ligands aligned by the program LigAlign. The cumulative and density distribution of aCSM signatures, fully explained in the next section, are presented in the same colors in (b), considering normalized (top) and absolute values (bottom).

Other challenging factor is the several poses adopted by ligands in different pockets and its solvent accessibility when bound. Figure 2 presents an example where Flavin Adenine Dinucleotide (FAD) is bound to three pockets with very different degrees of solvent accessibility. It is clear that these factors could dramatically affect pocket shape and size for the same ligand, which may impose severe limitations to methods that rely solely on structural alignments. In our case, it is also a considerable source of noise for the proposed signatures, which are based on atomic distance patterns.

To deal with these challenges and eliminate inherent noise, we apply a Singular Value Decomposition (SVD)-based noise and dimensionality reduction strategy. A detailed description about the technique as well as references can be found in supplementary material. Figure 1(b) presents the proposed signatures for the NAD binding sites before noise reduction. We can see that, despite the similarity in the curves profile, we still see a considerable variability among them what is reduced by the data normalization achieved after SVD preprocessing. This preprocessing step is essential do extract from the original signatures, the components which are the

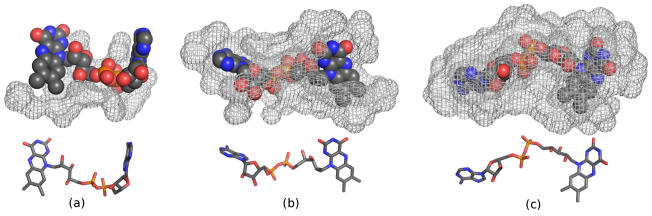

Fig. 2. Ligand solvent accessibility diversity. The figure shows FAD molecules with different degrees of solvent accessibility and its impact in the respective pocket (calculated using a distance threshold of 5Å). The PDB ids used were (a) 1O26:A, (b) 1AHV:A and (c) 1H83:C. Pymol CPK coloring scheme: Cs in gray, Ns in blue, Os in red and Ss in yellow.

most important to describe the pockets discarding redundancy and non-conserved dimensions.

1.2 Related works

In order to describe protein pockets to either compare them and / or predict their ligands, several methods have been proposed in the literature. Some of them are based on a paradigm of pocket similarity metrics. In (Davies *et al.*, 2007), the authors introduced a matching score for binding sites based on a probabilistic model and compare their metric with the Tanimoto Index. Protein binding pockets were compared by Spherical Harmonic Decompositions (Morris *et al.*, 2005), technique also used in a study of their shape variation (Kahraman *et al.*, 2007). More recently, in (Hoffmann *et al.*, 2010) the authors proposed a method to quantify pocket similarity by representing them as clouds of atoms, and comparing the resulting alignments with a convolution kernel. A measure of similarity was also derived in a recent work (Ueno *et al.*, 2012) from radial distribution functions (RDFs) of physicochemical properties of catalytic sites, information that was then used to cluster enzymes by function. In (Gonçalves-Almeida *et al.*, 2012), pockets are compared using hydrophobic patches represented by geometric centroids, and their conservation is detected despite sequence and structure dissimilarity.

Another set of methods attempts to compare ligand-binding sites based on multiple alignments. Similarity metrics were derived from the alignments of binding sites or cavity fingerprints represented by its physicochemical or topological properties in (Shulman-Peleg *et al.*, 2008; Schalon *et al.*, 2008) while the author of (Spitzer *et al.*, 2011) proposed a surface-based approach. There are also efforts that use multiple graph alignments and clique-based matching algorithms (Weskamp *et al.*, 2007; Najmanovich *et al.*, 2008) in order to perceive receptor-ligand interactions. Other alternative approaches in the study of binding mechanism include the use of Docking and Quantitative Structure-Activity Relationships techniques (QSARs) (Sippl, 2000).

1.3 Summary of results

We showed the proposed signatures successfully deal with the challenging aspects of large-scale ligand prediction achieving an Area Under ROC Curve (AUC) of 0.92 for a data set composed by more than 35,000 enzyme pockets. As far as we are concerned, no other method was tested with a data set of comparable volume. Despite the prominent variability of NAD, our methodology was

able to retrieve their pockets with an AUC up to 0.96. We recovered as well FAD sites presenting molecules with different solvent accessibilities achieving an AUC of 0.99. When compared to state-of-the-art methods, our approach achieves comparable or better results. Finally, we present a case study where we predict novel ligands for proteins from *Trypanosoma cruzi*, the parasite responsible for Chagas disease and validate them in silico via a docking protocol showing the applicability of the method in a real-world scenario.

2 MATERIALS AND METHODS

In this section we explain the basis for defining our noise free graph signatures, describe the data sets used in the experiments and explain the evaluation strategies. First, we describe the CSM method. Our strategy to reduce noise and dimensionality turning aCSM precise and robust as well as scalable to very large data sets is explained in detail in supplementary material as well as how the classification algorithms used work and why they were chosen. Finally, we explain how the method was validated. Details about the used quality measures are also available in supplementary material. Figure 3 shows the aCSM-based ligand prediction workflow. It is divided into the following main steps: data collection, signature generation and noise/dimensionality reduction, supervised learning, ligand prediction and validation. A more detailed workflow can be found in Figure 1 of supplementary data.

2.1 aCSM-based signatures

The Cutoff Scanning Matrix (CSM) is a protein structural signature proposed in (Pires *et al.*, 2011) and successfully employed in large-scale protein function prediction and structural classification tasks. The original CSM workflow generates, for each protein, a feature vector that represents distance patterns between protein residues represented by centroids which are then used as evidence for the classification procedures. To reduce noise as well as data dimensionality, Singular Value Decomposition (SVD) (Golub and Reinsch, 1970) was used as a preprocessing step. Inter-residue distance patterns were also subject of study of our previous study (da Silveira *et al.*, 2009) and showed to be conserved across protein folds.

In the present work we extend the inter-residue signature to an atomic level (atomic CSM, or aCSM for short). The aCSM-based signatures are generated as follows: for each protein we create a feature vector. First, we compute the Euclidean distance between all pairs of atoms and define a range of distances (cutoffs) to be considered and a distance step. We scan through these distances, computing the frequency of pairs of atoms that are close according to this distance threshold, i.e., the atoms in contact.

Furthermore, we propose in this work three new different types of aCSM-based signatures using atomic physicochemical properties.

- **aCSM**: generates one value per cutoff, corresponding to the number of atoms in contact according to this distance threshold.
- **aCSM-HP**: generates three values per cutoff, i.e., the frequency of hydrophobic-hydrophobic, hydrophobic-polar and polar-polar contacts.
- **aCSM-ALL**: considers eight categories: hydrophobic, positive, negative, acceptor, donor, aromatic, sulfur and neutral. The combination of these atoms labels generates 36 values per cutoff. The atoms classification were obtained by the program PMapper at pH 7. PMapper perceives pharmacophoric properties of atoms in a given molecular structures.

Algorithm 1 shows the function that calculates the atomic version of CSM. To compute a signature one must supply the following input parameters: a set of proteins and the atomic categories to be considered, a cutoff range (D_{MIN} and D_{MAX}) and a cutoff step (D_{STEP}) in which each cutoff is discretized. In line 1, we define the prototype of the aCSM function. In line 2, we iterate through each i of the proteins of the input data set. Line 3 shows the initialization of a variable used to index the signature array. In line 4, we call a function which computes the pairwise distances between all pairs of atoms of a protein and return and store this data in $distMatrix$. The loop in line 5 controls the iterations used to scan the $distMatrix$ to compute the signature. In line 6, we iterate through the considered atom classes and finally in lines 7-8 we call a function that computes the frequency of contacts for the current distance, between atoms of the given classes and store it in the corresponding signature array position. The aCSM generation runs in quadratic time, i.e., has time complexity of $O(n^2)$, where n is the number of atoms of the pocket, due to the pairwise distance computation. It is important to point out that the method is easily parallelizable, an important and desired characteristic for its efficient use in multi-core processor architectures.

In the experiments presented in this work, we vary the distance threshold from D_{MIN}=0.0Å to D_{MAX}=30.0Å, with a D_{STEP}=0.2Å, which generated for each type of signature a vector of 151, 453 and 5,436 entries for each protein. In Figure 2 from supplementary data, one can see through the pocket diameter distribution that 30.0Å accounts to approximately 95% of the pockets of the extensive enzyme data set.

The aCSM might also be presented as a graph-based signature, since the information regarding each cutoff distance represents the number of edges of an atomic contact graph assembled considering this cutoff. Notice also the method generality as it can be applied to predict both proteic and non-proteic ligands.

Algorithm 1 Atomic Cutoff Scanning Matrix calculation

```
1: function aCSM(ProteinSet, AtomClass, D_MIN, D_MAX, D_STEP)
2:     for all protein i ∈ (ProteinSet) do
3:         j = 0
4:         distMatrix ← calculateAtomicPairwiseDist(protein)
5:         for dist ← D_MIN; to D_MAX; step D_STEP do
6:             for all class ∈ (AtomClass) do
7:                 aCSM[i][j] ← getFrequency(distMatrix, dist, class)
8:                 j++
9:     return aCSM
```

2.2 Evaluation methodology

We evaluate the proposed method and compare it to other state-of-the-art algorithms using cross-validation computed metrics, specially AUC.

Cross-validation: is a traditional statistical analysis used to

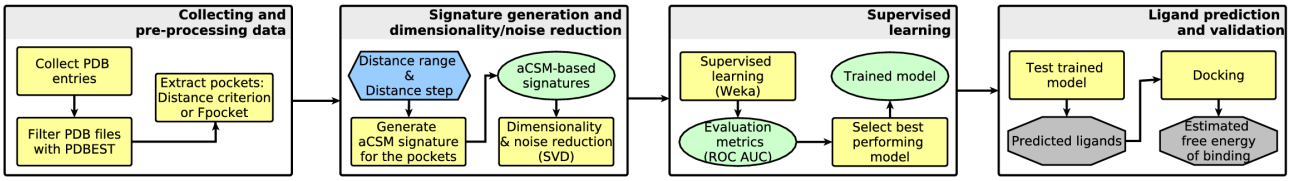

Fig. 3. aCSM-based ligand prediction workflow. The workflow is divided into four main steps: data collection, signature generation and noise/dimensionality reduction, supervised learning, ligand prediction and validation. Hexagonal blue boxes denote input files/parameters, ellipsoid green boxes are intermediate files generated, rectangular yellow boxes denote the intermediate steps, and the octagonal gray boxes the outputs, i.e. the predicted ligands and its estimated binding free energy.

estimate the performance of predictive models. It consists of partitioning the data set into two complementary subsets. The first is the training set used to build the model. The other is the test set used to measure the validity of the model. The data set is partitioned and the metrics are averaged over all the rounds. We use 10-fold cross validation in all the experiments but the comparative ones. For the comparisons, we use leave-one-out cross validation because this was the technique used by competitors works.

AUC: is the area under ROC curve. ROC curves are explained in more detail in supplementary material. They provide a visual tool for examining the trade-off between the ability of a classifier to correctly identify positive cases and the number of negative cases that are incorrectly classified. An interesting feature of these curves is that the area under curve (AUC) can be used as measure of accuracy in many applications. The AUC ranges from 0 to 1 and a random classifier would have an AUC of 0.5.

2.3 Data

2.3.1 Data sets In order to support multiple types of experiments to validate our method's quality, generality and real-world applicability, we used four different databases with different purposes:

a. **Large-scale enzyme data set:** In a previous work (Pires *et al.*, 2011) we have proposed a data set of the top 950 most-populated EC numbers, in terms of available structures, with at least 9 representatives per class, concerning 55,474 chains. This data set consists of reviewed enzymes from UniProt, i.e., the experimentally validated annotations from that database. Only ligands with 7 or more atoms were considered and also the pockets with less than 10 atoms were discarded. A total of 35,480 pockets were generated for 604 different ligands, with at least 10 representatives per ligand. This data set is used to show the applicability of the method for very diverse, real-world large enzyme database.
b. **Kahraman data set**: proposed by (Kahraman *et al.*, 2007), it comprises 100 protein binding sites that are non-evolutionary related, x-ray-solved, complexed with 10 different ligands with various sizes and flexibility (namely: AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD). This data set is used in comparisons of our method and its competitors.
c. **Hoffmann HD data set**: proposed by (Hoffmann *et al.*, 2010), it is formed by 100 protein pockets complexed with 10 different ligands of similar size and was assembled by the authors to complement the *Kahraman data set* since it have ligands of very different volumes.This data set is also used to compare our method with its competitors.
d. ***Trypanosoma cruzi* data set**: composed by *Trypanosoma cruzi* proteins. The criteria adopted for the protein selection was: proteins solved by x-ray crystallography, with resolution below 2.5Å. 104 PDB ids, comprising 200 chains, were gathered. We used a 5Å distance criteria to define the pockets. After removing crystallographic artifacts, 225 pockets were selected. This data set is used to rise candidate ligands, using the aCSM signatures, and validate them via a docking protocol.

2.3.2 Data preprocessing All protein structures were collected from the Protein Data Bank, filtered and preprocessed using the PDBest toolkit. The proteins chains were split in separate files and the binding pockets for each ligand were extracted.

2.3.3 Pocket computation Ligand pockets were extracted from protein structures in two ways:

- **Distance criterion:** considering a distance of 5Å, i.e., only atoms within 5Å from the ligand were selected. This criterion was used by the competitors works.
- **Geometric criterion:** using the software FPocket (Le Guilloux *et al.*, 2009) which is based on alpha-shapes theory. We chose the pocket which has the closest atom from the ligand, that is, the pocket which probably is more in contact with it.

3 RESULTS

In order to test and validate the ability of our signature to describe binding sites to support and aid in protein-ligand interaction prediction tasks we designed an extensive set of experiments. Firstly, we show our method can be used in large-scale ligand prediction and evaluate its precision in doing this task. Then, we compare the three proposed versions of aCSM signatures and evaluate which one presents the best descriptive power to ligand prediction. After that, we present the comparative results concerning state-of-the-art methods described in the literature and its respective data sets. Finally, we apply our methodology to pockets of *Trypanosoma cruzi* proteins and predict ligands to them, comparing the binding free energies of the ligands docked with receptors with those of real complexes available at the PDB, via a redocking protocol, and with ligands randomly chosen.

3.1 Large-scale experiments

Figure 4 presents the AUC of our method for the large-scale enzyme data set composed by reviewed enzymes from UniProt from which, more than 35,000 pockets were extracted. We can see that the methods successfully predicted ligands in every type of experiment described with precisions going from 0.6 to 0.92. In the next sections, we explain the variations of the method that generated the results showed in the figure.

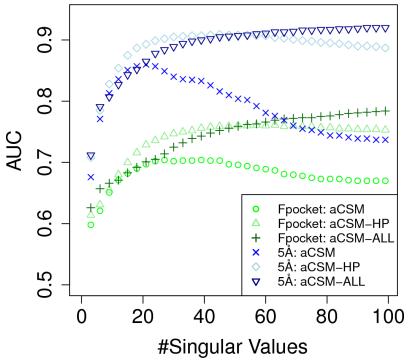

Fig. 4. Comparative prediction performance, in terms of the AUC metric, between two methods for defining the pocket: FPocket (green shades/three lower curves) and threshold distance (blue shades/three upper curves). For each method, the performance of the three signatures types proposed are also compared. The large-scale enzyme data set was employed in this experiment, as well as the KNN classifier.

3.1.1 Signature types evaluation In Figure 4, we compare aCSM, aCSM-HP and aCSM-ALL in terms of their AUC achieved with different number of singular values used to approximate the original matrix. In one hand, we can see that the more specific the signature is in terms of physicochemical atom properties the more precise it is in ligand prediction. With 100 singular values, for pockets extracted via a distance criterion (three upper curves), aCSM-ALL reaches an AUC of 0.92 as the basic aCSM presents an AUC of 0.75. On the other hand, with less than 20 singular values the difference between the different signatures is almost null reaching a high AUC score of 0.85.

It is interesting to notice that aCSM-ALL is the only one which seems to have benefited from the addition of many singular values. The first singular values respond to the higher data variability and are the most informative ones. As long as we add singular values to the signature, we add more and more noise to data as well as we demand more computational time. These results show that aCSM present an intrinsic limitation when 20 singular values are used and a pick of AUC of about 0.86 is achieved. aCSM-HP reaches more than 0.90 with about 40 singular values. aCSM-ALL AUC is improved when we add successive singular values and it does not converge until 100 singular values. It could indicate the absence of noise when we label atoms in a such a very specific way.

Table 1 from supplementary material shows the comparison of the three proposed signatures, for the large-scale enzyme dataset, in terms of several quality measures as well as in terms of mean execution time. The best results were achieved by aCSM-HP and aCSM-ALL after SVD pre-processing. aCSM-ALL is slightly better in terms of accuracy even though it takes twice the time to run in comparison with aCSM-HP. In conclusion, aCSM-ALL is the better choice being the most accurate and having a non-prohibitive execution time.

3.1.2 Pocket detection method influence Figure 4 also shows the comparison of the aCSM signatures computed for pockets delimited using distance and geometric criterion (three lower curves). We can see that using the geometric method the results are systematically about 15% worse than simply with distance criteria. This is probably due to loss of important molecular information when using Fpocket algorithm. Even if we aggregate more atoms than the ones that were in fact accessible in the pocket with 5Å cutoff, our method behaves robustly, being able to discard unnecessary or irrelevant information. Figure 3 of supplementary data shows that for cutoff distances greater than 5Å the predictive performance of our method increases.

We use 5Å as a cutoff criterion because it was the same adopted by the competitors works. However, this value seemed to be defined arbitrarily and not necessarily reflects the best possible cutoff for every method. In order to evaluate this hypothesis we contrast our signature performance according to pocket distance criterion for the large-scale enzyme data set. In fact, in Figure 4 of supplementary data, we show that the best distance criterion for the aCSM signature, using the KNN classifier, was actually 6.0Å. This value is in agreement with other authors that also have investigated about the best atomic cutoff when the network of contacts is computed using a heavy atom proximity criterion (Zhang *et al.*, 1997; Kamagata and Kuwajima, 2006). It is important to stress that with this cutoff, we have minimum noise in our signatures, as the best performing SVD cutoff is chosen.

3.2 Comparison with state-of-the-art methods

The experiments described below were performed in two data sets (*Hoffmann HD* and *Kahraman*) already used by several related studies in order to compare them to our protein pocket signature *aCSM*. Leave-one-out cross validation was used in all experiments regarding these two data sets, the same methodology was employed in the related works.

Table 1 summarizes the results obtained. The aCSM signature achieved better results, considering the AUC score, in comparison to the other methods with low standard deviation. It is important to stress that leave-one-out cross validation is computationally demanding and not suitable for a large-scale, real-world scenario. Moreover, the two aforementioned data sets are small (only 100 pockets) and, in the case of the Kahraman data set, divided into very unbalanced classes which makes the learning process of the classifiers very difficult.

3.3 Case study: predicting ligands for *T. cruzi* proteins

Chagas Disease is a tropical infection caused by the protozoan parasite *T. cruzi* that affects about 8 million people in Latin America (Rassi Jr. *et al.*, 2010) and is the leading etiology of non-ischemic heart disease worldwide. Unfortunately, the two available drugs for treatment (Nifurtimox and Benznidazole) have potential toxic side effects and variable efficacy (Canavaci *et al.*, 2010). The limitation of the current available treatment and interventions have been motivating several efforts towards the development of new

Table 1. Comparative results evaluated by the mean and standard deviation of the AUC score. The aCSM-ALL was the best performing signature for these experiments. AUC values were directly obtained from (Hoffmann *et al.*, 2010; Spitzer *et al.*, 2011) and the results for the aCSM signature were generated using Multinomial Logistic Regression.

Method	Dataset	AUC
Sequence	Kahraman	0.550 ± 0.08
MultiBind	Kahraman	0.715 ± 0.17
SHD	Kahraman	0.770
PSIM	Kahraman	0.790 ± 0.19
sup-PI	Kahraman	0.815 ± 0.13
sup-CK_L	Kahraman	0.861 ± 0.13
aCSM signature	**Kahraman**	$\mathbf{0.901 \pm 0.07}$
Sequence	Hoffmann HD	0.577 ± 0.09
MultiBind	Hoffmann HD	0.690 ± 0.14
sup-PI	Hoffmann HD	0.702 ± 0.19
sup-CK_L	Hoffmann HD	0.752 ± 0.16
PSIM	Hoffmann HD	0.760 ± 0.15
aCSM signature	**Hoffmann HD**	$\mathbf{0.804 \pm 0.13}$

drugs or a vaccine against *T. cruzi*. In fact, a recent study (Lee *et al.*, 2010) proposed a decision analytic Markov model that indicated that such vaccine could provide a substantial economic benefit. Some recent approaches to this problem described in the literature include screening efforts aiming inhibitors for *T. cruzi* known targets and development of high-throughput assays to validate anti-*T. cruzi* compounds (Canavaci *et al.*, 2010).

In this section, we used trained classification models in order to predict potential novel ligands to *T. cruzi* proteins with structures available at the PDB.

After an extensive analysis of the proposed signatures we selected the best performing model trained in the biggest data sets considered (more than 35,000 pockets). The pockets obtained from the *T. cruzi* proteins were tested against this model and a single ligand were predicted for each pocket. The KNN algorithm was used.

In order to validate the predictions we performed the docking of the ligands in the *T. cruzi* pockets using AUTODOCK. We compare the energies of binding of the predicted ligands with the ones from real ligands from the crystallographic complexes via a redocking protocol. To access the methods statistical significance, we compared our results with a null model. We selected for each pocket three independent random / null ligands from the pool of the training data set. The docking workflow adopted in the present work is shown in the Figure 6 of the supplementary data.

In Figure 5, we can see that the energy distribution for aCSM predictions is more similar in shape to the redocking energy profile than the profile of the null models. Paired t-tests reveal a high p-value (0.26) between aCSM prediction and redocking means, but a low p-value ($1.2e^{-9}$) for aCSM prediction and null models. This strongly suggests that ligands found by aCSM may have the same binding free energy profile of redocking ligands, but they may differ significantly from null ligands.

In summary, we showed that the binding free energies for ligands predicted by aCSM are better (lower) in comparison with those predicted by the null models. Furthermore, the energies from a redocking protocol are indistinguishable from those obtained for aCSM prediction.

4 CONCLUSIONS

In the present work we proposed a novel, scalable, graph-based pocket signature called aCSM. It prospects distance patterns from the atoms that compose the binding pockets generating a feature vector that represents a cumulative edge count of contact graphs defined for different cutoff distances, which are used as evidence by supervised learning algorithms. Singular Value Decomposition is also used as a preprocessing step to reduce dimensionality, lessen computational costs and grant scalability to the methodology, and also reducing the inherent noise of the data, which increased the success rate of the predictions. Some of the most remarkable advantages of the aCSM signatures is that it does not require any ligand information in its calculation and also is independent of molecular orientation. Additionally, our algorithm presents a notable signature generality since it may be applied to predict both proteic and non-proteic ligands to any type of biomolecular target (not only protein).

aCSM was successful when applied in ligand prediction tasks, presenting compatible or superior efficacy in comparison to state-of-the-art competitors. Besides, as a requirement and demand for its application to databases that are continuously growing, it proved scalable for large-scale scenarios, and was able to perform well in a dataset composed by more than 35,000 pockets. On top of that, we applied the methodology in order to predict potential novel inhibitors to *T. cruzi* proteins. The validation of this step via docking confirmed that inhibitors predicted represent good candidates for further experimental validation.

We intend to predict inhibitors for proteins from other pathogenic organisms of interest, a study already in progress in our group. Finally, we plan to expand the signature to ligand-based lead discovery.

5 ACKNOWLEDGEMENTS

This work was supported by the Brazilian agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

REFERENCES

Canavaci, A. M. C., Bustamante, J. M., Padilla, A. M., Perez-Brandan, C. M., Simpson, L. J., Xu, D., Boehlke, C. L., and Tarleton, R. L. (2010). In vitro and in vivo high-throughput assays for the testing of anti-trypanosoma cruzi compounds. *PLoS Neglected Tropical Diseases*, **4**(7), e740.

da Silveira, C. H., Pires, D. E. V., Melo-Minardi, R. C., Ribeiro, C., Veloso, C. J. M., Lopes, J. C. D., Meira Jr, W., Neshich, G., Ramos, C. H. I., Habesch, R., and Santoro, M. M. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and

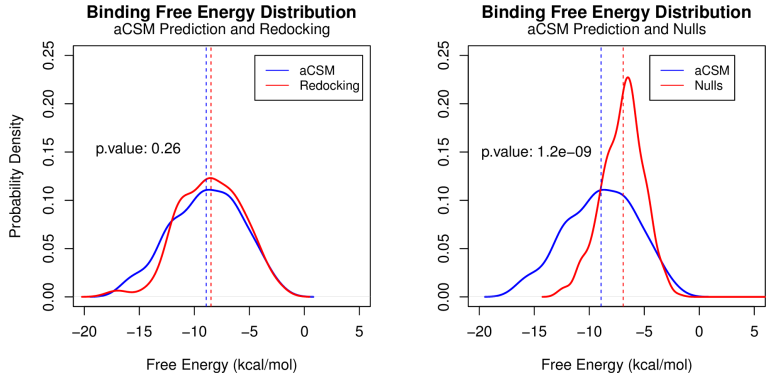

Fig. 5. Comparison of binding free energies distribution of the docked complexes for aCSM prediction, redocking and null models. Dashed lines indicate the mean values and the student t-test p-values for the significance of the means are also presented. Binding free energies for ligands predicted by aCSM are lower (better) in comparison with those predicted by the null models and indistinguishable from those obtained via a redocking protocol.

cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function and Bioinformatics*, **74**(3), 727–743.

Davies, J. R., Jackson, R. M., Mardia, K. V., and Taylor, C. C. (2007). The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, **23**(22), 3001–3008.

Golub, G. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**(5), 403–420.

Gonçalves-Almeida, V. M., Pires, D. E. V., de Melo-Minardi, R. C., da Silveira, C. H., Meira Jr, W., and Santoro, M. M. (2012). HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, **28**(3), 342–349.

Hoffmann, B., Zaslavskiy, M., Vert, J. P., and Stoven, V. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics*, **11**, 99.

Kahraman, A., Morris, R. J., Laskowski, R. A., and Thornton, J. M. (2007). Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*, **368**(1), 283–301.

Kamagata, K. and Kuwajima, K. (2006). Surprisingly high correlation between early and late stages in non-two-state protein folding. *Journal of Molecular Biology*, **357**(5), 1647–1654.

Koshland Jr, D. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **44**(2), 98.

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

Lee, B. Y., Bacon, K. M., Connor, D. L., Willig, A. M., and Bailey, R. R. (2010). The potential economic value of a trypanosoma cruzi (chagas disease) vaccine in latin america. *PLoS Neglected Tropical Diseases*, **4**(12), e916.

Monod, J., Changeux, J., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of molecular biology*, **6**(4), 306–329.

Morris, R. J., Najmanovich, R. J., Kahraman, A., and Thornton, J. M. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **21**(10), 2347–2355.

Najmanovich, R., Kurbatova, N., and Thornton, J. M. (2008). Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, **24**(16), i105–111.

Pires, D. E. V., Melo-Minardi, R. C., Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira Jr, W. (2011). Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12 Suppl 4**, S12.

Rassi Jr., A., Rassi, A., and Marin-Neto, J. A. (2010). Chagas disease. *Lancet*, **375**(9723), 1388–1402.

Schalon, C., Surgand, J. S., Kellenberger, E., and Rognan, D. (2008). A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Structure, Function and Bioinformatics*, **71**(4), 1755–1778.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. (2008). MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Research*, **36**(Web Server issue), W260–264.

Sippl, W. (2000). Receptor-based 3D QSAR analysis of estrogen receptor ligands - merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *Journal of Computer-Aided Molecular Design*, **14**, 559–572.

Spitzer, R., Cleves, A. E., and Jain, A. N. (2011). Surface-based protein binding pocket similarity. *Proteins: Structure, Function and Bioinformatics*, **79**(9), 2746–2763.

Ueno, K., Mineta, K., Ito, K., and Endo, T. (2012). Exploring functionally related enzymes using radially distributed properties of active sites around the reacting points of bound ligands. *BMC Structural Biology*, **12**(1), 5.

Weskamp, N., Hullermeier, E., Kuhn, D., and Klebe, G. (2007). Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**(2), 310–320.

Zhang, C., Vasmatzis, G., Cornette, J. L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology*, **267**(3), 707–726.

Apêndice C

Classificação de Átomos para Assinaturas aCSM

Código do Resíduo	**Código do Átomo**	**Caráter**
ALA	CB	Hidrofóbico
ARG	CB	Hidrofóbico
ARG	CG	Hidrofóbico
ARG	CD	Hidrofóbico
ASN	CB	Hidrofóbico
ASP	CB	Hidrofóbico
CYS	CB	Hidrofóbico
GLN	CB	Hidrofóbico
GLN	CG	Hidrofóbico
GLU	CB	Hidrofóbico
GLU	CG	Hidrofóbico
HIS	CB	Hidrofóbico
HIS	CG	Hidrofóbico
HIS	CD2	Hidrofóbico
HIS	CE1	Hidrofóbico
ILE	CB	Hidrofóbico
ILE	CG1	Hidrofóbico
ILE	CG2	Hidrofóbico
ILE	CD1	Hidrofóbico
LEU	CB	Hidrofóbico
LEU	CG	Hidrofóbico
LEU	CD1	Hidrofóbico

LEU	CD2	Hidrofóbico
LYS	CB	Hidrofóbico
LYS	CG	Hidrofóbico
LYS	CD	Hidrofóbico
MET	CB	Hidrofóbico
MET	CG	Hidrofóbico
MET	CE	Hidrofóbico
PHE	CB	Hidrofóbico
PHE	CG	Hidrofóbico
PHE	CD1	Hidrofóbico
PHE	CD2	Hidrofóbico
PHE	CE1	Hidrofóbico
PHE	CE2	Hidrofóbico
PHE	CZ	Hidrofóbico
PRO	CB	Hidrofóbico
PRO	CG	Hidrofóbico
PRO	CD	Hidrofóbico
THR	CG2	Hidrofóbico
TRP	CB	Hidrofóbico
TRP	CG	Hidrofóbico
TRP	CD1	Hidrofóbico
TRP	CD2	Hidrofóbico
TRP	CE2	Hidrofóbico
TRP	CE3	Hidrofóbico
TRP	CH2	Hidrofóbico
TRP	CZ	Hidrofóbico
TRP	CZ2	Hidrofóbico
TRP	CZ3	Hidrofóbico
TYR	CB	Hidrofóbico
TYR	CG	Hidrofóbico
TYR	CD1	Hidrofóbico
TYR	CD2	Hidrofóbico
TYR	CE1	Hidrofóbico
TYR	CE2	Hidrofóbico
TYR	CZ	Hidrofóbico
VAL	CB	Hidrofóbico
VAL	CG1	Hidrofóbico

VAL	CG2	Hidrofóbico
ARG	NH1	Positivo
ARG	NH2	Positivo
HIS	ND1	Positivo
HIS	NE2	Positivo
LYS	NZ	Positivo
ASP	OD1	Negativo
ASP	OD2	Negativo
GLU	OE1	Negativo
GLU	OE2	Negativo
ALA	O	Aceptor
ARG	O	Aceptor
ASN	O	Aceptor
ASN	OD1	Aceptor
ASP	O	Aceptor
ASP	OD1	Aceptor
ASP	OD2	Aceptor
CYS	O	Aceptor
GLN	O	Aceptor
GLN	OE1	Aceptor
GLU	O	Aceptor
GLU	OE1	Aceptor
GLU	OE2	Aceptor
GLY	O	Aceptor
HIS	O	Aceptor
ILE	O	Aceptor
LEU	O	Aceptor
LYS	O	Aceptor
MET	O	Aceptor
PHE	O	Aceptor
PRO	O	Aceptor
SER	O	Aceptor
THR	O	Aceptor
TRP	O	Aceptor
TYR	O	Aceptor
VAL	O	Aceptor
ALA	N	Doador

ARG	N	Doador
ARG	NE	Doador
ARG	NH1	Doador
ARG	NH2	Doador
ASN	N	Doador
ASN	ND2	Doador
ASN	OD1	Doador
ASP	N	Doador
CYS	N	Doador
GLN	N	Doador
GLN	NE2	Doador
GLU	N	Doador
GLY	N	Doador
HIS	N	Doador
HIS	ND1	Doador
HIS	NE2	Doador
ILE	N	Doador
LEU	N	Doador
LYS	N	Doador
LYS	NZ	Doador
MET	N	Doador
PHE	N	Doador
PRO	N	Doador
SER	N	Doador
SER	OG	Doador
THR	N	Doador
THR	OG1	Doador
TRP	N	Doador
TRP	NE1	Doador
TYR	N	Doador
TYR	OH	Doador
VAL	N	Doador
HIS	CG	Aromático
HIS	ND1	Aromático
HIS	CD2	Aromático
HIS	CE1	Aromático
HIS	NE2	Aromático

PHE	CG	Aromático
PHE	CD1	Aromático
PHE	CD2	Aromático
PHE	CE1	Aromático
PHE	CE2	Aromático
PHE	CZ	Aromático
TRP	CG	Aromático
TRP	CD1	Aromático
TRP	CD2	Aromático
TRP	NE1	Aromático
TRP	CE2	Aromático
TRP	CE3	Aromático
TRP	CZ2	Aromático
TRP	CZ3	Aromático
TRP	CH2	Aromático
TYR	CD1	Aromático
TYR	CD2	Aromático
TYR	CE1	Aromático
TYR	CE2	Aromático
TYR	CG	Aromático
TYR	CZ	Aromático
CYS	S	Enxofre
MET	SD	Enxofre

Tabela C.1: Tabela de categorias de átomos para cálculo das assinatura aCSM-HP e aCSM-ALL. Classificação obtida a partir do programa PMapper [ChemAxon, 2012] em pH 7. Átomos não contemplados na tabela são considerados *neutros*.