

**GRASP: UMA ESTRATÉGIA DE
APRENDIZAGEM SUPERVISIONADA BASEADA
EM GRAFOS DE VIZINHANÇA DE RESÍDUO
PARA PREVISÃO DE SÍTIO DE LIGAÇÃO**

CHARLES ABREU SANTANA

**GRASP: UMA ESTRATÉGIA DE
APRENDIZAGEM SUPERVISIONADA BASEADA
EM GRAFOS DE VIZINHANÇA DE RESÍDUO
PARA PREVISÃO DE SÍTIO DE LIGAÇÃO**

Tese apresentada ao Programa Interunidades 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.

ORIENTADORA: SABRINA DE AZEVEDO SILVEIRA
COORIENTADORA: RAQUEL CARDOSO DE MELO MINARDI

Belo Horizonte

Agosto de 2021

© 2021, Charles Abreu Santana.
Todos os direitos reservados.

Abreu Santana, Charles

D1234p / Charles Abreu Santana. — Belo Horizonte, 2021
xiv, 97 f. : il. ; 29cm

Tese (doutorado) — Federal University of Minas
Gerais

Orientador: Sabrina de Azevedo Silveira

1. proteína. 2. sitio de ligação. 3. grafo.
4. aprendizagem supervisionada. I. Título.

CDU 519.6*82.10



UNIVERSIDADE FEDERAL DE MINAS GERAIS
 Instituto de Ciências Biológicas
 Programa de Pós-graduação em Bioinformática

ATA DA DEFESA DE TESE

CHARLES ABREU SANTANA

Às quatorze horas do dia **31 de agosto de 2021**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho de Charles Abreu Santana intitulado: "**GRaSP: uma estratégia de aprendizagem supervisionada baseada em grafos de vizinhança de resíduo para previsão de sítio de ligação**", requisito para obtenção do grau de Doutor em **Bioinformática**. Abrindo a sessão, a Presidente da Comissão, **Dra. Sabrina de Azevedo Silveira**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	Indicação
Dra. Sabrina de Azevedo Silveira - Orientadora	Universidade Federal de Viçosa	Aprovado
Dra. Maria Goreti de Almeida Oliveira	Universidade Federal de Viçosa	Aprovado
Dr. Lucas Bleicher	Universidade Federal de Minas Gerais	Aprovado
Dra. Mariana Torquato Quezado de Magalhaes	Universidade Federal de Minas Gerais	Aprovado
Dr. Carlos Henrique da Silveira	Universidade Federal de Itajubá	Aprovado

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

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

Belo Horizonte, 31 de agosto de 2021.



Documento assinado eletronicamente por **Sabrina de Azevedo Silveira, Usuário Externo**, em 31/08/2021, às 17:14, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

Documento assinado eletronicamente por **Mariana Torquato Quezado de Magalhaes, Professora do Magistério Superior**, em 31/08/2021, às 17:30, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Lucas Bleicher, Professor do Magistério Superior**, em 31/08/2021, às 17:32, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Carlos Henrique da Silveira, Usuário Externo**, em 31/08/2021, às 17:35, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maria Goreti de Almeida Oliveira, Usuário Externo**, em 01/09/2021, às 10:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0934571** e o código CRC **24AE9502**.

À família, aos amigos e a todos que dedicarem seu tempo à esta leitura.

Agradecimentos

Agradeço primeiramente a oportunidade de desfrutar desta passagem chamada vida. Como uma vez foi dito, "a vida é curta para ser pequena"(Benjamin Disraeli), logo viverei grandiosamente. Já basta que a vida seja curta para cogitar a possibilidade de apequená-la.

Agradeço à minha família pelo apoio, em especial à minha mãe Eliene, pela preocupação e atenção, aceitando, mesmo sob a dor da distância, que seu filho desbrave o mundo.

Agradeço à Universidade Federal de Minas Gerais e ao Programa Interunidades de Pós-Graduação em Bioinformática da UFMG, junto aos seus professores, que sempre se dedicam ao máximo para passar o conhecimento com excelência. Agradeço especialmente à professora Sabrina de Azevedo Silveira, pela paciência, inteligência e dedicação em guiar-me, de forma sábia, pelo laborioso caminho acadêmico.

Agradeço ao financiamento fornecido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), fundamental para que eu me dedicasse exclusivamente a este trabalho.

*“God, grant me the serenity to accept the things I cannot change, courage to change
the things I can, and wisdom to know the difference.”*

(Serenity Prayer)

Resumo

As proteínas são macromoléculas importantes para a manutenção dos sistemas biológicos e que participam de processos vitais para a célula. A atividade proteica é desempenhada através de interações físico-químicas entre a proteína com outras moléculas denominadas ligantes. Sejam compostos orgânicos, íons metálicos, ácidos nucleicos ou até mesmo outras proteínas, os ligantes acoplam-se à proteína para que sua atividade seja devidamente desempenhada. A região na proteína onde essas interações ocorrem são chamadas de sítios de ligação. A identificação e caracterização dessas regiões é de suma importância para determinar a função de uma proteína, que é uma das etapas necessárias em áreas como o planejamento e desenvolvimento de novos fármacos. Devido a questões experimentais, a localização dessas regiões pode não ser trivial, sendo necessário o apoio de métodos automáticos para auxiliar na sua identificação. Nesta tese é proposto o GRaSP, uma estratégia baseada em aprendizagem de máquina para previsão de sítio de ligação que utiliza como insumo grafos de vizinhança de resíduo. A partir de experimentos usando bases de dados de estruturas de proteínas diversas, o GRaSP demonstrou ser robusto, apresentando resultados compatíveis ou superiores em relação às ferramentas já consolidadas pela literatura. Além disso, devido à modelagem simples e informativa disponibilizada pelos grafos, o algoritmo mostrou-se eficiente. Enquanto métodos já consolidados levam em torno de 5 horas de processamento, em seus servidores, para estruturas de proteínas com aproximadamente 300 resíduos, o GRaSP é capaz de processá-las com um tempo médio de 20 segundos.

Palavras-chave: Biologia Computacional, Proteínas, Sítios de Ligação, Teoria dos grafos, Aprendizado de Máquina Supervisionado.

Abstract

Proteins are macromolecules crucial for the maintenance of biological systems and participate in vital processes for the cell. Protein activity is performed through physicochemical interactions between the protein and other molecules called ligands. These ligands comprise organic compounds, metal ions, nucleic acids or even other proteins, in which attach to the protein so that its activity is properly performed. The region on the protein where these interactions take place is called binding sites. The identification and characterization of these regions is crucial to determine the function of a protein, which is one of the necessary steps in areas such as the design and development of new drugs. Due to experimental issues, the location of these regions may not be trivial, requiring the support of automatic methods to assist in their identification. In this thesis, the GRASP is proposed, a machine learning-based strategy for binding site prediction that uses residue neighborhood graphs as input. From experiments using databases of different protein structures, GRASP proved to be robust, presenting compatible or better results in relation to the tools already consolidated in the literature. Furthermore, due to the simple and informative modeling provided by the graphs, the algorithm proved to be efficient. While already consolidated methods take around 5 hours of processing for protein structures with approximately 300 residues, the GRASP is able to process them in an average time of 20 seconds.

Keywords: Computational Biology, Proteins, Binding Sites, Graph Theory, Supervised Machine Learning.

Lista de Figuras

1.1	Interações não-covalentes entre a proteína tirosina-quinase Bcr-Abl humana e o ligante imatinibe (PDB 2hyy)	16
2.1	Esquema das etapas seguidas por métodos energéticos para buscar sítios de ligação em proteínas.	22
2.2	Modelos dos mecanismos de interação proteína-ligante, (a) chave-fechadura e (b) encaixe induzido	23
2.3	Tela principal do PrankWeb.	24
2.4	Página de resultados preditivos ferramenta web COACH-D.	25
3.1	Interações do tipo ponte de hidrogênio (tracejadas em amarelo) estabelecidas para formar uma alfa-hélice.	28
3.2	Esquema um resíduo ρ e sua vizinhança modelados como um grafo em uma proteína.	30
3.3	Esquema de construção do vetor de características de um resíduo a partir do seu grafo de vizinhança.	32
3.4	Esquema de amostragem de exemplos do repositório de proteínas para construção do modelo preditivo do GRaSP.	34
3.5	Exemplo de árvore de decisão induzida de uma base de dados para inferir a existência de uma ponte salina entre um par de átomos.	35
3.6	Esquema ilustrativo do dilema viés-variância. Em (a) é mostrado um esquema das predições realizadas por um classificador com viés, que apresenta um erro sistemático, deslocando suas predições do centro. Já em (b), é mostrado o esquema das predições de um classificador com variância, onde há uma dispersão nos resultados.	36
3.7	Estratégia de balanceamento e classificação mediada por voto majoritário. Os dados originais em (a) são devidamente particionados e balanceados (b), onde classificadores são inferidos em (c) e a classificação final é alcançada através do voto majoritário (d).	38

4.1	Representação estrutural das predições realizadas pelo GRaSP no experimento CASP10.	48
4.2	Métodos participantes do CASP10, junto com o GRaSP, ordenados pelo MCC médio de suas predições.	49
4.3	Estruturas da HIV protease, identificador PDB 4PHV em cor bege (holoproteína) acoplada ao ligante VAC e identificador PDB 3PHV em cor magenta (apoproteína) sobrepostas (a), acompanhadas da predição de seus sítios em cor alaranjada, 4PHV em (b) e 3PHV em (c).	51
4.4	Estrutura da quimotripsina, (a) complexo proteína-ligante (pdb 2CHA), (b) predição dos resíduos de sítio realizada pelo GRaSP.	52
4.5	Resultados do GRaSP para a estrutura 1G2O com múltiplos sítios.	53
4.6	Desempenho preditivo do GRaSP comparado a métodos centrados em cavidades.	54
4.7	Página inicial do GRaSPweb.	55
4.8	Página de submissão do GRaSPweb.	56
4.9	Fluxo de execução do GRaSPweb.	56
4.10	Exemplo de resultado apresentado pelo GRaSPweb após a predição.	57
4.11	Exemplo de sugestão de ligantes feita pelo GRaSPweb.	58

Lista de Tabelas

3.1	Critérios de distância (em Å) e propriedades físico-químicas dos átomos envolvidos em cada tipo de interação.	30
4.1	Resultados comparativos entre o GRaSP, COACH e seus métodos.	45
4.2	Comparação entre GRaSP e COACH usando a base de dados B44/U44. . .	45
4.3	Estruturas alvo do CASP10 e seus respectivos resíduos de sítio de ligação.	46
4.4	Resíduos de sítios preditos pelo GRaSP	47
4.5	Resultados do GRaSP para bases de dados diversos.	50

Sumário

Agradecimentos

Resumo

Abstract

Lista de Figuras

Lista de Tabelas

1	Introdução	15
1.1	Objetivos	18
1.2	Organização do Texto	19
2	Revisão da Literatura	20
2.1	Métodos preditivos para sítio de ligação	20
2.1.1	Métodos baseados em estrutura	21
3	Métodos	27
3.1	Construindo grafos de vizinhança	27
3.1.1	Construindo vetores de características	31
3.2	Construção da base de exemplos	32
3.3	Aprendizagem supervisionada	33
3.4	Tratando o desbalanceamento dos dados	37
3.5	Métricas de avaliação	38
3.6	Bases de Dados	41
4	Resultados	43
4.1	Experimentos centrados no resíduo	44
4.1.1	Comparando GRaSP com o método COACH	44

4.1.2	Experimento CASP 10	46
4.1.3	Desempenho do GRaSP em dados diversos	49
4.2	Experimentos centrados na cavidade	52
4.3	GRaSPweb	55
5	Considerações Finais	59
	Referências Bibliográficas	61
	Apêndice A Artigo publicado em periódico	67
	Apêndice B Informações adicionais	77
B.1	Propriedades físico-químicas dos átomos	78
B.2	Resultados: base de dados COACH	79
B.3	Resultados: base de dados B44	89
B.4	Resultados: base de dados U44	90
B.5	Resultados do experimento CASP10	91
B.6	Resultados: base de dados ASTEX	96

Capítulo 1

Introdução

As proteínas são macromoléculas imprescindíveis para a vida. A diversidade de funções desempenhadas por essa classe de biomoléculas faz com que elas estejam envolvidas em inúmeros processos biológicos que norteiam o ciclo vital dos organismos, como catálise enzimática, metabolismo, sinalização e manutenção da integridade estrutural da célula (Zhao et al., 2020). Apesar de serem consideradas engrenagens essenciais para sustentar a máquina biológica dos seres vivos, as proteínas, em sua maioria, não trabalham de forma independente e precisam interagir com outros componentes para conseguir desempenhar seu papel biológico. Tais elementos interagentes, que trabalham em conjunto com as proteínas, são comumente denominados como ligantes, que por sua vez englobam classes de moléculas de pequena magnitude, como compostos orgânicos, íons, fármacos, e também macromoléculas, como ácidos nucleicos e até mesmo outras proteínas (Macari et al., 2019).

A interação proteína-ligante emerge através do reconhecimento molecular que é mediado por interações de alta especificidade e afinidade para formar um complexo. Essas interações são compostas em sua maioria por ligações não-covalentes, como pontes de hidrogênio, interações eletrostáticas e forças atrativas oriundas do efeito hidrofóbico, todas elas detentoras de padrões específicos e intrinsecamente complementares (Du et al., 2016). A Figura 1 mostra como o ligante *imatinibe* acopla-se à proteína tirosina-quinase Bcr-Abl, formando uma série de interações não-covalentes, como pontes de hidrogênio e contatos hidrofóbicos. O imatinibe inibe a atividade enzimática da Bcr-Abl, prevenindo a transdução de sinais necessários para a proliferação celular no tratamento da leucemia mieloide crônica (Azevedo et al., 2017).

Radivojac et al. (2013) define o conceito de função proteica como aquilo que descreve aspectos químicos, celulares e fenotípicos de eventos moleculares que envolve a proteína, incluindo como esta interage com outros elementos do ambiente. Caracterizar

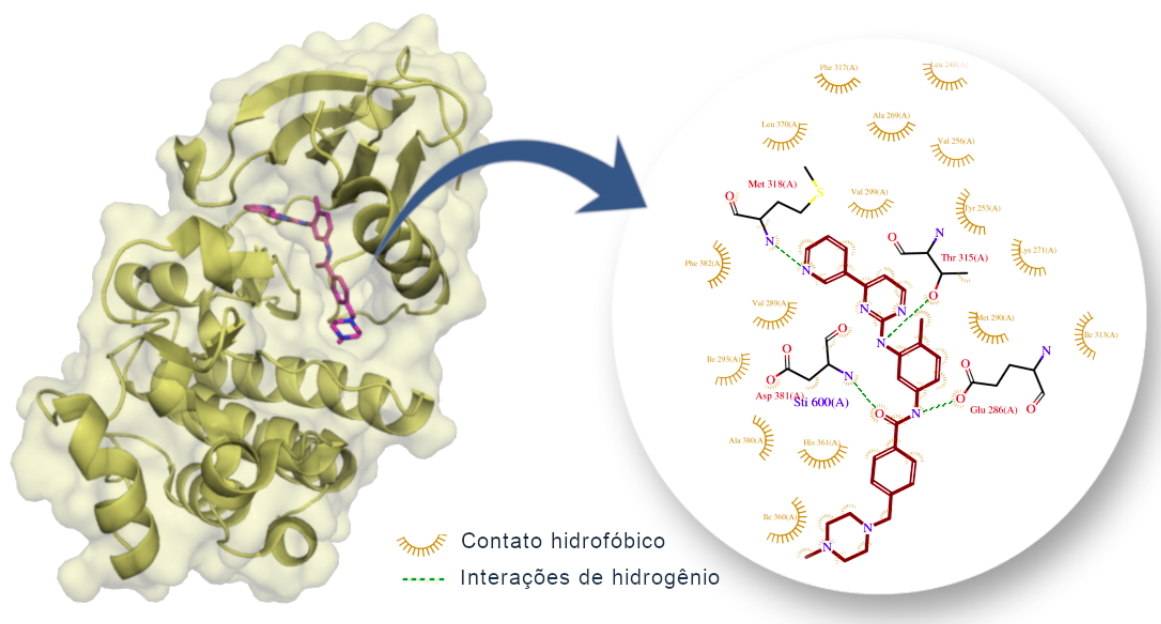


Figura 1.1. Interações não-covalentes entre a proteína tirosina-quinase Bcr-Abl humana e o ligante imatinibe (PDB 2hyy)

a função de uma proteína é importante para entender a vida do ponto de vista molecular e tem implicações diretas nas áreas biomédicas e farmacêuticas (Somody et al., 2017; Tran-Nguyen et al., 2018; Kana & Brylinski, 2019). Entretanto, como o número de genomas sequenciados cresce rapidamente, existe uma dificuldade inerente ao problema de anotação funcional, isso devido a discrepância entre o ritmo em que as sequências proteicas vêm sendo disponibilizadas nos bancos de dados e o ritmo em que elas são devidamente caracterizadas.

O custo para que a caracterização experimental possa escalar de forma a atender o volume de dados produzido é muito alto, portanto, a grande maioria das proteínas só podem ser anotadas computacionalmente (Radivojac et al., 2013). Tomemos como exemplo o Uniprot (Bateman et al., 2020), uma base de dados onde estão catalogados todas as sequências de proteínas conhecidas, possuindo cerca de 190 milhões de registros, onde somente meio milhão desses registros foram devidamente anotados e revisados. Outra base de dados, o Pfam (Mistry et al., 2021), que possui em seu catálogo famílias e domínios de proteínas, em sua versão 33.1 continha 23% (4244) de todos os seus registros classificados como domínios de função desconhecida (DUF - *domains of unknown function*) ou família de proteína não caracterizada (UFP - *uncharacterized protein families*). Isso mostra a lacuna de informações ainda existente nos dados de proteínas, onde necessita-se de esforços contínuos através de novos estudos e métodos de caracterização para enriquecimento das bases de dados.

Um pré-requisito para a elucidação da função biológica de uma proteína é o entendimento dos mecanismos responsáveis pela sua interação com o ligante, onde é crucial a descrição, caracterização e quantificação dos tipos de interações para a formação do complexo. O entendimento detalhado da interação proteína-ligante, em nível molecular, e seus mecanismos de reconhecimento e ligação, facilitam o planejamento racional de fármacos, assim como o desenvolvimento e a descoberta de novos procedimentos terapêuticos (Govindaraj et al., 2018).

O local da proteína onde os ligantes acoplam-se para formar um complexo é conhecido como sítio de ligação. Os sítios são regiões específicas, geralmente caracterizados por cavidades, oriundas do processo de enovelamento que é dirigido e estabilizado por forças termodinâmicas (Balchin et al., 2016). O sítio de ligação, assim como seu respectivo ligante, oferecem pistas sobre a função biológica da macromolécula proteica, sendo que essa relação estrutura-função tem sido um importante objeto de estudo da biologia estrutural. No entanto, encontrar a localização dessas cavidades na superfície proteica não é uma tarefa trivial. Geralmente a determinação experimental dos padrões de interação entre proteínas e seus ligantes, assim como seus respectivos sítios de ligação, via procedimentos *in vitro* ou *in vivo* é extremamente custoso, tanto em tempo quanto em recursos de laboratório (Roche et al., 2015).

Como citado anteriormente, existe uma lacuna de informação nas bases de dados biológicos quando comparamos o número de sequências catalogadas em relação ao número de registros devidamente anotados. Além disso, existe uma discrepância entre a quantidade de sequências proteicas conhecidas e o número de estruturas determinadas experimentalmente, principalmente devido às limitações intrínsecas às técnicas experimentais. Por exemplo, o PDB (Protein Data Bank) (Berman et al., 2000), uma base de dados de estruturas resolvidas experimentalmente, possui um pouco mais de 180 mil estruturas catalogadas (acesso em julho de 2021), frente às 150 milhões de sequências presentes no UNIPROT. Devido ao número restrito de estruturas experimentalmente determinadas o uso de abordagens *in silico* capazes de criar modelos estruturais aproximados tornou-se necessário. Na ausência de uma estrutura experimental, abordagens teóricas como modelagem comparativa por homologia, *threading*, e predição *ab initio*, são capazes de construir modelos que podem ser usados na investigação de eventos da interação proteína-ligante (Govindaraj et al., 2018).

Esses desafios enfrentados na determinação de sítios de ligação, tanto na ausência de ligantes em estruturas experimentais, quanto em relação ao desconhecimento dos sítios em estruturas oriundas da modelagem *in silico*, podem ser enxergados como oportunidades para o desenvolvimento de estratégias computacionais. A demanda por ferramentas eficientes e escaláveis que contribuam na construção de evidências neces-

sárias à compreensão do reconhecimento molecular é crescente. Desse modo, métodos e algoritmos de inteligência artificial podem ser utilizados para mitigar os custos, preencher lacunas deixadas por questões experimentais, além de extrair conhecimento útil e não trivial de bases de dados biomoleculares.

Como definido em Gallo Cassarino et al. (2014), a predição de um sítio de ligação consiste em estimar resíduos de aminoácidos de uma proteína com potencial de interagir com algum ligante biologicamente relevante. Métodos tradicionais de predição de sítios de ligação utilizam cálculos custosos como, por exemplo, o alinhamento múltiplo de estruturas, cálculos de funções de campo de força e cálculos de geometria computacional, que são procedimentos inviáveis num processamento em larga escala. Para aplicações em cenários reais, essas técnicas devem ser escaláveis. Além disso, algumas estratégias preditivas possuem limitações, como a incapacidade de processar proteínas com múltiplas cadeias. Adicionalmente, no contexto da aprendizagem de máquina, algumas estratégias trabalham com modelos preditivos que se apresentam em forma de caixa preta, deixando a desejar na interpretabilidade dos dados.

Essa tese propõe o GRaSP (Graph-based Residue neighborhood Strategy to Predict binding sites), uma estratégia para predição de sítios de ligação em proteínas. A partir da estrutura tridimensional da proteína, o GRaSP combina modelagem em grafos com técnicas de aprendizagem supervisionada para identificar resíduos de aminoácidos que pertencem a sítios de ligação em proteínas. GRaSP modela os átomos da proteína como nós do grafo, enquanto as interações intermoleculares realizadas entre esses átomos são denotadas pelas arestas do grafo. Como forma de tornar a modelagem compreensível aos algoritmos de aprendizagem de máquina, esses grafos passam por uma transformação, onde cada resíduo da proteína é convertido em um vetor de características oriundo da sua respectiva modelagem em grafo. Tais vetores de características alimentam o modelo preditivo, que irá responder se um determinado resíduo pertence ou não a um sítio de ligação. A partir dos experimentos realizados com diversas bases de dados, a estratégia mostrou-se escalável, robusta e com desempenho preditivo comparável ou superior aos métodos do estado da arte.

1.1 Objetivos

O objetivo deste trabalho é projetar, implementar e avaliar uma estratégia para caracterizar e prever potenciais resíduos do sítio de ligação em proteínas. Os objetivos específicos são:

- Modelar os resíduos da proteína e sua vizinhança como grafos no nível atômico,

rotulando seus nós e arestas com as propriedades físico-químicas dos átomos e interações.

- Modelar o problema de prever resíduos de sítio de ligação como um problema binário de classificação, definindo quais são os dados de entrada, saída e classe.
- Construir um repositório de dados contendo proteínas com sítio de ligação conhecidos para que possam ser usadas como exemplos no modelo preditivo;
- Coletar diversos conjuntos de dados de proteínas que serão usados em experimentos comparativos com outros métodos do estado da arte;
- Comparar a estratégia proposta com outros métodos a partir de protocolos e métricas consolidadas na literatura.
- Disponibilizar publicamente o código fonte da estratégia desenvolvida, assim como as bases de dados utilizadas.
- Desenvolver um servidor web para hospedar a estratégia, disponibilizando os resultados de forma visual e inteligível.

1.2 Organização do Texto

No Capítulo 2, Revisão da Literatura, foi feito um estudo sobre os trabalhos correlatos para predição de sítios de ligação. O Capítulo 3, Métodos, descreve cada uma das etapas, assim como seus conceitos teóricos, necessários para a construção da estratégia. Os resultados são apresentados no Capítulo 4, através de diversos experimentos comparativos. Finalmente, no Capítulo 5, estão as considerações finais a respeito do trabalho desenvolvido.

Capítulo 2

Revisão da Literatura

Neste capítulo são apresentadas as diferentes abordagens relacionadas à predição de sítios de ligação em proteínas. Serão descritas algumas técnicas relevantes, que foram frequentemente citadas durante a etapa de revisão da literatura, pontuando suas principais características, aspectos positivos e negativos.

2.1 Métodos preditivos para sítio de ligação

Identificar sítios de ligação em proteínas é uma tarefa chave para caracterizar a função proteica, e por esse motivo, diversas iniciativas como o Critical Assessment of Protein Structure Prediction (CASP) (Gallo Cassarino et al., 2014), o Continuous Automated Model Evaluation (CAMEO) (Haas et al., 2018), e o Critical Assessment of Function Annotation (CAFA) (Radivojac et al., 2013) fomentaram ou vêm fomentando o desenvolvimento de métodos para predição de sítios de ligação e anotação de função para proteínas nos últimos anos. Por meio desses projetos, as técnicas preditivas para sítios de ligação puderam se desenvolver a partir da padronização de métodos de avaliação, assim como a unificação de conceitos e definições pertinentes a esta área de pesquisa (Zhao et al., 2020).

Os métodos preditivos para sítio de ligação podem ser classificados em três categorias segundo Roche et al. (2015): baseados na sequência, baseados na estrutura e métodos híbridos. Os métodos baseados na sequência exploram a informação contida na conservação evolutiva das sequências de proteínas similares. Já os métodos baseados na estrutura aproveitam-se da informação contida no arranjo tridimensional dos átomos da proteínas para predizer regiões de sítio de ligação. Os métodos híbridos tentam extrair vantagens de ambas abordagens anteriores através da combinação de seus resultados.

Métodos baseados em sequência levam vantagem em relação aos que são baseados em estrutura sob a ótica do volume de dados disponível, especialmente em casos onde a proteína de interesse não possui um modelo tridimensional ou estrutura experimentalmente determinada. No entanto, os trabalhos da literatura que alcançam melhores resultados preditivos são em sua maioria de métodos estruturais (Gallo Cassarino et al., 2014). Vale ressaltar que, apesar da discrepância no volume de dados disponíveis, nos últimos anos, os dados estruturais também têm seguido uma tendência de crescimento, e a existência de repositórios de dados públicos como o PDB (Berman et al., 2000) e o Biolip (Yang et al., 2012) facilitam o acesso dos pesquisadores à informação usada para experimentação, avaliação e validação de suas técnicas.

Nesta tese, a descrição dos métodos preditivos terá um foco maior nas abordagens que se baseiam na estrutura, já que a estratégia proposta compartilha das mesmas características. Para uma informação mais detalhada sobre o métodos baseados em sequência recomenda-se os trabalhos (Macari et al., 2019; Roche et al., 2015).

2.1.1 Métodos baseados em estrutura

Os métodos de predição de sítio baseados em estrutura trabalham com dados oriundos do arranjo tridimensional dos átomos da estrutura proteica. A ideia é combinar padrões de distância espacial com propriedades geométricas e topológicas, assim como características físico-químicas dos átomos, para buscar por regiões candidatas a serem sítios de ligação. Como a informação estrutural de uma proteína é mais conservada do que sua sequência, ou seja, sequências com pouca identidade podem se enovelar em estruturas parecidas, é possível buscar por sítios de ligação em proteínas mesmo nos casos onde não seja factível a inferência funcional pela identidade de sequência (Izidoro et al., 2015). Pela forma que esses métodos apresentam os resultados preditivos, podemos distingui-los em dois grupos: (i) métodos centrados na cavidade e (ii) métodos centrados no resíduo.

O primeiro grupo, métodos centrados na cavidade, buscam por regiões na superfície da proteína, geralmente cavidades, classificando-as a partir de uma pontuação específica do método. Geralmente essas regiões são representadas como um conjunto de pontos, ou um centroide desses pontos no espaço tridimensional, de localidades próximas à superfície proteica. Fpocket (Le Guilloux et al., 2009), por exemplo, é um método que explora a geometria da superfície da proteína na busca por cavidades. Apesar de conseguir realizar a busca de forma rápida, Fpocket retorna um número grande de resultados para uma mesma proteína e, apesar de encontrar sítios de ligação dentro do conjunto de resultados, nem sempre consegue classificá-las no topo da lista,

gerando muitos falsos positivos (Krivák & Hoksza, 2015).

O SITEHOUND (Gherzi & Sanchez, 2009), é um método clássico baseado em energia, como esquematizado na Figura 2.1.1. Inicialmente a proteína é coberta por uma grade (Figura 2.1.1(a)), então sondas de carbono e fosfato são espalhadas pelos pontos da grade de forma que as forças de interação entre essas moléculas e a proteína são calculadas em cada ponto (Figura 2.1.1(b)). Os pontos de grade com maior energia de interação são extraídos e agrupados para formar uma cavidade em potencial (Figura 2.1.1(c)). Uma desvantagem dos métodos que exploram a estrutura tridimensional, como os baseados em geometria e energia, é que eles dependem fortemente do estado conformacional em que a proteína se encontra.

Como forma de ilustrar a desvantagem dos métodos geométricos podemos exemplificar alguns dos modelos propostos para explicar os mecanismos da interação proteína-ligante, como, por exemplo, o modelo chave-fechadura (Figura 2.2 (a)) e o modelo de encaixe induzido (Figura 2.2 (b)). Enquanto primeiro considera que a proteína e o ligante são rígidos e com interfaces de ligação perfeitamente complementares, o que seria o cenário ideal para abordagens geométricas, o segundo assume que o sítio de ligação da proteína é flexível e a interação com o ligante acaba induzindo mudanças conformacionais no sítio de ligação. Sendo assim, algumas cavidades podem ser induzidas pela interação proteína-ligante quando a proteína se encontra no estado acoplado ao ligante, e não podem ser encontradas por métodos sensíveis à conformação quando a proteína é processada no estado desacoplado, onde esta cavidade não existe Zhao et al. (2020).

Jiménez et al. (2017), por sua vez, apresentou o DeepSite, uma técnica baseada em redes neurais convolucionais que trata a estrutura da proteína como uma imagem tridimensional (3D), inspirado pela visão computacional. As proteínas são discreti-

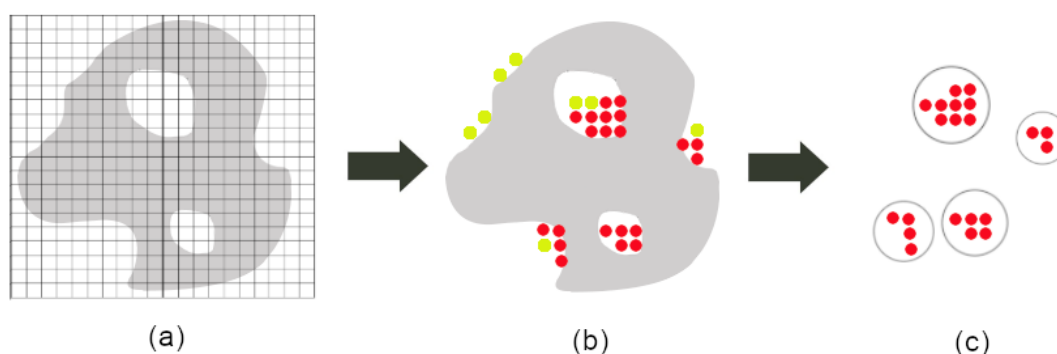


Figura 2.1. Esquema das etapas seguidas por métodos energéticos para buscar sítios de ligação em proteínas.

zadas em *voxels*, onde cada um deles é composto por um conjunto de propriedades físico-químicas em nível atômico. A palavra *voxel* origina-se da combinação entre as palavras "volume" e "pixel", e, em contraste ao pixel que é o menor componente em uma imagem bidimensional, o *voxel* representa um valor em uma grade no espaço tridimensional. Finalmente, subconjuntos de *voxels* são amostrados, e suas características são usadas como entrada para a rede neural convolucional. DeepSite apresentou melhor desempenho preditivo em relação a outros métodos focados em localizar cavidades, como Fpocket e Concavity (Capra et al., 2009).

Outro método, o P2Rank (Krivák & Hoksza, 2018), utiliza o algoritmo de florestas aleatórias para aprender com exemplos de sítios de ligação já conhecidos, e identificar cavidades em novas proteínas. Inicialmente, o P2Rank identifica pontos espalhados pelo solvente que estão próximos à superfície da proteína, e então calcula características a partir de propriedades físico-químicas e geométricas da vizinhança onde esses pontos estão localizados. Com base nessas características o P2Rank busca por regiões com alta concentração de pontos com potencial de "ligabilidade", classificando-as como sítios. Em seus experimentos, P2Rank alcançou desempenho preditivo superior a métodos como Fpocket, Sitehound (Gherzi & Sanchez, 2009), MetaPocket 2.0 (Zhang et al., 2011) e o DeepSite. A Figura 2.3 apresenta a tela principal do PrankWeb (Jendele et al., 2019), uma aplicação web que provê interface para o método preditivo P2Rank.

A outra classe de métodos baseados na estrutura é aquela que compreende estratégias que são centradas no resíduo, ou seja, usam os resíduos de aminoácidos da proteína como objeto de predição. Em contraste aos métodos centrados na cavidade, não há uma lista de classificação, informação de pontos no espaço, ou a forma geométrica da cavidade. Na perspectiva dos métodos centrados no resíduo, o problema de predição de sítios é enxergado como uma classificação binária de cada um dos resíduos de aminoácidos da proteína, se este pertence ou não a um sítio de ligação.

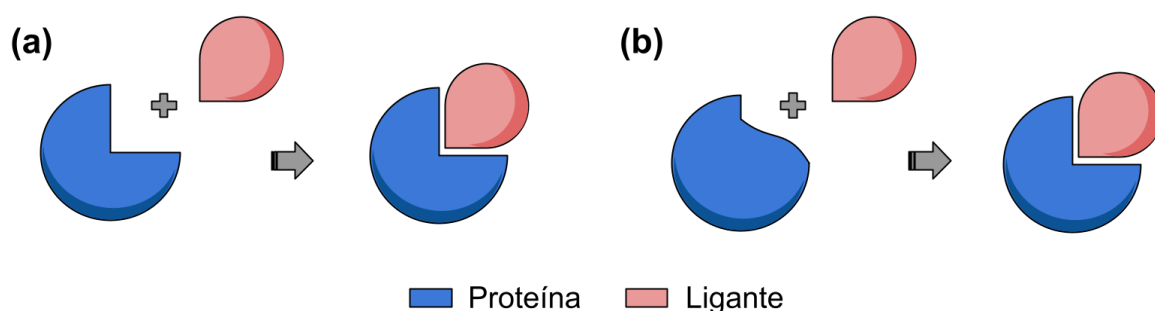
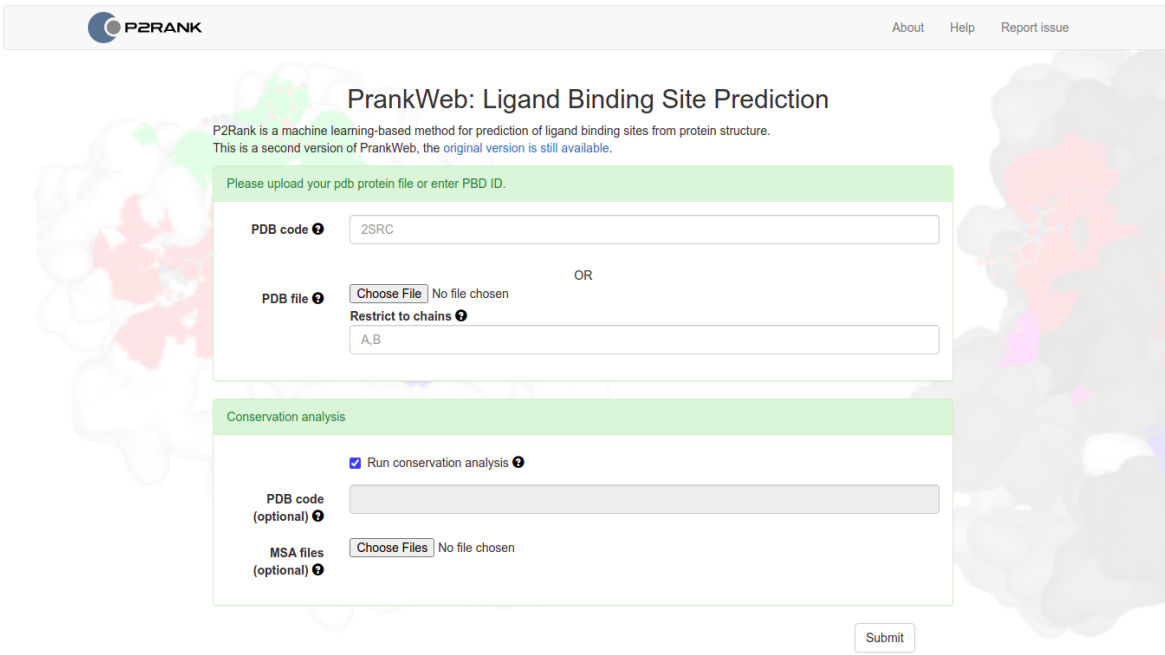


Figura 2.2. Modelos dos mecanismos de interação proteína-ligante, (a) chave-fechadura e (b) encaixe induzido

O GASS (Izidoro et al., 2015; Moraes et al., 2017), por exemplo, é um algoritmo genético que procura por sítios de catalíticos em enzimas a partir de um repositório estrutural de proteínas com sítios já conhecidos. O método envolve gerar populações de sítios catalíticos e simular efeitos evolutivos como mutação e cruzamento nessas populações, com o intuito de selecionar sítios candidatos a partir de uma função de avaliação similar ao RMSD, aplicada entre os exemplos e os sítios candidatos encontrados pelo GASS. Esse processo é repetido por um número específico de gerações até que uma condição de parada seja alcançada.

Vale ressaltar que o sítio catalítico compreende os resíduos diretamente responsáveis pela reação enzimática, enquanto que, por definição, o sítio de ligação compreende resíduos que fazem contato com o ligante de alguma forma. Como definido por Gallo Cassarino et al. (2014), se o resíduo está a uma distância específica do ligante, ele é considerado como resíduo do sítio de ligação, mesmo que não tenha participação direta na reação catalítica. Dessa forma, os resultados encontrados pelo GASS se restringem a um conjunto menor de resíduos de aminoácidos em comparação às predições feitas por outros métodos.

Também existem métodos híbridos, como o COACH (Yang et al., 2013), que combina resultados de diversos algoritmos preditivos, dentre eles: TM-SITE, S-SITE, COFACTOR (Roy et al., 2012), FINDSITE (Brylinski & Skolnick, 2008) e ConCavity (Capra et al., 2009). Os dois primeiros métodos são oriundos do próprio COACH,



The screenshot shows the PrankWeb interface. At the top left is the P2RANK logo. At the top right are links for 'About', 'Help', and 'Report issue'. The main heading is 'PrankWeb: Ligand Binding Site Prediction'. Below this is a brief description: 'P2Rank is a machine learning-based method for prediction of ligand binding sites from protein structure. This is a second version of PrankWeb, the original version is still available.' The main form area is titled 'Please upload your pdb protein file or enter PDB ID.' It contains two input fields: 'PDB code' with the value '2SRC' and 'PDB file' with a 'Choose File' button and the text 'No file chosen'. Below these is a section for 'Restrict to chains' with a text input field containing 'A,B'. A 'Submit' button is located at the bottom right of the form. Below the main form is a 'Conservation analysis' section with a checked checkbox 'Run conservation analysis', an optional 'PDB code' input field, and an optional 'MSA files' input field with a 'Choose Files' button and the text 'No file chosen'. A 'Submit' button is also present at the bottom right of this section.

Figura 2.3. Tela principal do PrankWeb.

enquanto os demais são métodos de terceiros. COACH introduz dois algoritmos: TM-SITE, que é baseado em estrutura e encontra resíduos de aminoácidos de sítio de ligação usando um repositório de exemplos estruturais; e o S-SITE, que é um algoritmo baseado na sequência, usado na busca por sítios de ligação através de comparações entre sequências. COACH foi considerado o método estado da arte para predição de resíduos de sítio de ligação nos trabalhos de Liu et al. (2020), superando outros métodos em um experimento usando uma base de dados de 500 proteínas não redundantes. Uma versão aprimorada do método, denominada COACH-D (Figura 2.1.1), foi proposta recentemente por Wu et al. (2018), em forma de ferramenta web, visando propor ligantes para os sítios encontrados e, posteriormente, aplicando o AutoDock Vina para ancorar e refinar as poses do ligante no sítio encontrado.

Apesar de ser o método com melhor desempenho preditivo dentre os demais, o COACH possui algumas limitações. Primeiramente, a ferramenta só consegue processar cadeias simples de proteína. Em termos de usabilidade, isso é um problema, já que é incômodo para o usuário ter que realizar um pré-processamento para particionar a proteína em cadeias antes de usar a ferramenta. Além disso, é provável que o método não consiga encontrar resíduos de sítio provenientes de regiões onde as cadeias fazem interseção, já que estas não são consideradas conjuntamente. Outra desvantagem é o tempo necessário para realizar as predições. O servidor do COACH leva, em média, de 2 a 4 horas para processar uma proteína de aproximadamente 300 resíduos. É um tempo muito alto para quem busca por métodos escaláveis, capazes de processar grandes quantidades de dados.

Ao estudar os métodos preditivos que foram descritos, notou-se a necessidade de

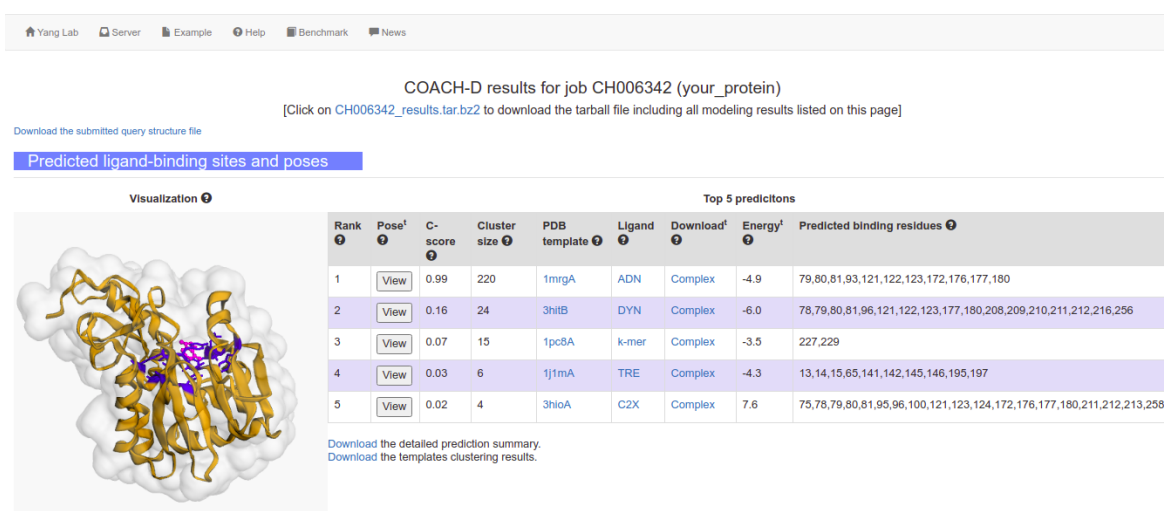


Figura 2.4. Página de resultados preditivos ferramenta web COACH-D.

desenvolver um novo método que buscasse cobrir algumas das lacunas deixadas pelas abordagens existentes. Então, neste trabalho, propõe-se o GRaSP. Uma estratégia para prever resíduos de sítio de ligação que é capaz de processar diferentes tipos de proteínas, inclusive de múltiplas cadeias, que é escalável, interpretável, e apresenta desempenho preditivo compatível ou superior aos métodos existentes.

Capítulo 3

Métodos

Esta tese propõe o GRaSP, uma estratégia para prever resíduos de aminoácidos pertencentes ao sítio de ligação em proteínas, usando modelagem em grafos e aprendizagem supervisionada. Nesta seção, serão descritas as etapas percorridas para construção da estratégia, estas que contemplam o processo de modelagem dos grafos, as bases de dados utilizadas para experimentos e os protocolos e métricas empregados na avaliação dos resultados.

3.1 Construindo grafos de vizinhança

Dados são observações de fenômenos do mundo real. Os algoritmos de aprendizagem de máquina alimentam-se desses dados, almejando construir modelos preditivos que melhor descrevam os relacionamentos entre as diferentes características dos mesmos. Entretanto, na maioria dos casos, os dados brutos não são interpretáveis pelos algoritmos, e quando são, a forma como estão anotados não captura os padrões necessários para gerar um modelo preditivo eficaz.

Os dados estruturais de proteínas contemplam um conjunto de arquivos contendo coordenadas espaciais dos átomos. Esse tipo de dado não é suficientemente informativo para gerar previsões de qualidade quando se trata de encontrar sítios de ligação em proteínas. Em casos como esse é necessário o uso de técnicas que fazem parte de um ramo da análise de dados denominado engenharia de características (*feature engineering*). Uma característica é a representação numérica de um dado bruto. Um dos objetivos deste trabalho, compreende a modelagem de características mais apropriadas para alimentar os algoritmos preditivos, levando em consideração os dados disponíveis, o tipo de algoritmo de classificação utilizado, e a tarefa na qual o algoritmo será aplicado (Zheng & Casari, 2018).

Como dito anteriormente, a estrutura de uma proteína é crucial para determinar sua função. A estabilização estrutural da macromolécula proteica provém de fatores entrópicos em conjunto com as interações não-covalentes que, além de determinar o arranjo tridimensional dos átomos da proteína, também são responsáveis por mediar a interação proteína-ligante. Através de forças atrativas e repulsivas as interações não-covalentes orientam o processo de enovelamento, onde resíduos de aminoácidos da proteína se organizam em uma estrutura tridimensional compacta para cumprir sua função biológica (Balchin et al., 2016; Dobson, 2003). Um exemplo clássico são os padrões formados pelas pontes de hidrogênio na formação de estruturas secundárias da proteína, como a alfa-hélice ilustrada pela Figura 3.1. Dessa forma um resíduo não pode ser observado de forma isolada, já que esse está sofrendo influência constante dos demais resíduos que o cercam. Cada resíduo é influenciado pela sua vizinhança através de uma rede de interações interatômicas, assumindo o comportamento de um sistema complexo.

A ideia fundamental então é enxergar a proteína pela perspectiva de um sistema complexo, ou seja, considerar também as interações que os átomos dos resíduos fazem entre si, e não somente as coordenadas espaciais dos mesmos. Em um nível abstrato, os átomos dos resíduos de uma proteína podem ser considerados como uma série de nós que são conectados entre si por arestas, onde cada aresta representa uma interação não-covalente entre um par de átomos. Estes conjuntos de nós e arestas formam a estrutura denominada grafo (Barabasi & Oltvai, 2004). Os grafos são exhaustivamente usados para modelar objetos e suas interações. Diversos trabalhos promissores têm

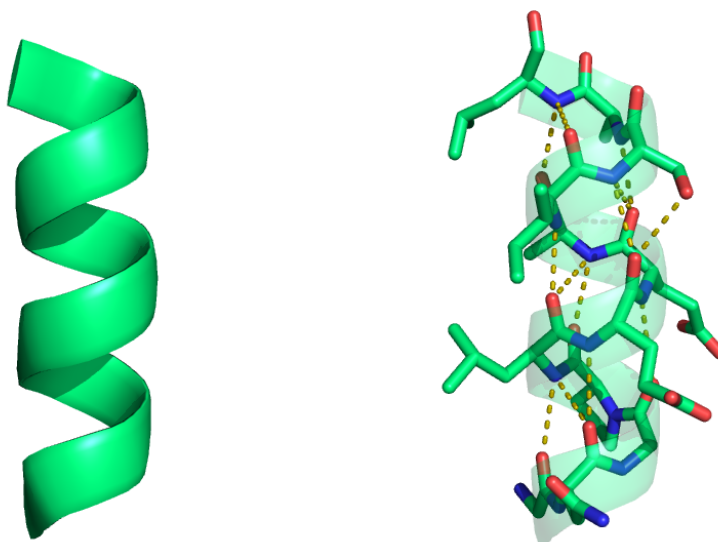


Figura 3.1. Interações do tipo ponte de hidrogênio (tracejadas em amarelo) estabelecidas para formar uma alfa-hélice.

explorado esse tipo de modelagem, onde os átomos da proteína representam os nós, e as interações entre os átomos correspondem às arestas (Fassio et al., 2017; Medina et al., 2017; Monteiro et al., 2018; Campelo et al., 2019; Santana et al., 2016; Ribeiro et al., 2020).

A estratégia proposta neste trabalho usufrui do poder descritivo e simplificado dos grafos para interpretar o problema de predição do sítios de ligação em proteínas. Dado a estrutura tridimensional de uma proteína, cada um dos seus resíduos de aminoácidos, assim como a vizinhança destes resíduos, são representados como um grafo. A representação da vizinhança em forma de grafo permite que a estratégia considere, simultaneamente, vários aspectos do ambiente que envolve o resíduo, capturando particularidades dos relacionamentos em nível molecular que são perdidas em abordagens locais, quando o resíduo é considerado individualmente.

A modelagem das interações entre o resíduo e sua vizinhança em um grafo é construída em nível atômico, onde cada nó do grafo representa um átomo, e cada aresta denota a existência de uma interação não-covalente entre o par de átomos de resíduos vizinhos. Para cada resíduo de uma proteína são anotadas as propriedades físico-químicas de seus átomos constituintes, podendo ser classificados como hidrofóbico, aromático, acceptor, doador, positivo e negativo, de acordo com os trabalhos (Gonçalves-Almeida et al., 2012; Fassio et al., 2017, 2019). A informação completa sobre as propriedades atribuídas a cada tipo de átomo está descrita no Apêndice B1. As interações entre os átomos dos resíduos vizinhos são construídas a partir das propriedades anotadas no passo anterior, observando critérios de distância euclidiana entre os pares de átomos, também de acordo com (Gonçalves-Almeida et al., 2012; Fassio et al., 2017, 2019), como pode ser observado na Tabela 3.1. A topologia do grafo é construída usando uma estrutura de dados chamada de árvore k -dimensional e que está implementada no Biopython (Yang et al., 2012). A árvore k -d, como é abreviada, organiza pontos no espaço de forma que seja possível calcular os contatos dentro de um raio arbitrário sem a necessidade de considerar todos os átomos da proteína. A complexidade de busca na estrutura é de $O(n^{\frac{2}{3}} + m)$, em comparação com a busca exaustiva por contatos de $O(n^2)$, onde n é o número de átomos na proteína, e m denota o número de átomos reportados na busca dentro do raio escolhido (Van Kreveld et al., 2000).

A vizinhança de um resíduo é dividida em dois subconjuntos disjuntos: a primeira camada e a segunda camada. A Figura 3.2 mostra uma representação esquemática de uma proteína, onde os círculos representam seus resíduos. Seja um resíduo específico ρ , nesse caso o resíduo central em cor preta na Figura 3.2 (a), a primeira camada, em tom cinza escuro, é definida pelo subconjunto de resíduos que realizam interações não-covalentes com ρ . Já a segunda camada, em tom cinza claro, é composta pelo sub-

conjunto de resíduos que interagem diretamente com os elementos da primeira camada, através de interações não-covalentes. O grafo de vizinhança para o resíduo ρ é então construído com base nestas duas camadas de resíduos vizinhos, como esquematizado na Figura 3.2 (b).

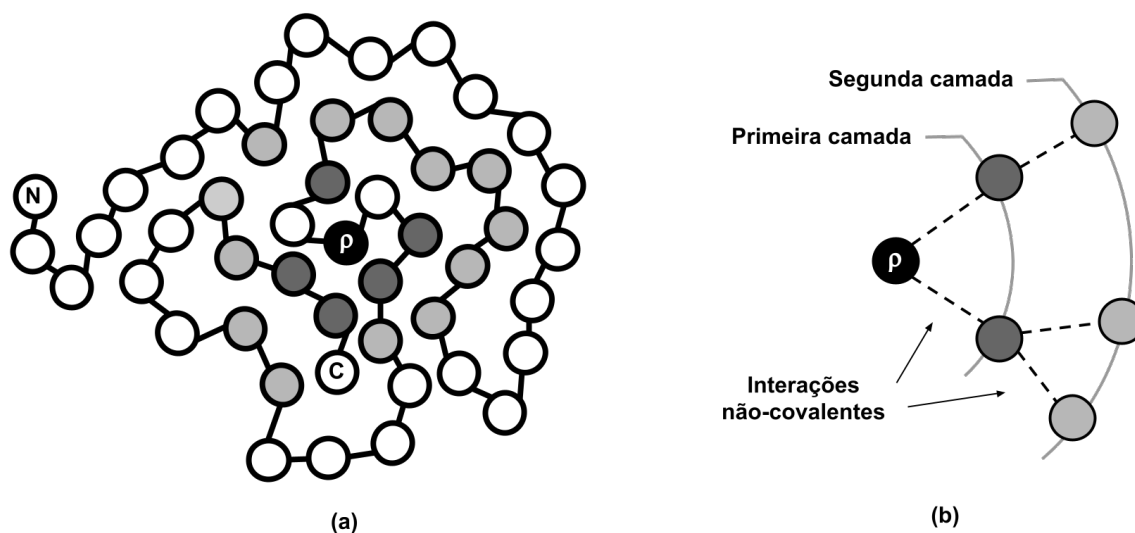


Figura 3.2. Esquema um resíduo ρ e sua vizinhança modelados como um grafo em uma proteína.

Em outras palavras, cada resíduo de aminoácido é modelado como um grafo, que é uma abstração das interações presentes no ambiente local do respectivo resíduo. Logo, uma proteína P é representada pelo seu conjunto de resíduos constituintes, isto é, formada pelos grafos de vizinhança de cada resíduo $\rho \in P$. Com esta modelagem os dados brutos de coordenadas espaciais das estruturas de proteínas são transformados em grafos de interações, enriquecido com propriedades em seus vértices e arestas, para a subsequente tarefa de predição.

Tabela 3.1. Critérios de distância (em Å) e propriedades físico-químicas dos átomos envolvidos em cada tipo de interação.

Tipo de interação	Tipo do átomo	Distância	Distância
		Mínima	Máxima
Empilhamento aromático	dois átomos aromáticos	1.5	3.5
Interação de hidrogênio	um átomo acceptor e um doador	2.0	3.0
Interação hidrofóbica	dois átomos hidrofóbicos	2.0	3.8
Repulsivo	dois átomos com a mesma carga	2.0	6.0
Ponte salina	dois átomos com cargas opostas	2.0	6.0

3.1.1 Construindo vetores de características

Como forma de tornar o conjunto de grafos inteligível aos algoritmos de aprendizagem supervisionada, cada grafo foi codificado como um vetor de características. Os vetores podem ser compilados em uma matriz, que por sua vez, pode ser usada como conjunto de treinamento para o modelo de classificação. Um vetor de características resume as propriedades físico-químicas e topológicas presentes no grafo de vizinhança através de uma contagem simples. São consideradas as seguintes características para construção do vetor:

- **Exposição do resíduo ao solvente (2 características):** através da métrica Half-Sphere Exposure (HSE) desenvolvida por Hamelryck (2005) e implementada na biblioteca Biopython (Yang et al., 2012) do Python. O HSE consiste em definir uma esfera de raio arbitrário em torno do carbono alfa do resíduo. Então a esfera é cortada por um plano que a separa em dois hemisférios: um na orientação da cadeia lateral do resíduo, e o segundo contemplando o lado oposto. A densidade de outros resíduos que cercam o carbono alfa é computada para cada hemisfério, resultando em uma dupla de valores (HESu, HSEd) que correspondem ao hemisfério orientado para a cadeia lateral (up), e o hemisfério oposto (down).
- **Propriedades atômicas (6 características):** onde são computados os tipos de átomos contidos no resíduo a partir das propriedades: aromático, acceptor, doador, positivo, negativo e hidrofóbico.
- **Interações com a primeira camada (5 características):** onde são computados os tipos de interação que o resíduo realiza com sua primeira camada de vizinhos. As interações possíveis são: ponte de hidrogênio, ponte salina, contato hidrofóbico, repulsiva ou empilhamento aromático.
- **Propriedades atômicas + exposição do resíduo ao solvente da primeira camada (8 características):** onde são computados a média tanto do grau de exposição (HESu, HSEd), quanto dos tipos de átomos contidos nos resíduos que fazem parte da primeira camada de vizinhança.
- **Propriedades atômicas + exposição do resíduo ao solvente da segunda camada (8 características):** o mesmo do item anterior, porém computado para os resíduos presentes na segunda camada.

base de dados contém informação referente aos resíduos de sítio de ligação de todos os complexos catalogados. Esse cenário interessante é interessante para a construção de uma base de exemplos, já que, dotado de um repositório de dados devidamente curado, e com informação sobre o sítio de ligação, nos resta a tarefa de codificar esse repositório em grafos de vizinhança, e assim foi feito.

A princípio foi necessário realizar um filtro nas informações disponíveis. O GRaSP foi projetado com o intuito de prever sítios de ligação para compostos orgânicos não proteicos. A partir dessa definição foram removidas as informações referentes aos sítios de íons metálicos, sítios de peptídeos e sítios de ácidos nucleicos, restando 53.278 entradas para compor o repositório final. Cada complexo passou pela modelagem de grafos de vizinhança e tiveram suas respectivas matrizes de características geradas. Esse repositório de matrizes será a fonte de alimentação para treinamento do modelo preditivo. O objetivo inicial foi usar todos os dados disponíveis para treinar um conjunto de classificadores estáticos. Entretanto o processo de classificar um resíduo de sítio de ligação é bastante complexo, onde até os melhores métodos de predição alcançam resultados preditivos medianos, com pouca expressividade na magnitude das métricas, que serão abordadas com mais detalhe na Seção 3.5. A partir de experimentos iniciais observou-se a existência de uma heterogeneidade nos dados que dificultava obter bons resultados.

Como forma de contornar essa característica heterogênea, passou-se a implementar modelos dinâmicos, construídos por demanda, a partir de uma amostragem do repositório em função da proteína de interesse. Como esquematizado na Figura 3.4, cada vez que uma proteína é submetida ao processo de predição dos seus resíduos de sítio, uma amostragem do repositório de complexos proteicos é realizada usando o BLAST. As sequências que apresentarem maior identidade são selecionadas, e somente esse subconjunto dos dados será utilizado para alimentar a construção do modelo preditivo, exclusivamente para a proteína submetida. Esse processo torna o conjunto de treinamento mais homogêneo, facilitando o trabalho do algoritmo de aprendizagem em detectar os padrões responsáveis por caracterizar o sítio de ligação.

3.3 Aprendizagem supervisionada

Dado que o conjunto de treinamento foi definido e a matriz de resíduos contendo as características foi construída, o próximo passo é usar esse conjunto de dados como entrada para o algoritmo de aprendizagem supervisionada. O método de classificação usado para compor a estratégia foi o *Extra-tree* (Extremely Randomized Trees) (Geurts



Figura 3.4. Esquema de amostragem de exemplos do repositório de proteínas para construção do modelo preditivo do GRaSP.

et al., 2006), implementado no *scikit-learn 0.20.2*, uma biblioteca de aprendizado de máquina para Python (Pedregosa et al., 2011). Esse algoritmo pertence ao grupo de classificadores *ensemble*, isto é, modelos que realizam a tarefa de classificação ao agregar as respostas de diversos classificadores e obter os resultados através do voto majoritário. O classificador base usado em um *Extra-tree* é a árvore de decisão.

A árvore de decisão é um método de aprendizado não-paramétrico, definido por uma estrutura hierárquica, criada para organizar um conjunto de perguntas e respostas a respeito dos atributos das instâncias de dados, com o intuito de classificá-las (Tan et al., 2016). O processo de classificação usando árvores de decisão se dá em duas etapas. Inicialmente o conjunto de dados de entrada, denominado conjunto de treinamento, é usado para construção do modelo preditivo. Num segundo momento, esse modelo é usado para classificar um conjunto novas instâncias de dados, denominado conjunto teste, que não foi usado na etapa de construção do modelo.

A estrutura hierárquica da árvore de decisão é formada pelo nó raiz, aquele que possui mais alta hierarquia; por nós internos, que denotam os testes condicionais em relação aos atributos; e os nós folha, que são associados a uma das classes. Na Figura 3.5 é mostrado uma pequena base de dados de pares de átomos contendo somente dois atributos, distância em Å e tipo de carga dos átomos envolvidos, acompanhada de um possível modelo de árvore de decisão que é induzido a partir do próprio conjunto de dados e usada para discriminar as instâncias em duas classes. Para o exemplo em questão, a classe positiva denota a existência teórica de uma ponte salina entre um par de átomos, enquanto a classe negativa denota a ausência desse tipo de interação.

Para construir uma árvore de decisão é necessário dividir o conjunto de dados em partições binárias e de forma recursiva. O objetivo é separar as instâncias que possuem características diferentes, de forma que seja alcançada uma homogeneidade das classes nos nós terminais (Chen & Ishwaran, 2012). Contudo, à medida que a árvore cresce

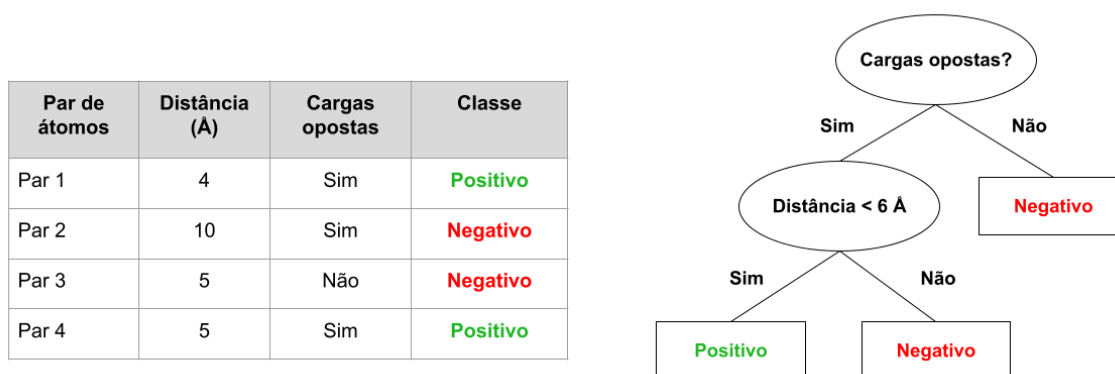


Figura 3.5. Exemplo de árvore de decisão induzida de uma base de dados para inferir a existência de uma ponte salina entre um par de átomos.

e ganha profundidade ela fica suscetível a *overfitting*, um termo usado para descrever uma forte adequação do modelo ao conjunto de treinamento, o que empobrece seu poder de generalização, isto é, um aumento no erro de classificação no conjunto de teste utilizado. Esse fenômeno está relacionado com o aumento da variância de um modelo, onde existe uma alta dispersão das predições feitas pelo classificador.

Apesar de não poder ser evitado, o *overfitting* pode ser amenizado. Para que isso aconteça, a árvore pode ser podada durante sua construção, impedindo que ela cresça até a maior profundidade possível de forma a gerar uma árvore simplificada. Entretanto, um modelo muito simples tem um erro de treinamento maior, isto é, predições equivocadas em relação ao próprio conjunto de treinamento. Esse fenômeno é conhecido como *underfitting* e está associado ao viés do modelo, denotado pela diferença entre o valor esperado e o que foi predito pelo modelo de classificação.

Suponha que o modelo de classificação seja um arqueiro tentando acertar o centro de um alvo (predição correta). Como ilustrado na Figura 3.6 (a), um classificador com viés possui um erro sistemático, deslocando suas predições do centro. Já um classificador com variância tem uma dispersão dos resultados pela superfície do alvo (Figura 3.6 (b)). O viés e a variância são dois componentes que influenciam no erro de classificação. Existe um dilema quando se quer amenizar os erros, de modo que, à medida que se tenta diminuir a variância, há um aumento no viés, e vice-versa. Algumas estratégias são capazes manipular essas componentes e obter melhoras significativas no erro de classificação, como, por exemplo, a diversificação e combinação de vários classificadores.

Um conjunto de classificadores, ou *ensemble*, é uma forma usada para obter melhorias na qualidade preditiva. A motivação de combinar vários classificadores é ate-

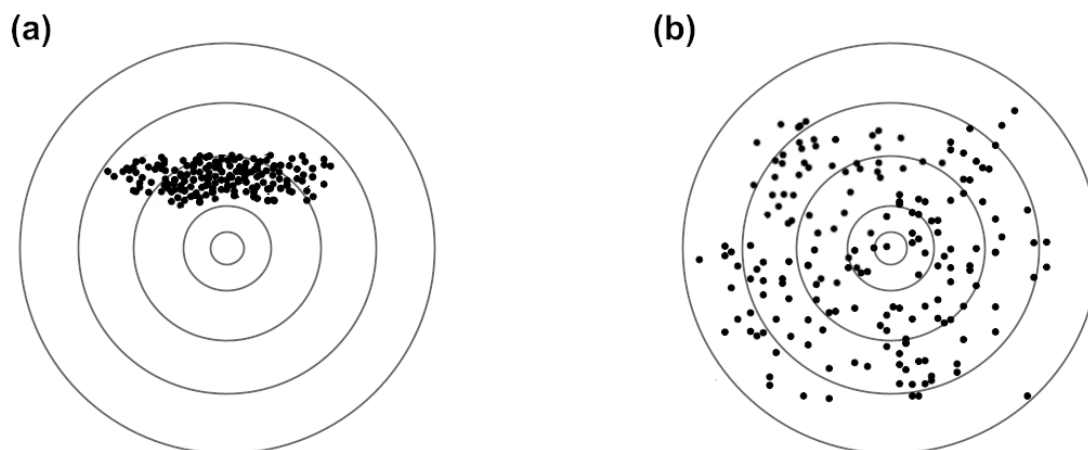


Figura 3.6. Esquema ilustrativo do dilema viés-variância. Em (a) é mostrado um esquema das predições realizadas por um classificador com viés, que apresenta um erro sistemático, deslocando suas predições do centro. Já em (b), é mostrado o esquema das predições de um classificador com variância, onde há uma dispersão nos resultados.

nuar o erro de classificação quando a predição é feita individualmente (Okun & Skarlas, 2011). Em teoria, o resultado ideal é alcançado quando o erro feito pelos classificadores são independentes, porém, na prática, assumir a independência não é uma tarefa trivial. O algoritmo de floresta aleatória (FA) (Breiman, 2001) é um exemplo de método *ensemble*. O FA funciona como uma coleção de árvores, onde cada árvore é construída usando uma amostragem, com reposição, do dado original (*bootstrap*), e a predição é feita com base no voto majoritário sobre o conjunto de árvores. O grande diferencial está na perturbação introduzida durante a construção de cada árvore. Além da amostragem do dado original, uma segunda camada de aleatorização é inserida durante o particionamento de um nó. Normalmente, os nós são particionados visitando todo o conjunto de atributos disponíveis, na busca por aquele que maximize a homogeneidade de classes dentro das partições. Diferentemente, o FA seleciona somente um subconjunto aleatório dos atributos para realizar a partição. Essa aleatorização minimiza a correlação entre as árvores construídas, implicando na diminuição da variância dos resultados no modelo global (Chen & Ishwaran, 2012).

O algoritmo de árvores extremamente aleatórias (*extra-tree*) utiliza os mesmos princípios do FA, porém vai um passo além no processo de aleatorização dos modelos. Na partição dos nós, o algoritmo *extra-tree*, incrementa uma camada de aleatorização, onde os limiares responsáveis pela regra da divisão do nó também são gerados aleatoriamente. Esse procedimento torna a construção do modelo mais eficiente computacionalmente, além de reduzir ainda mais a variância do modelo global em detrimento

de um aumento do viés nos modelos locais. Durante os experimentos, esse e outros algoritmos de classificação como o Naive Bayes, Nearest Neighbors e Support Vector Machines (Tan et al., 2016), assim como o próprio FA, foram testados para classificar resíduos de sítio de ligação, e o *extra-tree* destacou-se sendo mais rápido e apresentando resultados compatíveis com os demais métodos.

3.4 Tratando o desbalanceamento dos dados

Conjuntos de dados referentes aos resíduos de aminoácidos em uma proteína são intrinsecamente desbalanceados. A classe positiva, que é composta por resíduos de aminoácidos pertencentes aos sítios, tem uma proporção muito inferior à distribuição dos resíduos que não compõem sítios de ligação, denotados pela classe negativa. Essa configuração é desfavorável no âmbito da criação do modelo de classificação. As métricas utilizadas tanto na construção do classificador, quanto na validação do poder generalizador do mesmo, são fortemente influenciadas pela discrepância entre as classes, podendo gerar valores completamente equivocados.

Suponha que em uma proteína de 100 resíduos exista apenas um único resíduo de sítio de ligação. É fácil perceber que um modelo preditivo trivial, que classifique qualquer instância de resíduo de aminoácido como negativo, terá uma taxa de acerto de 99% para essa proteína. Devido a esse contexto, a predição correta de uma instância da classe positiva deve ter mais peso em relação à predição de uma classe negativa, de forma que a métrica não camufle eventuais erros e nos forneça uma interpretação fidedigna da qualidade do modelo preditivo.

Como forma de ajustar a distribuição das classes dentro do conjunto de dados, muitos trabalhos utilizam a estratégia de subamostragem (*undersampling*) (Jiménez et al., 2017; Shi et al., 2019; Haixiang et al., 2017). Essa técnica consiste na retirada aleatória de amostras da classe majoritária, equiparando as distribuições das classes. Considere R o conjunto de dados contendo os registros de todos os resíduos de aminoácidos, onde R^+ denota todas as instâncias positivas, e R^- todas as instâncias negativas. Na Figura 3.7 (a) é mostrado o esquema de um conjunto de dados desbalanceado, onde as instâncias positivas estão representadas por pequenos quadros cinzas. O subconjunto R^- é segmentado em k partições, $R^- = \bigcup R_i^-$, sendo $i = 1, 2, \dots, k$, e necessariamente $R_i^- \cap R_j^- \neq \emptyset \Rightarrow R_i^- = R_j^-$. O tamanho de cada partição R_i^- é igual ou próximo ao tamanho de R^+ . Em seguida, como esquematizado na Figura 3.7 (b), é construído um conjunto de novas matrizes, $E = (E_1, E_2, \dots, E_k)$, onde $E_i = R^+ \cup R_i^-$.

Cada submatriz $E_i \in E$ agora possui distribuição balanceada de suas classes, e

pode ser usada como entrada pra treinar um classificador (Figura 3.7 (c)). O algoritmo *Extra-tree* foi usado para construir k classificadores, que serão utilizados para estimar a classe dos resíduos de aminoácidos. Dada a instância de um resíduo de aminoácido qualquer como consulta, cada classificador irá prever se esse resíduo pertence a um sítio de ligação ou não. A decisão final é dada pelo voto majoritário dentre os classificadores, formando assim em espécie de combinação de *ensembles* (Figura 3.7 (d)).

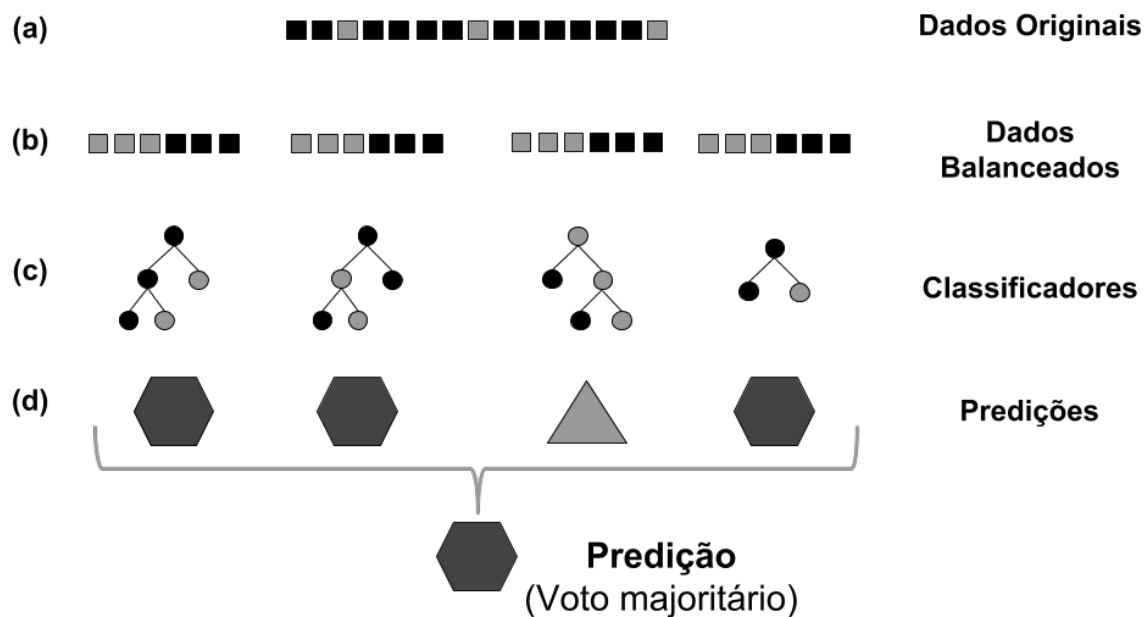


Figura 3.7. Estratégia de balanceamento e classificação mediada por voto majoritário. Os dados originais em (a) são devidamente particionados e balanceados (b), onde classificadores são inferidos em (c) e a classificação final é alcançada através do voto majoritário (d).

3.5 Métricas de avaliação

Nesta seção são detalhadas as métricas usadas para avaliar a estratégia proposta e comparará-la com outras técnicas do estado da arte.

Como forma de avaliar o desempenho de um modelo preditivo, é necessário submetê-lo ao teste de classificação usando instâncias ainda desconhecidas pelo modelo, ou seja, um conjunto de dados à parte do que foi usado na construção do modelo. Para essa finalidade existem técnicas que particionam o conjunto de dados aleatoriamente e de forma estratificada, formando dois subconjuntos disjuntos denominados conjunto de treinamento e conjunto de teste. Os dados do conjunto de treinamento são usados na construção do modelo de classificação, enquanto o conjunto de teste é

usado para avaliar a capacidade de generalização do modelo construído. Finalmente o desempenho preditivo dos classificadores é avaliado de forma quantitativa a partir do conjunto de teste.

Para entender as métricas utilizadas é preciso resumir as instâncias preditas, correta e incorretamente, pelo modelo de classificação. São usados os seguintes termos para denotar as instâncias preditas:

- **Verdadeiro positivo** (*true positive* - TP): que corresponde ao número de exemplos positivos preditos corretamente pelo classificador.
- **Falso Negativo** (*false negative* - FN): que corresponde ao número de exemplos positivos preditos erroneamente como negativos pelo classificador.
- **Falso positivo** (*false positive* -FP): que são os exemplos negativos erroneamente preditos como positivos pelo classificador.
- **Verdadeiro negativo** (*true negative* - TN): que são os exemplos negativos preditos corretamente pelo classificador.

No contexto deste trabalho, estimar corretamente um resíduo de aminoácido que pertence ao sítio de ligação é mais relevante do que a predição dos resíduos que não são de sítios de ligação. As métricas de precisão e revocação são muito usadas quando o sucesso da predição de uma classe é mais significativo do que da outra (Tan et al., 2016). A precisão (p) corresponde à fração de verdadeiros positivos no grupo de positivos preditos pelo classificador, sendo calculada pela fórmula $p = TP/(TP + FP)$. Quanto maior a precisão, menor é o número de falsos positivos. Já a revocação (r), é a fração de instâncias originalmente positivas que foram corretamente estimadas pelo modelo de classificação, ou seja, $r = TP/(TP + FN)$. O principal desafio para os algoritmos de classificação é a construção de modelos que maximizam ambos precisão e revocação (Tan et al., 2016).

A Curva Característica de Operação do Receptor, ou, do inglês, *Receiver Operating Characteristic* (ROC) é uma ferramenta gráfica usada na análise do custo benefício entre a taxa de verdadeiros positivos ($TPR = TP/(TP + FN)$), e a taxa de falsos positivos ($FPR = FP/(TN + FP)$). Na curva ROC, o TPR é plotado no eixo das ordenadas, enquanto a FPR é mostrada no eixo da abscissas. A área sob a curva ROC, ou, do inglês *Area Under Curve* (AUC), permite a comparação relativa entre classificadores. Seu valor está no intervalo entre 0 e 1, inclusive, sendo que, se um modelo de classificação é perfeito, então sua área sob a curva ROC é 1. Já um modelo que

classifica aleatoriamente tem sua AUC próxima de 0.5. Valores de AUC próximos de zero denotam predição inversa.

Uma das formas de avaliação da qualidade da predição de resíduos de sítio de ligação no experimento CASP10 (Gallo Cassarino et al., 2014), é o coeficiente de correlação de Matthews (MCC), que é dado pela equação 3.1.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (3.1)$$

O MCC varia de -1 (predição inversa) até 1 (predição perfeita), onde o valor 0 corresponde a uma predição aleatória. Essa é uma métrica robusta quando as classes são desbalanceadas. Nesse caso, há um grande número de resíduos que não são parte do sítio de ligação e um pequeno número que faz parte do sítio (Gallo Cassarino et al., 2014).

O MCC considera em seus cálculos apenas informações sobre o que foi observado e o que foi estimado, desconsiderando a natureza tridimensional da proteína no cálculo da métrica. Ao considerar o fator distância euclidiana, por exemplo, um resíduo predito como parte do sítio incorretamente, mas que está bem próximo do sítio pode ser penalizado de forma diferente de um resíduo predito como parte do sítio incorretamente e que está distante do sítio de ligação. Para contornar essa limitação do MCC, em Roche et al. (2010), foi desenvolvida a métrica *Binding-site Distance Test* (BDT), que considera em seus cálculos a distância no espaço 3D entre o sítio de ligação estimado pelo modelo preditor e o sítio de ligação observado. Essa métrica compreende o intervalo entre 0 e 1, onde preditores corretos se aproximam do valor 1, enquanto predições distantes do sítio de ligação observado têm valores próximos de zero. Para calcular o BDT é preciso obter o S-score entre o resíduo estimado i e o resíduo observado j usando a Equação 3.2, onde d_{ij} é a distância Euclidiana entre i e j , e d_0 é um limiar de distância. Em Roche et al. (2010), esse limiar é definido entre 1 e 3Å. O BDT é calculado usando a Equação 3.3, onde N_p é o número de resíduos preditos e N_0 é o número de resíduos observados.

$$S_{ij} = \frac{1}{1 + \left(\frac{d_{ij}}{d_0}\right)} \quad (3.2)$$

$$BDT = \frac{\sum_{i=1}^{N_p} \max(S_{ij})}{\max(N_p, N_0)} \quad (3.3)$$

Outra métrica denominada *Distance between the Center of the pocket and any ligand Atom* (DCA) foi usada para avaliar predições envolvendo cavidades. Ela calcula

a distância entre o centro de uma cavidade predita a os átomos do ligante acoplado ao sítio de ligação observado. O DCA considera que um sítio de ligação foi corretamente predito se a distância entre o o centro da cavidade e qualquer átomo de um ligante estiver abaixo de um limiar. Trabalhos anteriores, como em (Chen et al., 2011; Krivák & Hoksza, 2018), consideraram 4 Å como um limiar apropriado para avaliação.

3.6 Bases de Dados

Várias bases de dados foram usadas, tanto no processo de construção da estratégia, quanto nas etapas de avaliação e comparação com outros métodos do estado da arte. O intuito foi coletar conjuntos de estruturas proteicas que fossem diversas e dotadas de características que remetessem às adversidades encontradas no mundo real. As bases de dados utilizadas foram:

- **BioLip:** esse repositório de dados reportado em (Yang et al., 2012) contempla um conjunto de informações referentes às interações proteína-ligante consideradas biologicamente relevantes. Neste repositório está contido cerca de 70.000 complexos sem redundância. Por ser uma base de dados que contém a informação dos resíduos de sítios de todos os complexos catalogados, seus dados foram usados como exemplos para construir o conjunto de treinamento do modelo preditivo.
- **Base de dados COACH:** conjunto de 500 estruturas sem redundância e de cadeia simples provenientes de (Yang et al., 2013). É a base utilizada pelo próprio método COACH com a finalidade de avaliar o método.
- **Base de dados CASP10:** composto por 13 estruturas alvo, onde cada estrutura é acompanhada de 25 *templates* (Gallo Cassarino et al., 2014). Esse conjunto de dados é usado para comparar a estratégia proposta neste trabalho com outros 17 métodos participantes do CASP10.
- **B44/U44:** base de dados obtida a partir do trabalho (Krivák & Hoksza, 2018), contendo 44 proteínas em seu estado desacoplado do ligante, e mais 44 estruturas que correspondem ao estado acoplado das proteínas anteriores. O intuito de usar essa base de dados é verificar se a estratégia proposta é robusta contra flutuações conformacionais induzidas por diferentes estados da proteína, principalmente em consequência do encaixe induzido.
- **HOLO4K:** uma base com maior volume de dados, contendo 4543 complexos proteína-ligante obtidos a partir do trabalho (Krivák & Hoksza, 2018). Diferente

dos dados do COACH, estes possuem estruturas com múltiplas cadeias. O intuito aqui é verificar a capacidade de identificação de sítios entre cadeias, assim como avaliar o potencial de escala da estratégia proposta, já que é uma base de dados mais volumosa.

- **ASTEX:** uma base de dados construída para validar experimentos de ancoragem molecular, contendo 85 estruturas de alta qualidade experimental e com características drogáveis Hartshorn et al. (2007). A proposta dessa base de dados é a diversificação dos ligantes, contendo diversos complexos proteína-ligante com diferentes tipos de reconhecimento molecular.

Capítulo 4

Resultados

Este capítulo discute os resultados obtidos na avaliação do GRaSP em comparação com outros métodos usando diferentes bases de dados. Como forma de validar o GRaSP, os experimentos foram realizados visando responder as seguintes perguntas:

- O desempenho preditivo do GRaSP é consistente em cenários adversos, onde proteínas apresentam diferentes características?
- O GRaSP é capaz de processar proteínas de múltiplas cadeias?
- Como o GRaSP se comporta com proteínas nos estados acoplado ao ligante e desacoplado aoligante?
- O GRaSP apresenta resultados compatíveis a outros métodos do estado da arte?
- Como as previsões do GRaSP podem ser comparadas com resultados de métodos populares baseados em cavidade?

Com o intuito de avaliar se a estratégia aqui proposta é mesmo capaz de identificar resíduos de sítio de ligação, na Seção 4.1, o GRaSP é comparado com outros métodos que buscam por resíduos de sítio, considerados estado da arte, incluindo métodos participantes do CASP 10 (Seção 4.1.2). Com o objetivo de verificar a consistência das previsões em cenários adversos, onde proteínas apresentam características distintas, foram realizados experimentos em diferentes bases de dados, cada qual com sua respectiva peculiaridade, como alvos drogáveis (base de dados ASTEX), proteínas no estado acoplado e desacoplado do ligante (base de dados B44/U44), e estruturas diversas contendo proteínas de cadeia simples e múltiplas cadeias (base de dados HOLO4K) (Seção 4.1.3). Na Seção 4.2 avalia-se a estratégia a partir de uma perspectiva voltada para a busca de cavidades, comparando as previsões calculadas pelo GRaSP com as

previsões de métodos baseados em cavidades. E, finalmente, na Seção 4.3 é apresentada uma aplicação web implementada para hospedar a estratégia proposta neste trabalho.

4.1 Experimentos centrados no resíduo

Nesta seção são apresentados os resultados referentes aos experimentos realizados para comparar o GRaSP com outros métodos preditivos com o foco em classificar os resíduos de aminoácidos de uma proteína de forma individual, discriminando se esses pertencem ou não a um sítio de ligação. O formato de predição dessa categoria de classificadores consiste em listar aminoácidos na proteína alvo com potencial de interação com ligantes biologicamente relevantes. Esse tipo de avaliação não inclui pontuação ou probabilidade para as predições, de modo que os resíduos são classificados de forma binária, em positivos, caso pertençam ao sítio de ligação, e negativos, caso não pertençam.

Neste trabalho, todos os dados de proteínas utilizados como teste tem seus sítios de ligação anotados de acordo com a definição utilizada nos trabalhos de Gallo Casarino et al. (2014) e Yang et al. (2012). Um sítio de ligação observado é definido pelo conjunto de resíduos de uma proteína em que pelo menos um átomo, exceto hidrogênio, está a uma distância $d_{i,j}$ de qualquer átomo do ligante considerado biologicamente relevante, onde:

$$d_{i,j} \leq r_i + r_j + c \quad (4.1)$$

sendo $d_{i,j}$ a distância entre o átomo do resíduo i e o átomo do ligante j ; r_i e r_j os raios de Van der Waals dos átomos envolvidos; e c uma distância de tolerância de 0,5 Å.

4.1.1 Comparando GRaSP com o método COACH

O trabalho de destaque na literatura para classificar resíduos de sítio de ligação é o COACH, desenvolvido por Yang et al. (2013), que combina a predição de cinco métodos diferentes. Os resultados provenientes de cada método são combinados através de um modelo de aprendizagem de máquina para buscar por resíduos de sítio em proteínas de cadeia simples. Com o intuito de comparar GRaSP com COACH foram usadas estruturas de proteínas da base de dados desenvolvida pelos próprios autores em (Yang et al., 2013). Essa base contém 500 estruturas de cadeia simples e sem redundância.

Usando os resultados reportados por Liu et al. (2020), também foi possível comparar o GRaSP com os demais métodos utilizados pelo COACH. A tabela 4.1 mostra os resultados obtidos com os experimentos. GRaSP equiparou-se aos demais métodos, considerando o MCC como métrica de referência. O MCC é um coeficiente de correla-

Tabela 4.1. Resultados comparativos entre o GRaSP, COACH e seus métodos.

	TM-SITE	S-SITE	COFACTOR	FINDSITE	ConCavity	COACH	GRASP
MCC	0.51	0.45	0.46	0.44	0.33	0.60	0.61
Precision	0.59	0.45	0.61	0.45	0.26	0.59	0.69
Recall	0.51	0.58	0.41	0.51	0.62	0.70	0.61

ção entre as classes observadas e as previstas, apresentando uma avaliação mais robusta em relação às medidas de precisão e revocação usadas individualmente, segundo Yang et al. (2013). Resultados detalhados a respeito de cada alvo do conjunto de dados estão descritos no Apêndice B2.

Adicionalmente foi feita uma comparação exclusiva entre GRaSP e COACH utilizando os dados da base B44/U44. A Tabela 4.2 apresenta os valores alcançados por cada técnica para precisão, revocação e MCC. Em termos de desempenho preditivo, os resultados apresentados por GRaSP e COACH são equiparáveis. Entretanto, é importante destacar que, em média, o GRaSP leva 12 segundos para processar cada estrutura proteica, sendo que para a base de dados completa, levou-se 9 minutos. Esse processamento foi realizado em uma máquina Intel Core i5, 1,7 GHz, quatro núcleos, 12 GB de memória RAM e uma capacidade de armazenamento de 500 GB. Já quando uma tarefa é submetida para o COACH, o servidor reporta uma estimativa de até 4 horas para finalizar o experimento com uma única proteína.

Adicionalmente, para algumas estruturas, o processamento demorou mais do que 4 horas para ser finalizado, levando até dias. Por exemplo, o servidor do COACH levou mais de uma semana para completar o processamento com as instâncias 1DWD, 1NPC, e 1PDY. Além disso, vale destacar novamente que o COACH não trabalha com estruturas de múltiplas cadeias, conseguindo processar somente cadeia simples. Em contraste, o GRaSP é capaz de processar proteínas compostas por uma única cadeia

Tabela 4.2. Comparação entre GRaSP e COACH usando a base de dados B44/U44.

Dataset	Method	MCC	Precision	Recall
B44	GRaSP	0.67	0.64	0.77
	COACH	0.64	0.59	0.77
U44	GRaSP	0.67	0.61	0.80
	COACH	0.67	0.64	0.75

ou por múltiplas cadeias conectadas. Os resultados detalhados deste experimento estão presentes no Apêndice B3, para a base de dados B44, e no Apêndice B4, para a base de dados U44.

4.1.2 Experimento CASP 10

Nesta seção o GRaSP é avaliado com base no trabalho desenvolvido por Gallo Cassarino et al. (2014), onde é reportado resultados preditivos de diferentes métodos participantes do CASP 10 (*critical assessment of protein structure prediction*). Nesta edição, especificamente na categoria "predição de sítios de ligação (predição de função, FN)", o autor avalia 17 estratégias preditivas para resíduos de sítio de ligação. O conjunto de estruturas utilizadas como teste para avaliar os métodos é composto por apenas 13 proteínas alvo. Certamente não é um número que ofereça alguma significância estatística, porém, é possível tomar como base os protocolos e métricas empregados nesse tipo de avaliação para avaliar o GRaSP também em outros cenários. A Tabela 4.1.2 lista as proteínas alvo e seus respectivos resíduos de sítio, denotados pelos seus números de sequência.

Para esse experimento, especificamente, o conjunto de treinamento utilizado foi o mesmo proposto no CASP 10, contendo 25 estruturas modelos para cada uma das 13 enzimas. Dessa forma, cada método usufrui destes 25 modelos como ponto de partida

Tabela 4.3. Estruturas alvo do CASP10 e seus respectivos resíduos de sítio de ligação.

Alvo	Sítio de ligação (número do resíduo)
T0652	74, 79, 80, 99, 100, 101, 102, 103, 104, 165, 180, 182, 183
T0657	121, 132, 133, 143
T0659	43, 48, 63
T0675	21, 24, 37, 42, 49, 52, 65, 70
T0686	28, 30, 103
T0696	18, 69, 104
T0697	91, 150, 151, 152, 190, 243, 245, 247, 272, 274, 301, 303, 304, 351
T0706	25, 27, 99, 101, 129, 130
T0720	32, 34, 35, 62, 99, 113, 114, 115, 182, 188, 191, 194, 197, 200
T0721	10, 12, 13, 14, 33, 34, 35, 36, 37, 38, 39, 42, 45, 46, 60, 78,79, 80, 109, 110, 111, 114, 126, 136, 235, 237, 268, 269, 277, 278, 281
T0726	273, 277, 307
T0737	37, 40, 41, 42, 44, 45, 49, 78, 83, 114, 117, 118, 120, 121, 123, 124, 128, 130, 135, 138, 174, 237
T0744	22, 23, 24, 26, 58, 61, 120, 121, 122, 124, 196, 214, 216, 270, 271, 272, 273, 314, 316

para encontrar os sítios dos 13 alvos. Para o GRaSP isso foi uma desvantagem, já que a informação utilizada para construir o modelo preditivo foi reduzida. Além disso, nem todos as estruturas disponíveis para o experimento contêm informação sobre a localização do sítio, já que alguns dos arquivos estavam desprovidos de ligantes. Com a qualidade dos dados de entrada comprometida, e como o desempenho do GRaSP está sujeito a este dados de entrada, há uma defasagem nas predições realizadas, como veremos a seguir.

A Tabela 4.4 apresenta a lista de resíduos classificados pelo GRaSP como positivos, ou seja, pertencentes ao sítio, para cada alvo. A Figura 4.1 apresenta as predições realizadas pelo GRaSP a partir da representação tridimensional dos alvos. Em laranja estão destacados os resíduos verdadeiros positivos; em magenta estão destacados os falso positivos; e em cor azul estão destacados os falso negativos.

A Figura 4.2 mostra o desempenho do GRASP em relação aos demais métodos na categoria FN do experimento CASP10. Barras amarelas representam estratégias que utilizaram algum tipo de validação manual por especialistas (humano), enquanto as barras em lilás representam preditores com características automáticas (servidor). Os métodos são ranqueados de acordo MCC médio dentre as predições feitas nas 13 estruturas alvos. O GRASP foi ranqueado na sétima posição, com MCC médio de 0.58, onde, segundo Gallo Cassarino et al. (2014), as diferenças entre os 10 primeiros métodos não são estatisticamente significativas. Dentre os métodos automáticos, o GRaSP assume a terceira posição. Uma descrição completa a respeito dos resultados do GRaSP e demais métodos participantes do CASP10 está disponível no Apêndice B5.

Tabela 4.4. Resíduos de sítios preditos pelo GRaSP

Alvo	Resíduos de sítio de acordo com o GRaSP
T0652	100, 101, 102, 104
T0657	121, 132, 133, 143
T0659	-
T0675	21, 24, 37, 42, 49, 52, 65, 70
T0686	28, 30, 103
T0696	18, 69
T0697	150, 151, 152, 155, 190, 246, 247, 272
T0706	-
T0720	35, 99, 113, 115, 188, 191
T0721	10, 12, 14, 17, 46, 136, 278
T0726	44, 95, 273, 277, 307
T0737	40, 41, 42, 45, 46, 49, 83, 86, 87, 114, 117, 118, 120, 121
T0744	120, 122, 243, 282, 316

Para os alvos T0659 e T0706, o GRASP não identificou resíduos de sítio de ligação, então são atribuídos o valor zero para o MCC da predição destes alvos. De todos os 25 modelos para o alvo T0659, somente um deles contém ligante relevante, o que resultou em um conjunto de treinamento pobre de exemplos positivos. Para superar

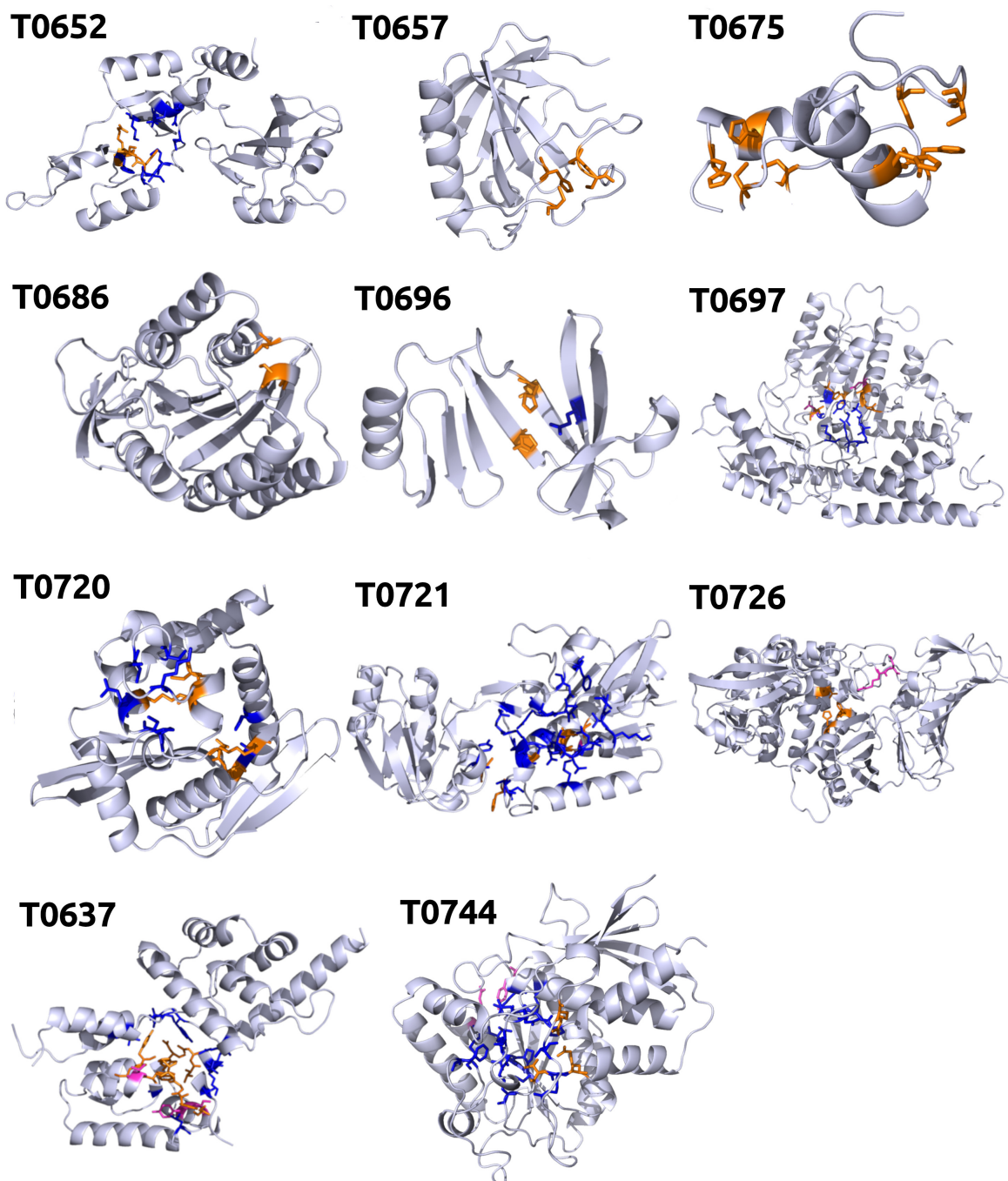


Figura 4.1. Representação estrutural das predições realizadas pelo GRASP no experimento CASP10.

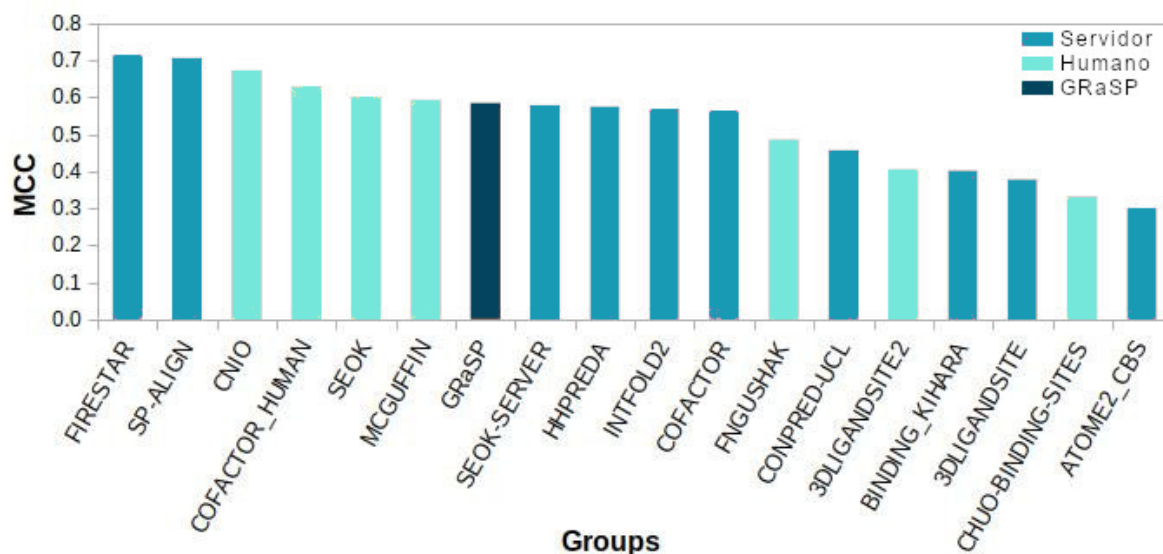


Figura 4.2. Métodos participantes do CASP10, junto com o GRASP, ordenados pelo MCC médio de suas predições.

essa limitação, foram usados modelos de outros alvos como conjunto de treinamento para criar o modelo de predição do alvo T0659. O resultado foi notável, com uma predição perfeita com $MCC = 1$. Ao levar em conta esse resultado, o MCC médio do GRASP sobe para 0.66, alcançando uma quarta posição dentre todos os métodos do experimento CASP10. Para o alvo T0706, apesar de serem usados 20 de seus respectivos exemplos na construção do modelo de classificação, não foi possível obter um bom MCC. Aproveitando a mesma estratégia usada no alvo T0659, foram usados exemplos de todos os outros alvos para construir um novo modelo de classificação, que também não conseguiu resultados significativos.

De acordo com a categoria FN do experimento CASP10, os alvos T0657 e T0659 foram os alvos mais desafiadores para os métodos participantes do experimento, obtendo valores baixos de MCC. Para o alvo T0657 o GRASP foi capaz de prever corretamente os resíduos de sítio de ligação ($MCC = 1$), assim como o alvo T0659, mencionado anteriormente, que obteve $MCC=1$ usando modelos dos demais alvos como conjunto de treinamento para o modelo preditivo.

4.1.3 Desempenho do GRASP em dados diversos

Para avaliar a qualidade das predições do GRASP em diferentes cenários, nesta seção, são descritos os experimentos usando conjuntos de dados proteicos com diferentes características, tais como: a base ASTEX, composta por moléculas drogáveis; o conjunto

B44/U44, contendo proteínas em seus estados acoplado e desacoplado ao ligante; e o conjunto HOLO4K, contendo maior volume de estruturas, incluindo proteínas com múltiplas cadeias. A Tabela 4.5 mostra os valores de MCC, precisão e revocação alcançados pelo GRaSP nos diferentes experimentos. A pequena variabilidade nos resultados mostra como o GRaSP consegue ser consistente mediante dados diversificados, evidenciando a robustez do método. O tempo necessário para processar cada estrutura nesse experimento foi em média 20 segundos.

Começando pela base de dados ASTEX, composta por 85 instâncias, o GRaSP alcançou um MCC médio de 0.66. Esse conjunto de proteínas tem a particularidade de ser constituído por alvos drogáveis. Tratando-se de um experimento preditivo de resíduos de sítio, uma precisão de 0,74 é um valor expressivo, refletindo uma baixa ocorrência de falsos positivos. Essa é uma característica importante para um classificador de sítios drogáveis. Por exemplo, no processo de busca por novos fármacos, quando não é conhecida a localização do sítio de ligação *a priori*, uma ancoragem molecular é executada em diversas regiões da proteína que foram consideradas inicialmente como candidatas a serem sítios de ligação. Uma baixa taxa de falsos positivos por parte do classificador evita experimentos de ancoragem molecular em regiões desnecessárias, otimizando o processo como um todo. Uma descrição mais detalhada dos resultados reportados pelo GRaSP para cada proteína do conjunto ASTEX encontra-se no Apêndice B6.

Já a base de dados B44/U44 é composta por 88 estruturas proteicas em seu estado acoplado e desacoplado do ligante, ou seja, são 44 pares de estruturas. O GRaSP obteve mesmo MCC médio de 0,67 para ambos os conjuntos de proteínas, acopladas (B44) e desacopladas (U44). Esse é um aspecto positivo da estratégia, já que encontrar sítios de ligação em estruturas desacopladas é uma tarefa desafiadora. Várias proteínas sofrem o ajuste induzido no momento que interagem com seus ligantes, sendo que, no modo desacoplado, e na ausência da indução, as cavidades podem apresentar um formato diferente, inclusive desapropriado para a ligação (Vajda et al., 2018; Du et al., 2016). Essa é uma característica importante para um classificador, que é o poder de generalização, onde, no contexto da predição de sítios, o classificador deve ser capaz de ignorar variações no dobramento da proteína, atentando somente aos padrões físico-

Tabela 4.5. Resultados do GRaSP para bases de dados diversos.

	ASTEX	B44	U44	HOLO4K
MCC	0.66	0.67	0.67	0.61
Precision	0.74	0.64	0.61	0.68
Recall	0.65	0.77	0.80	0.58

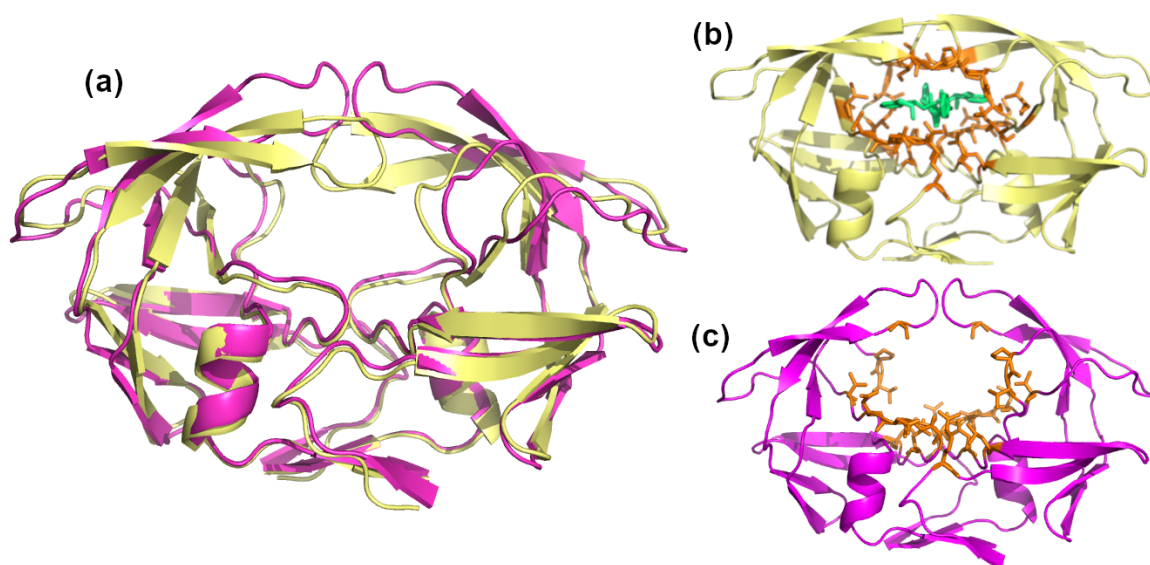


Figura 4.3. Estruturas da HIV protease, identificador PDB 4PHV em cor bege (holoproteína) acoplada ao ligante VAC e identificador PDB 3PHV em cor magenta (apoproteína) sobrepostas (a), acompanhadas da predição de seus sítios em cor alaranjada, 4PHV em (b) e 3PHV em (c).

químicos da cavidade. Normalmente, métodos geométricos e energéticos de predição, por serem extremamente sensíveis às variações conformacionais da proteína, sofrem com esse tipo de característica nos dados. Resultados detalhados para cada proteína dos conjuntos B44 e U44 encontram-se nos Apêndices B3 e B4, respectivamente.

A Figura 4.3 mostra um exemplo de predição feita pelo GRaSP para a HIV protease. Uma sobreposição é feita na Figura 4.3 (a) entre as estruturas do estado acoplado ao ligante de pdbid 4PHV (em bege) e o estado sem ligante com identificador pdb 3PHV (em magenta). É nítida a variação conformacional entre as estruturas, sendo que a 4PVH está mais compacta devido o encaixe induzido proporcionado pela interação com o ligante. A predição do sítios para 4PHV que está acoplada ao ligante VAC (em cor verde) é apresentada na Figura 4.3 (b) através de resíduos em cor alaranjada. Já na Figura 4.3 (c) está o sítio predito em cor laranja para a estrutura 3PHV.

Finalmente, o GRaSP foi usado para buscar por sítios de ligação na base de dados HOLO4K, um conjunto composto por 4.543 proteínas. O objetivo principal de usar este conjunto de estrutura é verificar a capacidade preditiva do GRaSP para cavidades, que será apresentado na Seção 4.2. Entretanto, é possível usufruir da diversidade contida nesta base de dados para avaliar o desempenho preditivo do GRaSP em proteínas com múltiplas cadeias. É preciso ressaltar que o COACH, então considerado estado da arte, não é projetado para processar proteínas com múltiplas cadeias.

A Figura 3.1 (a) mostra um exemplo de estrutura com múltiplas cadeias a par-

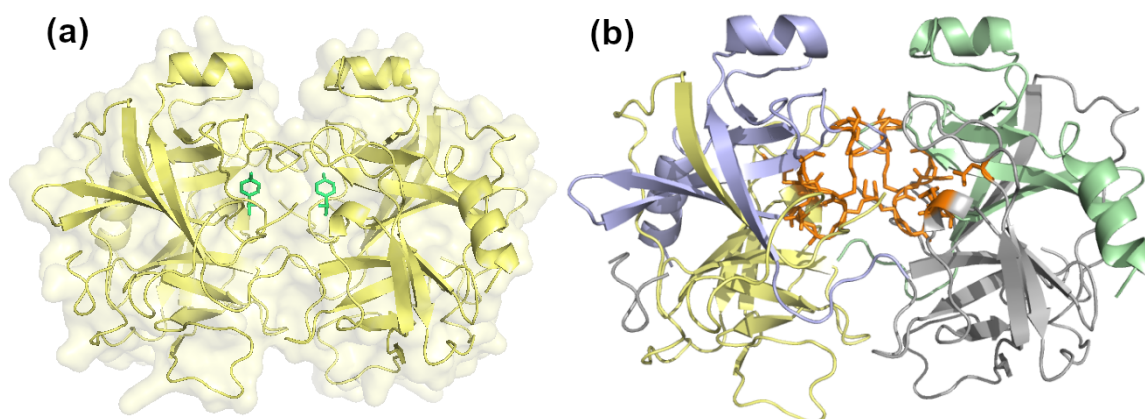


Figura 4.4. Estrutura da quimotripsina, (a) complexo proteína-ligante (pdb 2CHA), (b) predição dos resíduos de sítio realizada pelo GRASP.

tir da enzima digestiva quimotripsina (identificador pdb 2CHA). Como mostra a Figura 4.4 (b), GRASP foi capaz de encontrar o sítio de ligação (em laranja) para a estrutura 2CHA, localizado entre as superfícies de quatro cadeias polipeptídicas. Isso evidencia a capacidade que a estratégia aqui proposta tem de encontrar resíduos de sítio que compartilhem a superfície de múltiplas cadeias. Para o conjunto de dados HOLO4K como um todo o GRASP obteve um MCC médio de 0,61, considerado satisfatório e consistente com os demais experimentos.

4.2 Experimentos centrados na cavidade

Como mencionado anteriormente, as técnicas de predição baseadas em estrutura podem focar na classificação de resíduos ou na identificação de cavidades. O GRASP é uma estratégia centrada no resíduo, onde uma classificação binária é realizada, identificando se cada resíduo pertence ou não a um sítio de ligação. Existe um conjunto de métodos populares da categoria baseada em cavidade, como o FPocket, Metapocket 2.0, SiteHound, P2Rank e DeepSite, já descritos no Capítulo 2. Apesar de adotarem uma abordagem diferente em relação ao GRASP, há uma sobreposição nos objetivos, seja encontrando resíduos individualmente ou localizando cavidades, a essência do problema está em identificar uma região específica da proteína capaz de interagir com ligantes. Surgiu então uma questão importante: é possível comparar os resultados do GRASP com esses métodos de predição?

A solução reportada pelo GRASP é composta por uma lista de resíduos. Geralmente, aqueles preditos como sendo de sítio de ligação encontram-se próximos espacialmente ou, no caso de múltiplos sítios em uma mesma proteína, esses costumam

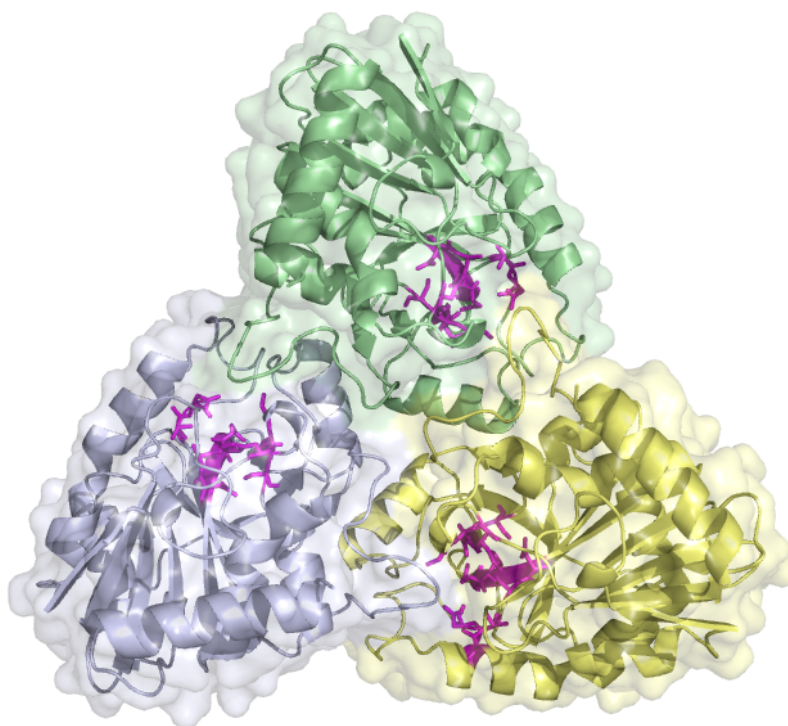


Figura 4.5. Resultados do GRaSP para a estrutura 1G2O com múltiplos sítios.

aparecer em grupos espalhados pela superfície. Por exemplo, na Figura 4.5, quando observa-se uma predição feita pelo GRaSP para a estrutura de identificador pdb 1G2O, é notável que os resíduos preditos (cor magenta) organizam-se em três regiões bem definidas. Com o intuito de emular a identificação de cavidades, optou-se então por agrupar os resíduos preditos pelo GRaSP a partir da sua proximidade no espaço, e, então, considerar cada um destes grupos como uma cavidade. Dessa forma, foi utilizado o algoritmo DBSCAN Schubert et al. (2017), um método de agrupamento baseado em densidade que separa as instâncias em conjuntos de alta densidade e baixa densidade. Assim, regiões da proteína onde se encontra uma alta densidade de resíduos preditos como positivos são consideradas como regiões de interação. Resíduos encontrados fora dessas regiões são descartados nesse tipo de avaliação. Para comparar o GRaSP com os métodos baseados em cavidade cada um dos grupos de resíduos computados pelo DBSCAN são considerados como uma cavidade.

Com o intuito de avaliar os diferentes métodos para encontrar cavidades, foi usada a métrica denominada DCA (Seção 3.5), assim como nos trabalhos (Chen et al., 2011; Krivák & Hoksza, 2018). Essa medida calcula a distância do centro da região predita para qualquer átomo do ligante original da proteína.

Seja uma proteína com n sítios de ligação, para cada método preditivo conside-

rado na análise foram também obtidos n cavidades preditas. Nesse caso específico, as cavidades selecionadas são aquelas melhores classificadas de acordo com a pontuação respectiva de cada método. Um sítio específico é corretamente predito caso a distância mínima entre seu ligante correspondente e qualquer uma das n predições de um determinado método está abaixo de um limiar. Trabalhos anteriores assumiram que um sítio de ligação é corretamente predita caso seu centro geométrico esteja no máximo a 4Å de distância de qualquer átomo do ligante (Chen et al., 2011; Jendele et al., 2019).

O desempenho de cada técnica foi avaliado usando as bases de dados COACH420 e HOLO4k, ambos obtidos de Jendele et al. (2019). COACH420 consiste de 420 proteínas de cadeia simples derivados da base de dados do COACH. Como alguns métodos não foram capazes de encontrar sítios de ligação para todas as estruturas, um subconjunto de 228 cadeias do COACH420 e 1.471 entradas do HOLO4K foram usadas no experimento, com o intuito de incluir todos os métodos na comparação.

A Figura 4.6 mostra o desempenho preditivo dos métodos de acordo com sua taxa de sucesso, que corresponde ao número de sítios corretamente classificados dividido pelo número total de sítios. Para evitar a seleção arbitrária de um limiar para o DCA, a taxa de sucesso foi calculada num intervalo entre 1Å e 20Å, inclusive. P2Rank superou os demais métodos, mas GRaSP tem um aumento de desempenho assumindo a segunda posição a partir do limiar de 4Å, que é o valor sugerido pela literatura. Tendo em mente que o GRaSP não foi devidamente projetado para predição de cavidades, mas sim para encontrar resíduos individualmente, o resultado é bastante satisfatório.

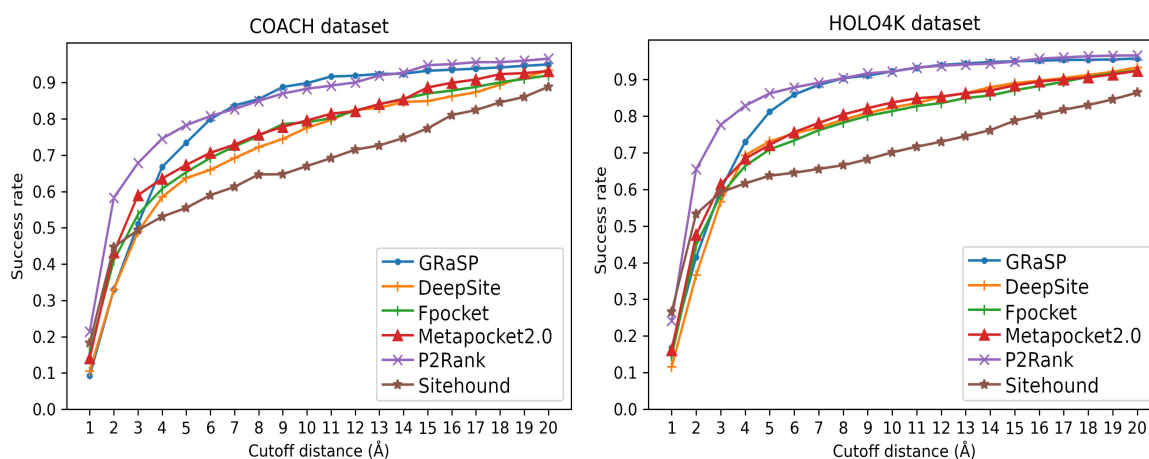


Figura 4.6. Desempenho preditivo do GRaSP comparado a métodos centrados em cavidades.

4.3 GRaSPweb

O GRaSP foi desenvolvido em Python3 e seu código fonte, assim como as bases de dados utilizadas nos experimentos, encontram-se disponíveis no repositório <https://github.com/charles-abreu/GRaSP>. A execução do GRaSP é feita através de linha de comando e os resultados são disponibilizados através de arquivos no formato *csv*. Essas características não são recomendáveis, pois o usuário pode não ter familiaridade com tais tecnologias, o que compromete sua experiência ao processar suas proteínas.

Para sanar tais brechas de usabilidade, se encontra em etapa de experimentação a ferramenta GRaSPweb, uma aplicação web que provê uma interface que possibilita aos usuários do GRaSP acessarem seus resultados de forma inteligível e visual. A existência de uma aplicação web isenta o usuário da laboriosa tarefa de executar linhas de comando e manipular arquivos, fazendo com que o mesmo concentre suas energias apenas na análise das predições. A Figura 4.7 e a Figura 4.8 correspondem a tela inicial e a tela de submissão do GRaSPweb, respectivamente. A aplicação atualmente roda em um servidor Apache e foi desenvolvida em Flask (Grinberg, 2018), com seu *front-end* baseado no framework Bootstrap (Spurlock, 2013).

Para conduzir a predição de resíduos de sítio de ligação, o usuário pode carregar um ou vários arquivos de estruturas de proteínas no formato *pdb*, ou informar um código de identificação de alguma estrutura contida na base de dados do Protein Data Bank. No último caso, o GRaSPweb se encarrega de recuperar as estruturas referentes aos códigos que foram informados pelo usuário.

O fluxo de execução do GRaSPweb está ilustrado na Figura 4.9. O primeiro passo é a submissão de uma ou várias proteínas à aplicação (Figura 4.9 (a)), onde, para cada

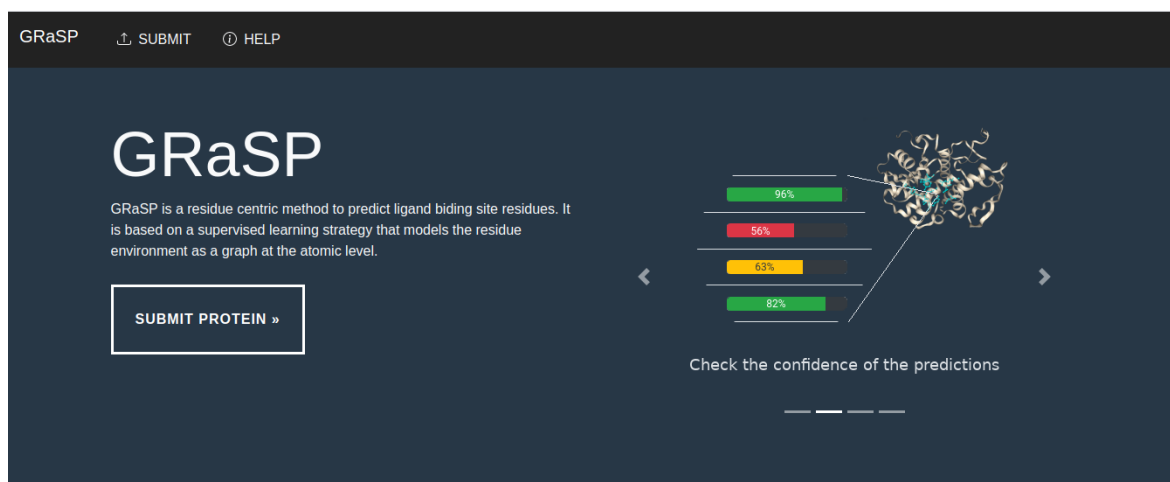


Figura 4.7. Página inicial do GRaSPweb.

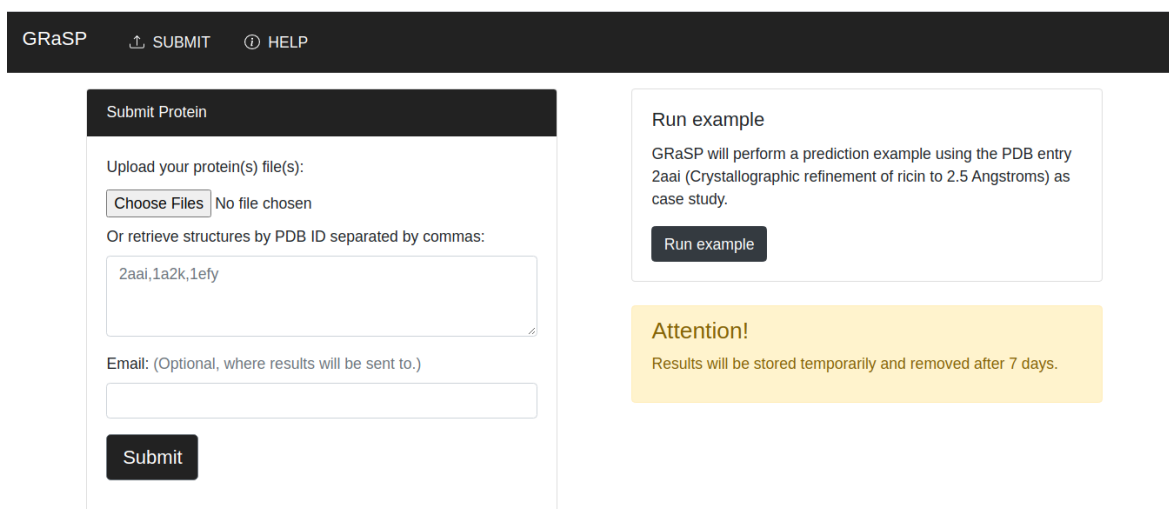


Figura 4.8. Página de submissão do GRaSPweb.

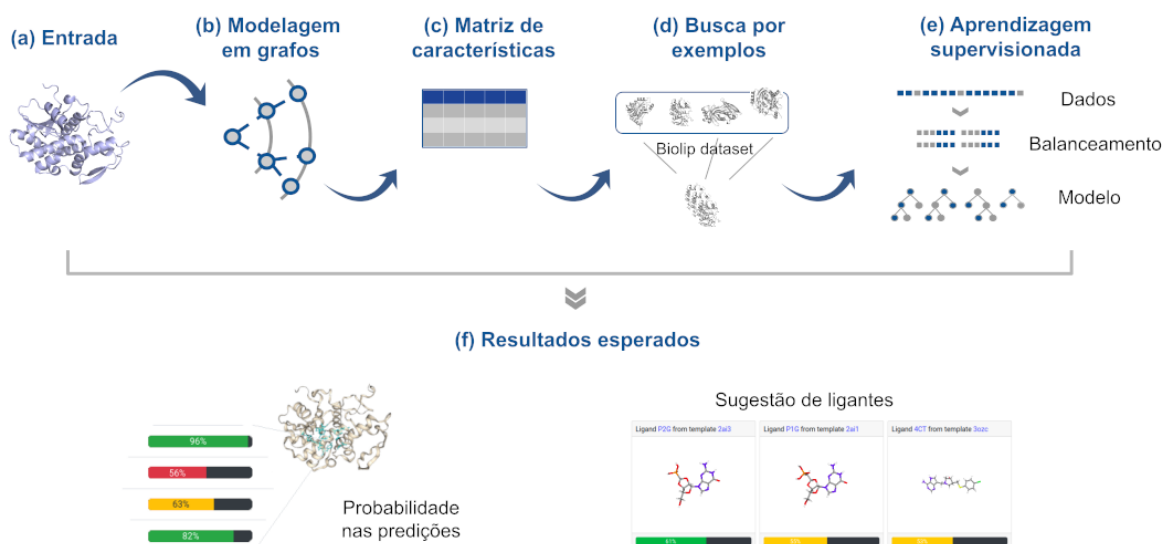


Figura 4.9. Fluxo de execução do GRaSPweb.

uma destas proteínas, os resíduos são modelados como grafos de vizinhança e codificados em uma matriz de características (Figura 4.9 (b,c)). A sequência de aminoácidos da proteína é utilizada na busca por exemplos presentes na base de dados de exemplos (Figura 4.9 (d)), estes que serão utilizados como conjunto de treinamento para a construção do modelo preditivo. Os dados são devidamente balanceados e o modelo de classificação é construído usando um conjunto de classificadores (Figura 4.9 (e)). Finalmente os resultados são apresentados ao usuário, de forma visual e inteligível (Figura 4.9 (f)).

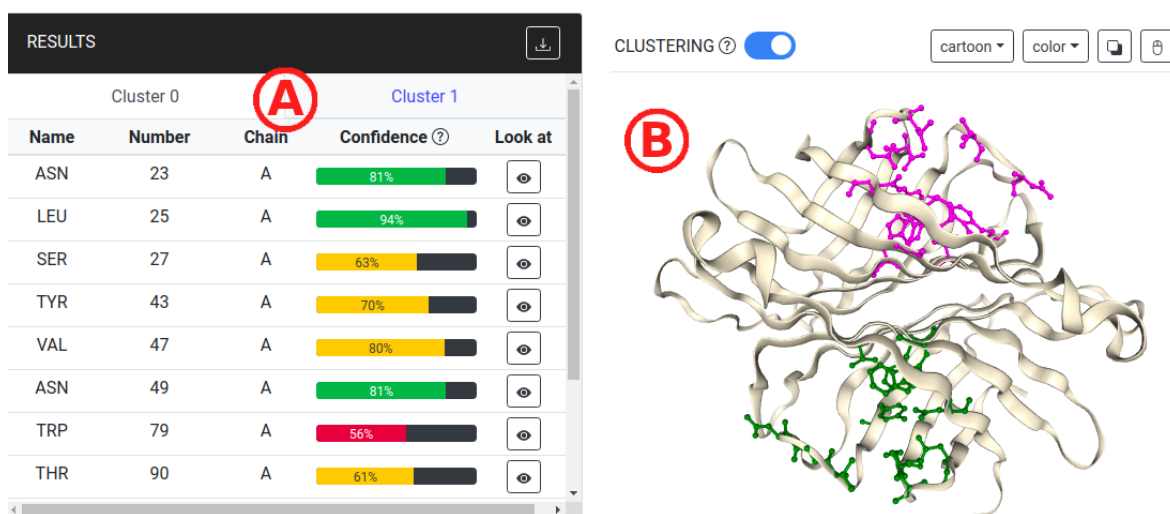


Figura 4.10. Exemplo de resultado apresentado pelo GRASPweb após a predição.

A interpretabilidade dos resultados é fundamental para auxiliar o usuário em seus experimentos com proteínas. Tornar as predições do GRASP mais informativas é um outro objetivo do GRASPweb, disponibilizando para o usuário ferramentas que o auxiliem na análise dos resultados. A saída padrão para cada proteína submetida ao GRASPweb é uma lista dos resíduos classificados como positivos, ou seja, os resíduos de sítio encontrados na proteína alvo, como está ilustrado na Figura 4.10 (A). A primeira funcionalidade que pode ser observada nessa lista é a inclusão de uma pontuação de confiança para cada resíduo predito. Com o intuito de deixar explícito a confiabilidade da predição, o GRASPweb provê a média da probabilidade de classificação, oriunda do percentual de classificadores que optaram pela classe positiva dentro do conjunto de preditores.

Para fazer com que o usuário tenha uma experiência visual dos resultados preditivos, o GRASPweb utiliza o visualizador molecular NGL (Rose & Hildebrand, 2015), uma aplicação compatível com os navegadores web modernos, que provê uma interface gráfica rica para a apresentação de moléculas, com a incremento de funcionalidades para manipulação e customização das apresentações. No contexto do GRASPweb, o NGL é responsável por mostrar a proteína alvo, onde os resíduos de sítio de ligação preditos são destacados, como mostra a Figura 4.10 (B).

Como já foi descrito anteriormente, o GRASP é um método centrado no resíduo e interpreta o problema de predição de sítios de ligação como um problema binário de classificação para cada resíduo individualmente. Para mostrar o sítio de ligação como uma unidade, e também discriminar as diferentes regiões de sítio numa mesma pro-

teína, o GRaSPweb implementa o algoritmo DBSCAN para agrupar resíduos próximos espacialmente e emular cavidades. Como mostrado na Figura 4.10, ao habilitar o botão *CLUSTERING*, localizado no canto superior esquerdo do visualizador molecular, o algoritmo de agrupamento separa áreas com diferentes densidades de resíduos preditos como positivos, apresentando-as com cores diferentes.

Um outra característica oferecida pelo GRaSPweb é a sugestão de ligantes para os sítios preditos com base na similaridade de cavidades. Como compostos similares tendem a interagir com cavidades similares, devido ao reconhecimento molecular, é possível aproveitar a informação sobre sítios de ligação contida na base de dados de exemplos para sugerir ligantes aos sítios preditos pelo GRaSP. Usando o algoritmo do PocketMatch (Yeturu & Chandra, 2008), GRaSPweb compara os sítios preditos com os sítios presentes na base de dados de exemplos. As cavidades que apresentarem maior pontuação na base de exemplos são recuperadas e seus respectivos ligantes são sugeridos ao usuário, como mostrado na Figura 4.11. Uma barra de pontuação também é provida para cada ligante sugerido, objetivando representar a pontuação reportada pelo PocketMatch para comparar o sítio predito com o sítio oriundo da base de exemplo.

O GRaSPweb já está operando com a maioria das suas características já implementadas e disponíveis em <https://grasp.ufv.br/>.

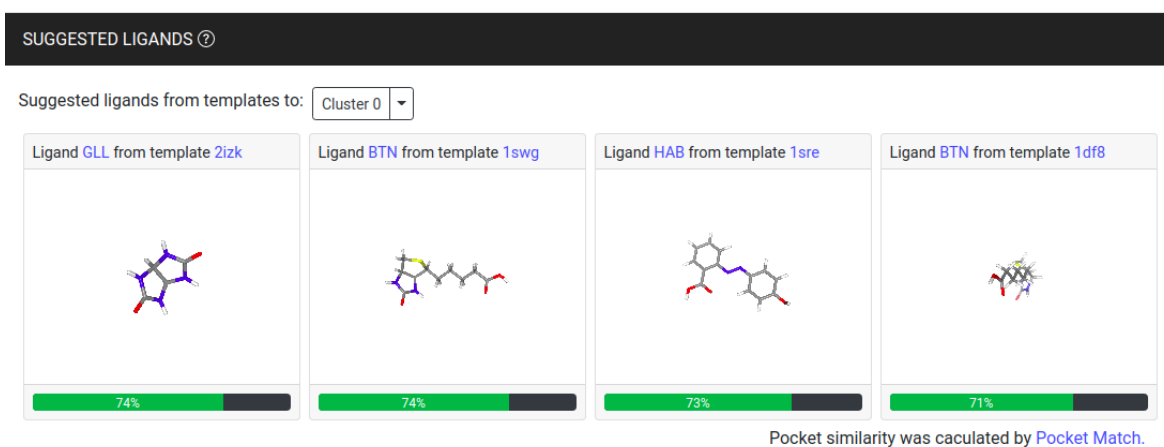


Figura 4.11. Exemplo de sugestão de ligantes feita pelo GRaSPweb.

Capítulo 5

Considerações Finais

Nesta tese foi proposta uma estratégia para a predição de resíduos de sítios de ligação denominada GRaSP. A estratégia faz uso de modelos baseados em grafos junto com aprendizagem supervisionada para classificar se os resíduos de uma proteína pertencem a algum sítio de ligação ou não. Cada resíduo da proteína é codificado em um grafo que tenta capturar as propriedades físico-químicas e topológicas do ambiente físico-químico entorno do resíduo. Nesse grafo denominado por grafo de vizinhança, os nós denotam os átomos e as arestas denotam interações não-covalentes entre os átomos do resíduo e os átomos da sua vizinhança de resíduos. Esses grafos são enriquecidos com propriedades físico-químicas, que por sua vez são codificados em um vetor de características. Esses vetores são combinados em uma matriz de características que alimenta o algoritmo de Árvores Extremamente Aleatórias usado para discriminar se os resíduos da proteína pertencem a um sítio de ligação ou não.

Os experimentos mostraram que o GRaSP foi capaz de obter resultados comparáveis ou superiores com a literatura, e com maior eficiência em relação ao tempo de processamento comparando-se ao método considerado como estado-da-arte. Além de sua velocidade, a estratégia é capaz de processar um conjunto heterogêneo de dados de proteínas, como, por exemplo, proteínas de cadeia simples ou compostas por múltiplas cadeias, proteínas no estado acoplado ao ligante ou desacoplado, e alvos drogáveis. É válido ressaltar que algumas dessas características são empecilhos para outros métodos. O GRaSP foi projetado para classificar resíduos individualmente, porém, ao agrupar os resíduos preditos usando um algoritmo de agrupamento, foi possível compará-lo a métodos centrados em prever cavidades, obtendo resultados satisfatórios e até mesmo superiores a estes métodos.

Como próximo passo no desenvolvimento da estratégia pretende-se implantar a ferramenta GRaSPweb para hospedar o algoritmo do GRaSP, de forma que, através de

uma interface amigável, o usuário possa usufruir das predições de uma maneira informativa. A ferramenta encontra-se em etapa experimental e com suas funcionalidades básicas já implementadas. Observando também os resultados promissores do algoritmo aqui proposto para predizer sítios referentes aos compostos não proteicos, têm-se em mente a possibilidade de usufruir da ideia central da estratégia, no caso a modelagem em grafos, para aplicar na predição de sítios no âmbito proteína-proteína. Com os experimentos realizados, acredita-se, em tese, que esse tipo de modelagem pode ter resultados promissores também em sítios de interação proteína-proteína.

Referências Bibliográficas

- Azevedo, L. D. d.; Bastos, M. M.; Oliveira, A. P. d. & Boechat, N. (2017). Sínteses e propriedades de fármacos inibidores da tirosina quinase bcr-abl, utilizados no tratamento da leucemia mieloide crônica. *Química Nova*, 40(7):791--809.
- Balchin, D.; Hayer-Hartl, M. & Hartl, F. U. (2016). In vivo aspects of protein folding and quality control. *Science*, 353(6294).
- Barabasi, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101--113.
- Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B. et al. (2020). Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235--242.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.
- Brylinski, M. & Skolnick, J. (2008). A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of sciences*, 105(1):129--134.
- Campelo, J. A. F. G.; Monteiro, C. R.; da Silveira, C. H.; de Azevedo Silveira, S. & de Melo-Minardi, R. C. (2019). Protein structural signatures revisited: Geometric linearity of main chains are more relevant to classification performance than packing of residues. Em *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 391--402. Springer.
- Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M. & Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. *PLoS Comput Biol*, 5(12):e1000585.

- Chen, K.; Mizianty, M. J.; Gao, J. & Kurgan, L. (2011). A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure*, 19(5):613--621.
- Chen, X. & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323--329.
- Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, 426(6968):884--890.
- Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L. & Liu, S.-Q. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144.
- Fassio, A. V.; Martins, P. M.; Guimarães, S. d. S.; Junior, S. S.; Ribeiro, V. S.; de Melo-Minardi, R. C. & Silveira, S. d. A. (2017). Vermont: a multi-perspective visual interactive platform for mutational analysis. *BMC bioinformatics*, 18(10):51-63.
- Fassio, A. V.; Santos, L. H.; Silveira, S. A.; Ferreira, R. S. & de Melo-Minardi, R. C. (2019). napoli: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(4):1317--1328.
- Gallo Cassarino, T.; Bordoli, L. & Schwede, T. (2014). Assessment of ligand binding site predictions in casp10. *Proteins: Structure, Function, and Bioinformatics*, 82:154-163.
- Geurts, P.; Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3--42.
- Gherzi, D. & Sanchez, R. (2009). Easymifs and sitehound. *Bioinformatics*, 25(23):3185-3186.
- Gonçalves-Almeida, V. M.; Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Meira, W. & Santoro, M. M. (2012). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342--349.
- Govindaraj, R. G.; Naderi, M.; Singha, M.; Lemoine, J. & Brylinski, M. (2018). Large-scale computational drug repositioning to find treatments for rare diseases. *NPJ systems biology and applications*, 4(1):1--10.

- Grinberg, M. (2018). *Flask web development: developing web applications with python*. "O'Reilly Media, Inc."
- Haas, J.; Barbato, A.; Behringer, D.; Studer, G.; Roth, S.; Bertoni, M.; Mostaguir, K.; Gumienny, R. & Schwede, T. (2018). Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387--398.
- Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220--239.
- Hamelryck, T. (2005). An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*, 59(1):38--48.
- Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N. & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein- ligand docking performance. *Journal of medicinal chemistry*, 50(4):726--741.
- Izidoro, S. C.; de Melo-Minardi, R. C. & Pappa, G. L. (2015). Gass: identifying enzyme active sites with genetic algorithms. *Bioinformatics*, 31(6):864--870.
- Jendele, L.; Krivak, R.; Skoda, P.; Novotny, M. & Hoksza, D. (2019). Prankweb: a web server for ligand binding site prediction and visualization. *Nucleic acids research*, 47(W1):W345--W349.
- Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S. & De Fabritiis, G. (2017). Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036--3042.
- Kana, O. & Brylinski, M. (2019). Elucidating the druggability of the human proteome with e findsite. *Journal of computer-aided molecular design*, 33(5):509--519.
- Krivák, R. & Hoksza, D. (2015). Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of cheminformatics*, 7(1):1--13.
- Krivák, R. & Hoksza, D. (2018). P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10(1):1--12.

- Le Guilloux, V.; Schmidtke, P. & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):1--11.
- Liu, Y.; Grimm, M.; Dai, W.-t.; Hou, M.-c.; Xiao, Z.-X. & Cao, Y. (2020). Cb-dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacologica Sinica*, 41(1):138--144.
- Macari, G.; Toti, D. & Polticelli, F. (2019). Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies. *Journal of computer-aided molecular design*, 33(10):887--903.
- Medina, S. G.; Fassio, A. V.; Silveira, S. d. A.; da Silveira, C. H. & de Melo-Minardi, R. C. (2017). Cali: A novel visual model for frequent pattern mining in protein-ligand graphs. Em *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 352--358. IEEE.
- Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L.; Tosatto, S. C.; Paladin, L.; Raj, S.; Richardson, L. J. et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412--D419.
- Monteiro, C.; Mendes, V.; Comarela, G. & Silveira, S. A. (2018). Using supervised learning successful descriptors to perform protein structural classification through unsupervised learning. Em *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 75--78. IEEE.
- Moraes, J. P.; Pappa, G. L.; Pires, D. E. & Izidoro, S. C. (2017). Gass-web: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic acids research*, 45(W1):W315--W319.
- Okun, O. & Skarlas, L. (2011). *Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations*, volume 445. Medical Information Science Reference.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825--2830.
- Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A. et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221--227.

- Ribeiro, V. S.; Santana, C. A.; Fassio, A. V.; Cerqueira, F. R.; da Silveira, C. H.; Romanelli, J. P.; Patarroyo-Vargas, A.; Oliveira, M. G.; Gonçalves-Almeida, V.; Izidoro, S. C. et al. (2020). visgremlin: graph mining-based detection and visualization of conserved motifs at 3d protein-ligand interface at the atomic level. *BMC bioinformatics*, 21(2):1--12.
- Roche, D. B.; Brackenridge, D. A. & McGuffin, L. J. (2015). Proteins and their interacting partners: An introduction to protein–ligand binding site prediction methods. *International journal of molecular sciences*, 16(12):29829--29842.
- Roche, D. B.; Tetchner, S. J. & McGuffin, L. J. (2010). The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics*, 26(22):2920--2921.
- Rose, A. S. & Hildebrand, P. W. (2015). Ngl viewer: a web application for molecular visualization. *Nucleic acids research*, 43(W1):W576--W579.
- Roy, A.; Yang, J. & Zhang, Y. (2012). Cofactor: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40(W1):W471--W477.
- Santana, C. A.; Cerqueira, F. R.; Da Silveira, C. H.; Fassio, A. V.; de Melo-Minardi, R. C. & Silveira, S. d. A. (2016). Gremlin: A graph mining strategy to infer protein-ligand interaction patterns. Em *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 28--35. IEEE.
- Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P. & Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1--21.
- Shi, H.; Liu, S.; Chen, J.; Li, X.; Ma, Q. & Yu, B. (2019). Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics*, 111(6):1839--1852.
- Somody, J. C.; MacKinnon, S. S. & Windemuth, A. (2017). Structural coverage of the proteome for pharmaceutical applications. *Drug discovery today*, 22(12):1792--1799.
- Spurlock, J. (2013). *Bootstrap: Responsive Web Development*. "O'Reilly Media, Inc."
- Tan, P.-N.; Steinbach, M. & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

- Tran-Nguyen, V.-K.; Da Silva, F.; Bret, G. & Rognan, D. (2018). All in one: Cavity detection, druggability estimate, cavity-based pharmacophore perception, and virtual screening. *Journal of chemical information and modeling*, 59(1):573--585.
- Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M. & Whitty, A. (2018). Cryptic binding sites on proteins: definition, detection, and druggability. *Current opinion in chemical biology*, 44:1--8.
- Van Kreveld, M.; Schwarzkopf, O.; de Berg, M. & Overmars, M. (2000). *Computational geometry algorithms and applications*. Springer.
- Wu, Q.; Peng, Z.; Zhang, Y. & Yang, J. (2018). Coach-d: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic acids research*, 46(W1):W438--W442.
- Yang, J.; Roy, A. & Zhang, Y. (2012). Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research*, 41(D1):D1096--D1103.
- Yang, J.; Roy, A. & Zhang, Y. (2013). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, 29(20):2588--2595.
- Yeturu, K. & Chandra, N. (2008). Pocketmatch: a new algorithm to compare binding sites in protein structures. *BMC bioinformatics*, 9(1):1--17.
- Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M. & Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, 27(15):2083--2088.
- Zhao, J.; Cao, Y. & Zhang, L. (2020). Exploring the computational methods for protein-ligand binding site prediction. *Computational and structural biotechnology journal*, 18:417--426.
- Zheng, A. & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. "O'Reilly Media, Inc."

Apêndice A

Artigo publicado em periódico

Proteins

GRaSP: a graph-based residue neighborhood strategy to predict binding sites

Charles A. Santana^{1,2,†}, Sabrina de A. Silveira^{3,4,*†}, João P. A. Moraes⁴, Sandro C. Izidoro⁴, Raquel C. de Melo-Minardi^{1,2}, Antônio J. M. Ribeiro⁵, Jonathan D. Tyzack⁵, Neera Borkakoti⁵ and Janet M. Thornton⁵

¹Department of Biochemistry and Immunology and ²Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil, ³Department of Computer Science, Universidade Federal de Viçosa, Viçosa 36570-900, Brazil, ⁴Institute of Technological Sciences (ICT), Advanced Campus at Itabira, Universidade Federal de Itabira, Itabira 35903-087, Brazil and ⁵European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: The discovery of protein–ligand-binding sites is a major step for elucidating protein function and for investigating new functional roles. Detecting protein–ligand-binding sites experimentally is time-consuming and expensive. Thus, a variety of *in silico* methods to detect and predict binding sites was proposed as they can be scalable, fast and present low cost.

Results: We proposed Graph-based Residue neighborhood Strategy to Predict binding sites (GRaSP), a novel residue centric and scalable method to predict ligand-binding site residues. It is based on a supervised learning strategy that models the residue environment as a graph at the atomic level. Results show that GRaSP made compatible or superior predictions when compared with methods described in the literature. GRaSP outperformed six other residue-centric methods, including the one considered as state-of-the-art. Also, our method achieved better results than the method from CAMEO independent assessment. GRaSP ranked second when compared with five state-of-the-art pocket-centric methods, which we consider a significant result, as it was not devised to predict pockets. Finally, our method proved scalable as it took 10–20 s on average to predict the binding site for a protein complex whereas the state-of-the-art residue-centric method takes 2–5 h on average.

Availability and implementation: The source code and datasets are available at <https://github.com/charles-abreu/GRaSP>.

Contact: sabrina@ufv.br

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The Uniprot database, the catalog of all known protein sequences, currently contains over 120 million entries, of which only about half a million have been curated by experts (Consortium, 2019). Pfam (El-Gebali *et al.*, 2019), a database of protein families, contains 22% (3961) of all entries labeled as domains of unknown function, despite continual efforts to annotate these structures.

The interplay between proteins and their small-molecule partners is important in a variety of biological processes like signal transmission, cellular defences and enzymatic catalysis, to name a few. A special type of proteins, the enzymes, interact with small

molecules to catalyze chemical reactions, regulating and synchronizing them, ensuring that they take place on the appropriate time and scale to make life possible (Warshel and Bora, 2016). Identifying protein–ligand-binding sites is important for elucidating protein function and for investigating new functional roles. However, determining protein interacting partners and their binding sites through *in vitro* and *in vivo* experimental methods is expensive and time-consuming (Ding *et al.*, 2017; Duan *et al.*, 2016; Nitsche and Otting, 2018; Roche *et al.*, 2015). To overcome these limitations, many computational methods have been developed to identify and predict protein–ligand-binding sites. These methods are particularly interesting as they can be fast and applied in large scale and at low cost (Komiya *et al.*, 2016).

Prediction methods for ligand-binding sites can be segmented into three major categories: sequence-based, structure-based and hybrid. The volume of available protein sequences is much larger than structures, thus sequence-based methods are useful especially when there is no structural data. Nevertheless, the top performing computational approaches to determine ligand-binding sites are based on structures (Gallo Cassarino *et al.*, 2014; Macari *et al.*, 2019; Schmidt *et al.*, 2011). Moreover, structural coverage of protein sequences is increasing. For example, structural coverage is approximately 70% for the human proteome and 95% if we consider human drug targets (Somody *et al.*, 2017). Nowadays a protein structure can be obtained through experimental or computational means (based on homologues) for the majority of proteins encoded by common model organism genomes (Schwede, 2013). Hence, in this work, we focus on methods that are mainly based on structure.

Among structure-based methods to predict ligand-binding sites, a subset is focused on pocket prediction. Some perform *ab initio* modeling of pockets while others use machine learning/artificial intelligence techniques based on examples. Two representative methods are DeepSite (Jiménez *et al.*, 2017) and P2Rank (Jendele *et al.*, 2019; Krivák and Hoksza, 2018). DeepSite is a technique based on deep convolutional neural networks that treats protein structures as 3D images, inspired by computer vision. The proteins are discretized into voxels and each voxel consists of a set of atom-level pharmacophoric properties, which are called channels. DeepSite outperformed fPocket (Le Guilloux *et al.*, 2009) and Concavity (Capra *et al.*, 2009), methods that were then considered as the state of the art. P2Rank, in turn, is a machine learning based technique that learns from examples. It is based on classification of points spread on the solvent accessible surface of the protein. These points are characterized by a vector consisting of physicochemical, geometric and statistical properties calculated from its local geometric neighborhood. P2Rank outperformed other compared methods, such as fPocket, Site Hound (Gherzi and Sanchez, 2009), MetaPocket 2.0 (Zhang *et al.*, 2011) and DeepSite.

Another subset of structure-based methods focuses on the prediction of ligand-binding residues. The cavity-centric approach is interesting as it may perform *ab initio* modeling of the ligand-binding site, especially the techniques that do not rely on examples. Also, it may be able to find binding sites when only distant-homologues are available. However, the residue-centric approach comprises highly accurate methods. Also, in initiatives that perform independent assessment of ligand-binding sites (Gallo Cassarino *et al.*, 2014; Haas *et al.*, 2018), the evaluation methodology is residue-centric. Thus, here, we are interested in methods primarily focused on predicting ligand-binding residues.

A variety of computational strategies have been proposed to identify protein binding site residues. Here we briefly comment on some representative examples. Firestar (López *et al.*, 2007; Lopez *et al.*, 2011) is a server that uses functional information in FireDB (Lopez *et al.*, 2007), a database of functionally important residues, to make predictions of ligand-binding and also catalytic residues in protein sequences. Predictions are inferred from local sequence conservation matches to small-molecule ligand-binding residues in FireDB and catalytic residues from CSA (Porter *et al.*, 2004).

FunFOLD3 (Roche *et al.*, 2011, 2013) is a method that calculates ligand-binding site, its residues, EC number and GO terms for a target protein as well as ligands that may bind to such target protein. It is composed by two main steps: (i) FunFOLD superposes a list of structural templates which contains ligands of biological relevance. The best templates are then superimposed on the target model and ligands from the templates are segmented into clusters. The potential ligand-binding site of a protein is the one that contains the largest cluster; (ii) FunFOLDQA assesses the quality of prediction generating confidence and evaluation scores.

COACH (Yang *et al.*, 2013) is a consensus approach that combines results of its own algorithms, TM-SITE and S-SITE, with other three third-party ligand-binding site prediction tools, COFACTOR (Roy *et al.*, 2012), FINDSITE (Brylinski and Skolnick, 2008) and ConCavity (Capra *et al.*, 2009), using a supervised learning

technique. COACH comprises two algorithms: (i) TM-SITE is structure-based and calculates ligand-binding sites from structure-related templates with the alignments built on binding-specific sub-structures matches; (ii) S-SITE is an algorithm to detect protein templates and the ligand-binding site through binding site specific, sequence profile-profile comparisons. In a similar manner to FunFold, COACH also outputs ligand-binding site, its residues, EC number and GO terms for a target protein and ligands that may bind to the target protein. An enhanced version of the method, named COACH-D (Wu *et al.*, 2018), was recently proposed. COACH-D uses COACH to predict ligand-binding sites. Then, AutoDock Vina is used to dock ligands informed by users or from templates into the binding pockets to refine the ligand-binding poses. COACH was considered one of the state-of-the-art protein-ligand-binding site prediction methods in Liu *et al.* (2020), outperforming 6 other methods in a dataset of 500 non-redundant single chains.

LigDig (Fuller *et al.*, 2015), in turn, uses ligands as the starting point for binding site prediction. The method creates a ligand interaction network to combine information from different databases to identify similar ligands as well as their binding proteins. The method is able to identify proteins that can potentially form complex with a specific ligand, which allows to predict potential protein-ligand-binding sites.

GASS (Izidoro *et al.*, 2015; Moraes *et al.*, 2017) is a genetic algorithm that searches for active site structural templates in unknown proteins. Given active site templates from CSA (Porter *et al.*, 2004), the method evolves a population of candidate active sites by simulating evolutionary effects, as crossover and mutation, according to user defined probabilities. Candidate solutions are assessed and ranked according to a fitness function, which is similar to an RMSD between the active site template and the candidate active site found by GASS. This process is repeated for a specified number of generations.

It is important to point out that the quality of methods has increased significantly over time, especially due to CASP (Gallo Cassarino *et al.*, 2014) and CAMEO (Haas *et al.*, 2018) initiatives, that allowed independent assessment of binding site residue predictions performed by different methods. For a detailed review of protein-ligand-binding site prediction methods, see Roche *et al.* (2015) and Macari *et al.* (2019).

As shown, several methods were proposed to characterize, understand and predict ligand-binding sites. Nonetheless, in spite of the relevant contributions of the majority of the works, methods that depend on multiple structural alignment might be prohibitively expensive for large scale processing. As the biological data has been growing in a fast pace, scalable techniques are pivotal in real-world scenarios. Also, some strategies predict ligand-binding sites only for single-chain structures, which means that they are not able to predict binding sites involving multiple chains, for instance in multimers and biological assemblies. Another important aspect is the lack of interpretability of some methods that work as black boxes. In techniques based on machine learning, for example, it is interesting to be able to point out the most relevant descriptors for the prediction, which can give insights to domain specialists.

To overcome these challenges, in this work, we propose Graph-based Residue neighborhood Strategy to Predict binding sites (GRaSP), which is a supervised learning strategy that represents a particular residue and its neighbors as a graph (also called network) at the atomic level to perceive residue environment information. For each residue of a protein, topological and physicochemical properties of its atoms and interactions are represented as a graph, which is encoded as a feature vector. The set of feature vectors that represents a protein serve as input for the supervised learning strategy. Our method is not based on sequence alignment nor structural alignment. GRaSP is residue centric, scalable and able to find binding sites across multiple chains. Also, it performs well when predicting binding sites for bound/unbound structures. One important aspect of GRaSP is the simplicity of the model, as the descriptors are interpretable and can be inspected to support users on the understanding of predictions. Our method shows good performance as it takes 10–20 s on average to predict a protein binding site, outperforming the

top compared method, that takes 2–5 h. GRaSP achieves compatible or superior results when compared to state-of-the-art methods. All the source code and datasets are freely available.

2 Materials and methods

This section details GRaSP, our supervised learning strategy based on neighborhood graphs to predict ligand-binding sites. We explain the problem modeling, the datasets used in experimental evaluation as well as the evaluation strategy. A workflow that summarizes GRaSP is presented in Figure 1.

2.1 GRaSP

2.1.1 Graph model

GRaSP is a supervised learning strategy that represents a particular residue and its structural neighbors as a graph to perceive residue environment information. For each residue of a protein, topological and physicochemical properties of its atoms and interactions are represented as a graph, which is encoded as a feature vector.

Given a residue from a protein structure, the non-covalent interactions established by this residue and its two first shells of neighbor residues are calculated based on physicochemical properties of atoms and distance criteria as in Fassio et al. (2017, 2019).

Figure 2 provides a schematic representation of a protein residue and its two shells of neighbor residues. So, for the residue 1 in Figure 2a, we calculate its relative solvent accessibility, the physicochemical properties of its atoms, the interactions established (by these atoms) and their types. The set of residues that interact with residue 1 in Figure 2b is considered as the first shell of neighbors (N1).

For each residue of the shell N1, we compute the same residue descriptors that are calculated for residue 1 in Figure 2a. Then we sum the values of each specific descriptor and divide it by the number of residues in the shell N1, averaging the values. Thus, for the shell N1, we have exactly the same residue descriptors of residue 1 in Figure 2a, but they represent a summary of the shell N1.

The set of residues that interact with residues of the first shell of neighbors (N1) is considered as the second shell of neighbors (N2). For each residue in N2, the same residue descriptors of residue 1 in Figure 2a are calculated. Then we add up the values of each descriptor and divide by the number of residues in N2, in a way that these descriptors represent a summary of the shell N2. To calculate interactions Euclidean distance was used. For details on atom types and distance criteria see Supplementary Tables S1 and S2.

Our graph model captures the physicochemical properties of the structural environment of each residue. The next step is to represent this model as a matrix for the subsequent supervised learning task.

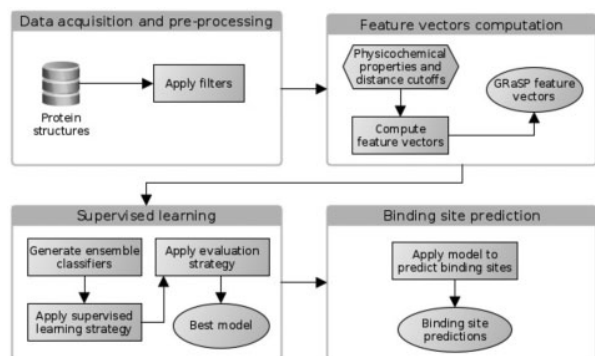


Fig. 1. The GRaSP workflow is segmented into four blocks: Data acquisition and pre-processing; Feature vectors computation; Supervised learning and Binding site prediction. Rectangles denote processing steps; ellipsoids represent output files; and hexagons are input files or parameters

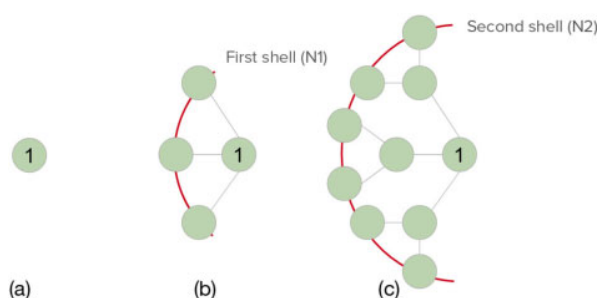


Fig. 2. GRaSP modeling. This scheme represents a residue and its neighborhood. (a) The residue being considered, which we named residue 1. (b) The first shell of neighbors of residue 1. (c) The second shell of neighbors of residue 1

2.1.2 Feature vectors

In this point, it is important to detail the set of descriptors calculated to characterize residues and how we generate a feature vector to represent each protein residue. The set of 14 types of descriptors calculated can be segmented in:

- Residue level: *solvent relative accessibility* is calculated using Naccess (Hubbard and Thornton, 1993).
- Atom level: atoms are labeled as *aromatic, acceptor, donor, hydrophobic, positive, negative*. Then, we compute how many atoms of each type each residue presents.
- Interaction level: non-covalent interactions are calculated based on the type of atoms and on the Euclidean distance between them using a kd-tree implementation from Biopython (Cock et al., 2009). In this way, we can avoid computing distances between all atom pairs in a protein, which is computationally expensive. The interaction types calculated are *aromatic stacking, disulfide bridge, hydrogen bond, hydrophobic, repulsive* and *salt bridge*.

The problem of predicting ligand-binding sites was modeled as follows. Given a residue r from a protein structure P , a feature vector is then generated for each residue of P based on the mentioned descriptors. Hence, to represent a whole protein we have a set of feature vectors. In short, to represent a set of protein structures we have a matrix, let say, G , in which each row encodes a residue, and each column represents a descriptor. This matrix serves as input for our supervised learning strategy, GRaSP, which aims to predict, for each residue, if it is in the binding site or not. Figure 3 presents a fragment of a matrix G . For example, in Figure 3d, the residue r is PRO: A:3 (proline numbered 3 in chain A), from the protein P , which is entry 1A59 from PDB; the sixth column, HB, represents how many hydrogen bonds this residue establishes.

The feature vector for a residue is represented by three sets of descriptors: (i) the first 14 refer to a protein residue (residue 1 in Fig. 2); (ii) the next 14 summarize the first shell of neighbors (N1 in Fig. 2); (iii) the last 14 summarize the second shell of neighbors (N2 in Fig. 2). Thus, there are 42 descriptors to represent each residue.

In short, we have a matrix, G , that represents the whole set of protein structures. Each row encodes a residue, and each column represents a property of the residue neighborhood. This matrix serves as input for a supervised learning strategy, which aims to predict, for each residue, if it is in the binding site or not.

2.2 Data

2.2.1 Datasets

Six datasets were used in the experimental evaluation of our method to show its generality, applicability in real-world scenario and to compare GRaSP with state-of-the-art methods.

1. COACH: a benchmark set from COACH (Yang et al., 2013) consisting of 500 non-redundant single-chain structures that

(a)	Residue being considered														First shell			Second shell			
	Residue	Interactions							Atoms						Residue	Interactions	...	Atoms	...		
(b)	res_name	Acc_rel	CYS	AS	DB	HB	HY	RP	SB	ARO	ACC	DON	HYD	POS	NEG	Acc_rel	AS	...	NEG	...	class
(c)	1a59_A_PRO_3	65.4	0	0	0	1	0	0	0	0	1	0	2	0	0	73.40	0.00	...	0.48	...	0
1a59_A_GLU_2	112.7	0	0	0	0	0	0	4	0	0	3	1	2	0	2	60.47	0.00	...	0.59	...	0
1a59_A_HIS_186	1.6	0	3	0	2	0	12	2	5	3	3	3	1	2	0	7.82	0.18	...	0.13	...	1
1a59_A ASN_189	3.7	0	0	0	4	0	0	0	0	2	2	2	1	0	0	12.37	0.23	...	0.09	...	1

Fig. 3. GRaSP feature vectors. (a) Starting top down, the first row is composed by three high-level columns, regarding the residue being considered, the first and second shell of neighbors respectively. (b) In the second row, each column groups descriptors in residue level, interaction level and atom level for each high-level column mentioned in (a). (c) The third row is segmented in columns: the identifier of each residue (*res_name*); In the residue level, we have relative accessibility (*Acc_rel*) and cysteine (*CYS*); In the interaction level, we have aromatic stacking (*AS*); disulfide bridge (*DB*); hydrogen bond (*HB*); hydrophobic (*HY*); repulsive (*RP*) and salt bridge (*SB*); In the atom level, we have aromatic (*ARO*); acceptor (*ACC*); donor (*DON*); hydrophobic (*HYD*); positive (*POS*); negative (*NEG*); We have the same descriptors for the first and second shell, which totals 42 descriptors. The last column represents whether the residue is in the binding site or not (*class*). (d) Hereafter each row represents a residue and each column represents a residue descriptor

contain natural, drug-like and metal ligands. This dataset is used to compare our method and COACH.

2. **CASP 10 dataset:** composed of 13 target structures and their correspondent 25 template structures (Gallo Cassarino *et al.*, 2014). This dataset is used to compare GRaSP with other 17 state-of-the-art methods that participated in CASP 10 (for details about the 13 target structures and the 17 competing methods, see Supplementary Tables S3–S6).
3. **CAMEO dataset:** composed of 31 targets obtained from CAMEO (Ligand Binding—LB—Targets for 1-week). For each target, we built a training dataset that consists of PDB structures with sequence similarity up to 75%. The training dataset is composed of 15 595 proteins (for details about the 31 target structures and their similar proteins, see Supplementary Tables S7 and S8). The purpose of this dataset is to compare our method with the RaptorX-Binding, the method that focuses on prediction of ligand-binding residues in the CAMEO initiative.
4. **B44/U44:** a dataset derived from (Krivák and Hoksza, 2018) that contains 44 protein structures in a bound and unbound state. Due to inconsistencies with PDB format, 4 entries were removed from the original dataset (contains 48 entries). This dataset is used to compare how our method performs on bound and unbound structures.
5. **HOLO4K:** a large benchmark of 4543 protein–ligand complexes obtained from (Krivák and Hoksza, 2018) that contains multi-chain structures from PDB. This dataset is used to show that GRaSP is able to detect binding site residues in multi-chain structures.
6. **ASTEX:** a docking validation dataset containing 85 high-quality experimental structures of drug-like complexes obtained from (Hartshorn *et al.*, 2007). This set includes diverse ligands, with distinct molecular recognition types, and it is used to show how our method perform on relevant and diverse drug-like targets.
7. **BioLip:** a semi-manually curated dataset of biologically relevant protein–ligand interactions (Yang *et al.*, 2012) that contains 71 448 non-redundant complexes. This dataset is used as training data by our method to build its machine learning model.

2.3 Experiment design

2.3.1 Ensemble classifier

Data imbalance is intrinsic to the problem of predicting binding site, as considering a protein, the number of non-binding site residues is greater than the number of binding site residues. Thus an ensemble classifier was devised to handle imbalanced data in GRaSP. Figure 4 presents a scheme of this ensemble classifier.

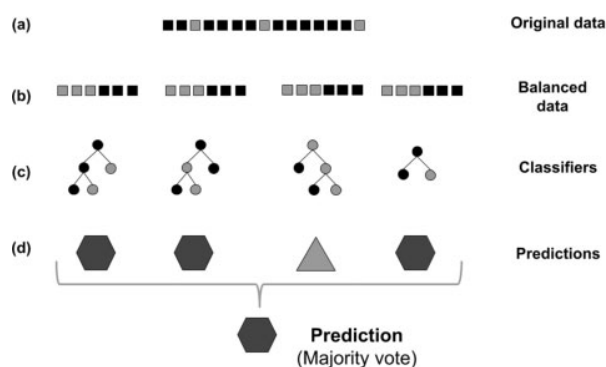


Fig. 4. Ensemble classifier scheme. (a) Imbalanced dataset. (b) Original dataset segmented in balanced partitions. Each partition contains the same set of binding site examples with approximately the same number of non-binding site examples. (c) Set of classifiers in the ensemble. Each balanced partition is the input training data for one classifier. (d) Voting scheme. To use the ensemble for prediction purposes, an unseen data instance is submitted to all classifiers. The resulting class is the one with the majority of votes

Consider the set of all residues from proteins in the training dataset that will be used for building a predictor. We can partition this set in two disjoint sets: the data instances that represent binding site residues B (light gray squares in Fig. 4a) and the data instances that represent non-binding site residues N (black squares in Fig. 4a). Due to the imbalance of classes, N is segmented in n partitions, T_1, T_2, \dots, T_n , whose size is approximately $|B|$. Each of these negative partitions is merged to the single positive partition B , as shown in Figure 4b. Then n classifiers are built so that for each one the input data is $E_i = B \cup T_i$, with $i = 1, \dots, n$, as shown in Figure 4c.

Finally, for prediction purposes, when the classifier is used in a real-world scenario, we use a voting scheme to determine if the residue is part of the binding site or not. The Figure 4d shows a majority vote scheme, in which each data instance unseen for GRaSP is queried against all n classifiers and the final answer is the most voted option (binding site or non-binding site). The classification algorithm applied was the Extremely randomized tree (Geurts *et al.*, 2006) from scikit-learn (Pedregosa *et al.*, 2011).

2.4 Evaluation strategy

We evaluate GRaSP and compare it with state-of-the-art methods using a set of robust metrics, which can be used with imbalanced data, especially in our case, as we are more interested in one of the classes (the binding site instead of nonbinding site).

Precision: the fraction of data instances that are actually positive in the group that the classifier predicted as positive. Precision ranges

from 0 to 1. The higher the precision, the lower the number of false positives ($p = TP/(TP + FP)$).

Recall: also called sensitivity or true positive rate, it measures the fraction of positive data instances that were predicted as positive by the classifier over the total amount of positive instances. Recall ranges from 0 to 1. The higher the recall, the lower the number of false negatives ($r = TP/(TP + FN)$).

MCC: Matthews correlation coefficient (MCC) is a quality measure for binary classification that can be used even with imbalanced data. It is a correlation coefficient between observed and predicted classifications that ranges from -1 to +1. MCC=-1 represents inverse predictions, MCC=0 represents random predictions and MCC = +1 represents the ideal classifier, with correct predictions. Equation 1 shows how to calculate MCC.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{(\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)})} \quad (1)$$

BDT: binding distance test (BDT) score ranges from 0 to 1 and was proposed by Roche *et al.* (2010) to take into account the distance between predicted and observed binding site residues. Hence, correct predictions achieves 1 and predictions distant from the observed binding site achieve scores closer to 0. To compute BDT, the S-score, S_{ij} , between predicted and observed residues i and j must be calculated first, using Equation 2, where d_{ij} is the Euclidean distance between residues i and j and d_0 is a distance threshold (recommended values are between 1 and 3 Å). BDT can be calculated as shown in Equation 3, where N_p is the number of predicted residues and N_0 is the number of observed residues.

$$S_{ij} = \frac{1}{1 + \left(\frac{d_{ij}}{d_0}\right)} \quad (2)$$

$$BDT = \frac{\sum_{i=1}^{N_p} \max(S_{ij})}{\max(N_p, N_0)} \quad (3)$$

DCA: distance between the center of the pocket and any ligand atom is a pocket-centric metric that considers a binding site as correctly predicted if the distance between its center and any atom of a ligand is below a threshold (Chen *et al.*, 2011). Previous works consider 4 Å as appropriate threshold.

For each experiment conducted, we used the same set of metrics as the competitors. Therefore, to compare our strategy with: (i) residue-centric methods, the metrics precision, recall and MCC were applied; (ii) 17 methods of CASP 10 competition, the main metric was MCC; (iii) a method of CAMEO initiative, MCC and BDT were used; (iv) pocket centric methods, DCA was employed; (v) diverse and popular datasets, precision, recall and MCC were selected; (vi) the state-of-the-art method in a controlled dataset, precision, recall and MCC were used.

3 Results and discussion

To validate the ability of our method in predicting binding sites we conducted a comprehensive set of experiments. First, we compare GRASP with other six residue-centric methods, including the one considered as the state-of-the-art. Then, we compare GRASP with

17 methods that participated in CASP10 and also with one method in CAMEO initiative. Next, we present comparative results regarding pocket-centric methods. After that, we show the performance of our method on datasets including drug-like complexes, same proteins on bound/unbound state and multi-chain structures. Finally, we present a comparison with the state-of-the-art method for a set of proteins on bound/unbound state.

3.1 GRASP results compared with state-of-the-art methods

3.1.1 Residue-centric methods

First, it is important to point out the main differences between COACH, considered the state-of-the-art method, and GRASP. COACH is a hybrid method that uses sequence and structure based approaches. Moreover, it is a consensus method that combines predictions of five other ligand-binding site predictors, two of them are in-house tools, TM-SITE and S-SITE, and the other three are the third-party tools COFACTOR, FINDSITE and ConCavity. These five methods are combined through a machine learning strategy to predict binding sites for single-chain structures. GRASP is a novel machine learning structure-based strategy that models each residue and its environment as a graph. This graph captures physicochemical properties of the structural environment of each residue and is encoded as a feature vector that serve as input for a classification strategy that points out binding residues for multi-chain structures.

To compare GRASP with COACH, we used the benchmark dataset COACH (Yang *et al.*, 2013) (dataset 1), a set of 500 non-redundant single-chain proteins. We compared our method with COACH consensus prediction and, also, with all the methods that compose COACH (TM-SITE, S-SITE, COFACTOR, FINDSITE and ConCavity), using the results reported in Liu *et al.* (2020). GRASP was run for each of the 500 structures, and, for all of them, BioLip (dataset 7) was used as training data.

Table 1 summarizes the results obtained. GRASP achieved better results in comparison with other methods, considering MCC, which is a correlation coefficient between observed and predicted classifications. MCC is the main measure used by COACH authors to compare their strategy with competitors as, according to them, MCC combines accuracy and coverage of the prediction presenting a better balance of both than precision and recall used individually (Yang *et al.*, 2013).

For completeness, we also present precision and recall. Our method outperformed competitors in terms of precision and, in terms of recall, COACH achieved better results. A possible interpretation of this result is that our method does not recover some of the binding site residues. However, among the residues it labels as binding site, we can be more confident in saying that they are actually part of the binding site because GRASP presents highest precision.

3.1.2 CASP 10 methods

GRASP was compared with the 17 methods submitted to the FN category of the CASP 10 competition. The dataset has 13 target enzymes and 25 binding site templates for each target. In this experiment, the templates of each target were used as training data to build one classification model to predict binding sites for each target, so in total, we have 13 models (for details about binding

Table 1. Comparative results for GRASP and COACH methods

	TM-SITE	S-SITE	COFACTOR	FINDSITE	ConCavity	COACH	GRASP
MCC	0.51	0.45	0.46	0.44	0.33	0.60	0.61
Precision	0.59	0.45	0.61	0.45	0.26	0.59	0.69
Recall	0.51	0.58	0.41	0.51	0.62	0.70	0.61

Note: GRASP was compared with COACH consensus predictions and to the methods that compose COACH (TM-SITE, S-SITE, COFACTOR, FINDSITE and ConCavity). The highest value for each metric was highlighted in bold.

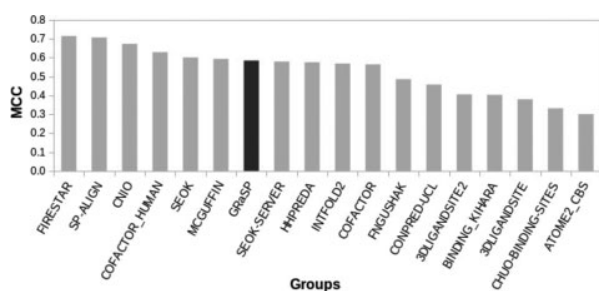


Fig. 5. Groups from CASP 10 (FN category) ranked in decreasing order by average MCC together with GRaSP

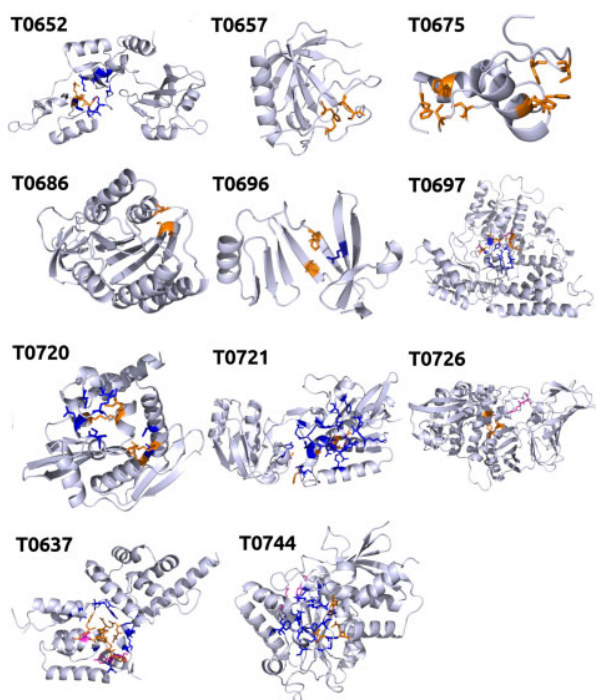


Fig. 6. In orange, we see residues observed to be part of the binding site and that were predicted as being binding site by GRaSP. In magenta, we see residues that GRaSP predicted as a binding site but were not observed to be so. Observed binding site residues that were not predicted by GRaSP are presented in blue. Targets T0659 and T0697 are not presented as GRaSP did not identify binding residues in these structures

residues from CASP 10 and binding residues found according to GRaSP, see [Supplementary Table S3](#)).

Figure 5 provides an overview of the performance of GRaSP and the 17 methods that participated in CASP 10 function prediction (FN) category. Methods are ranked according to average MCC standardized on each target. GRaSP was ranked in the seventh position, with average MCC 0.58 (according to CASP 10, differences among the first 10 methods were not statistically significant). Figure 6 presents GRaSP predictions for the targets of CASP 10 dataset.

For 2 targets, T0659 and T0706, our strategy did not identify binding residues, so their MCC was set to 0 according to CASP guidelines. Regarding T0659, there was only one protein template with just one ligand (a phosphate), which suggests that the set of templates for this target do not provide enough examples to train our predictor. To overcome this limitation, we used the templates of all 13 targets to build a model to predict the binding site for this target, which resulted in a correct prediction with MCC 1. If we considered this prediction, GRaSP would be ranked fourth (MCC 0.66) among CASP 10 participating methods. For target T0706, there

were 20 templates with ligands that could be used as training data for GRaSP. Nonetheless, even using the templates of all the 13 targets to build 1 model to try to predict the binding site for T0706 our method was not able to do so. In case we disregard the two targets GRaSP did not identify binding site residues, our strategy would be ranked third (MCC 0.69) among participating methods.

In accordance with FN category in CASP 10, targets T0657 and T0659 were the most challenging ones as participating methods obtained on average the lowest MCC values. GRaSP was able to predict the binding site for T0657 (MCC 1). For target T0659, as mentioned, MCC was set to 0 but our strategy is able to predict the binding site in case we use all templates (of all targets) as training data.

It is important to mention that in this experiment we used only templates provided by CASP 10 to build our predictor, while other tools make use of curated databases, coupled with sequence and structural alignments. Firestar, for instance, the top method in the assessment, uses the functional information in FireDB, as well as sequence and structural alignments to make predictions of ligand-binding residues and catalytic residues.

In addition, for each target, we used only its own templates as training data. In case we use different experiment design for each target, we can achieve better results. For example, if we use SVM, we have better MCC for target T0744 (MCC 0.39 instead of 0.26). If we use the whole template dataset to train a single classifier, targets T0659 and T0737 perform better. However, we decided to be more conservative, as we believe that for a fair comparison, we need to perform our experiments in a similar manner to CASP 10 participating tools. This means to consider one experimental design for all targets, without knowing the correct predictions in advance, which could allow fine tuning of parameters. We calculated the MCC, the cumulative confusion matrices, the MCC Z-scores and BDT in a similar manner to CASP 10 participants ([Supplementary Tables S4–S6](#)).

3.1.3 CAMEO

GRaSP was compared with the RaptorX-Binding method submitted to the LB category of the CAMEO competition. This choice is due to the fact that RaptorX-Binding resembles GRaSP, as it makes binary predictions (both predict if a residue is in the binding site or not).

In this experiment, proteins similar to each of the 31 target structures were used as training data to build one classification model to predict binding sites for each target. [Supplementary Table S7](#) shows the target structures and the number of similar proteins, and [Supplementary Table S8](#) shows the similar proteins identifiers.

Some targets in CAMEO—LB were incomplete or just informing the *P*-value of the residues as the predictions. Since RaptorX-Binding and GRaSP provide binary results, we used ProLigContact software ([Gallo Cassarino et al., 2014](#)) to define the residues that are considered as the observed binding sites for the 31 target structures. [Supplementary Table S10](#) shows all the 31 targets with binding residues according to RaptorX-Binding, ProLigContact and GRaSP.

[Supplementary Table S11](#) shows GRaSP and RaptorX-Binding results. GRaSP has better average values for MCC and BDT (0.656 and 0.632) than RaptorX-Binding (0.557 and 0.546). The true positives (TP) value for GRaSP (438) is significantly higher compared with RaptorX-Binding (365), and the false negatives (FN) value (29) is lower than RaptorX-Binding (102). [Supplementary Table S12](#) shows MCC and BDT values for all 31 target structures for both methods. Although in this experiment it was not possible to use the target directly from CAMEO, the results demonstrate the robustness and cohesion of our method.

3.2 GRaSP results compared with pocket-centric methods

As previously mentioned, structure-based methods to predict ligand-binding sites focus on pocket prediction or ligand-binding residue prediction. GRaSP is a residue-centric method that models the

ligand-binding site problem as a binary classification which aims to predict, for each residue, if it is in the binding site or not. Among the predictors centered on pockets, there are some popular methods, such as Fpocket (Schmidtke et al., 2010), Metapocket 2.0 (Zhang et al., 2011) and Sitehound (Hernandez et al., 2009), with emphasis on P2Rank (Jendele et al., 2019; Krivák and Hoksza, 2018) and DeepSite (Jiménez et al., 2017), which are state-of-the-art methods in the pocket-centric category. Thus a natural question that arises is how the predictions of GRaSP can be compared with the predictions of these methods.

The set of residues predicted as positive (residues that are part of the binding site) by GRaSP were clustered using DBSCAN (Schubert et al., 2017), a density-based clustering algorithm that groups together points in areas of high density separated by areas of low density. When residues predicted as positives form a region of high density, this region is considered as a pocket. Low density regions means far apart residues that were not taken into account in this assessment. Thus, to compare GRaSP pockets to those of pocket-centric methods, each group computed by DBSCAN was seen as a pocket.

To evaluate pocket predictors, we use the DCA measure, as it was applied in the state-of-the-art method P2Rank. This measure represents the distance from the center of the predicted pocket to any atom of the ligand (Chen et al., 2011). Given a protein with n binding sites, for every method considered we take n pocket predictions. A specific binding site is correctly predicted if the minimal distance between its corresponding ligand and any of the n predictions from a given method is below a threshold. Previous works assume that a predicted site is correct if its center is no farther than 4 Å to any atom of the ligand (Chen et al., 2011; Jendele et al., 2019).

The predictive performance of GRaSP was compared with competing methods using the datasets COACH420 and HOLO4K. These datasets, as well as the predictions of considered methods, were obtained from (Jendele et al., 2019). COACH420 consists of 420 single chains and was derived from COACH dataset. Some methods were not able to generate results for all entries in each dataset. Thus, a subset of 228 chains from COACH420 and 1471 entries from HOLO4K was used in this experiment because for this subset all competing methods generated results. As these datasets include biologically non-relevant ligands, such as the glycol molecule that is introduced by purification and crystallization procedures, we determined which ligands are relevant as in Jendele et al. (2019).

Figure 7 shows the success rate of the predictors, which is defined as the number of correctly predicted binding sites divided by the total number of binding sites. To avoid an arbitrary threshold selection, the success rates of the methods were computed in a range between 1 and 20 Å. P2Rank outperforms the other methods, but GRaSP is in the second position from the 4 Å threshold onwards, which is the value suggested in previous works. Having in mind that GRaSP was not devised with the aim of predicting pockets that are potential

Table 2. GRaSP results for diverse datasets

	ASTEX	B44	U44	HOLO4K
MCC	0.66	0.67	0.67	0.61
Precision	0.74	0.64	0.61	0.68
Recall	0.65	0.77	0.80	0.58

Note: Performance on datasets including drug-like complexes (ASTEX), same proteins on bound/unbound state (B44/U44) and multi-chain structures (HOLO4K).

binding sites, but with the aim of pointing out residues that are part of the binding site, we believe this result is very satisfactory.

3.3 GRaSP performance on diverse datasets

Previous experiments were conducted to compare GRaSP with residue-centric and with pocket-centric methods. This section comprises experiments to show how GRaSP performs on some popular datasets. Our method proved to be consistent in predicting binding sites for datasets with different characteristics, as shown in Table 2. The time required to process each protein complex in this experiment was between 10 and 20 s.

Our method was run for each of the 85 proteins in ASTEX dataset, resulting in an MCC 0.66. This set is particularly important due to its drug-like targets. In this scenario GRaSP reached a precision of 0.74, which is an expressive value in terms of residue binding site prediction. That means a low occurrence of false positives, which is very important with drugable binding sites, avoiding unnecessary docking experiments in negative regions.

Next, GRaSP performance was compared on 44 bound and unbound structures using B44/U44 (dataset 4), which contains the exact same protein structures in a bound and unbound state. GRaSP achieved the very similar results for both states (MCC 0.67). We believe it is a positive aspect of our method as detecting binding sites on unbound structures tend to be more challenging. Many proteins in their unbound state do not present surface pockets with appropriate size for drug binding (Vajda et al., 2018). We hypothesize that, as our strategy considers the residue environment, even though the protein structure presents different folding due to the unbound state, the properties of the residue environment remain similar to the bound state.

Finally, we used GRaSP to point out binding site residues for a set of 4543 multi-chain structures in HOLO4K (dataset 5). Our strategy achieved MCC 0.61, which we consider a satisfactory result taking into account that the state-of-the-art method is able to deal only with single-chain structures and achieved MCC 0.60 in its own benchmark dataset of 500 single-chain proteins.

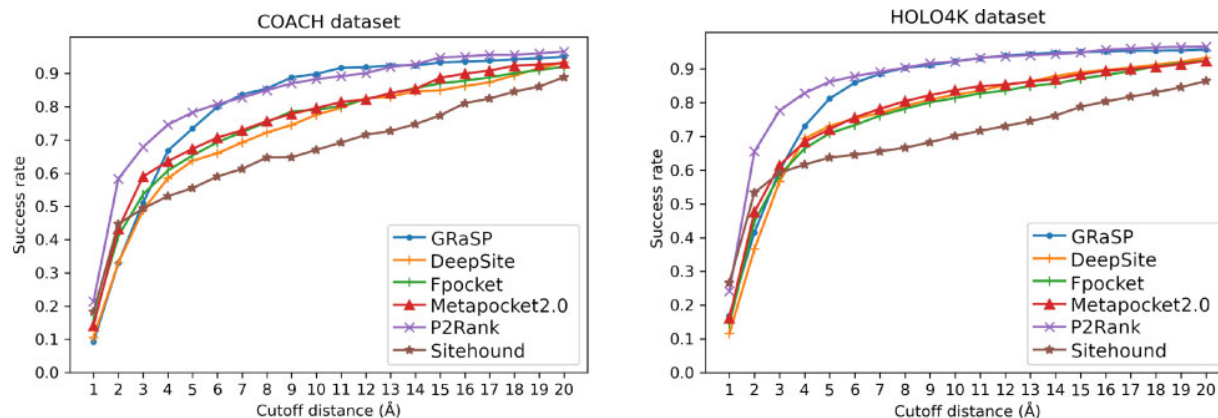


Fig. 7. Predictive performance of GRaSP compared with pocket-centric methods. GRaSP was compared with Fpocket, Metapocket 2.0, Sitehound, P2Rank and DeepSite on datasets COACH420 and HOLO4K with thresholds ranging from 1 to 20 Å

Table 3. GRaSP compared against COACH

Dataset	Method	MCC	Precision	Recall
B44	GRaSP	0.67	0.64	0.77
	COACH	0.64	0.59	0.77
U44	GRaSP	0.67	0.61	0.80
	COACH	0.67	0.64	0.75

Note: Performance on bound and unbound structures (B44/U44 dataset).

3.4 GRaSP x COACH

We also compared GRaSP with COACH using B44/U44 dataset. Table 3 presents average values of MCC, precision and recall for both methods (for detailed results, see the Supplementary Tables S13–S16).

It is interesting to notice that GRaSP presents results comparable to COACH. In addition, on average, our method took 12 s to process each structure in this experiment, and the processing of each dataset (B44 and U44) took less than 9 minutes. The time for each entry of the dataset to be processed is provided in Supplementary Tables S13 and S15.

When a task is submitted to COACH, the server reports that it can take up to 4 h to finish. Nevertheless, some structures took more than 4 h and, in some cases, this time was longer than a week. For instance, the server took more than a week to complete the processing of the PDB entries 1DWD, 1NPC and 1PDY. Moreover, it was unable to finish processing the structure 1IDA (for details, see Supplementary Tables S14 and S16).

Unlike GRaSP, if the numbering of residues in the submitted protein structure is not continuous or does not start from 1, COACH renames these residues and presents its result based on this new numbering. COACH works for single-chain structures only, whereas GRaSP works for both single and multi-chain structures.

4 Conclusion

In this work, we proposed a novel, residue centric, graph-based scalable method to predict ligand-binding site residues, which we named GRaSP. It is a supervised learning strategy that depicts each protein residue and its neighbors as a graph at the atomic level as a means to capture residue environment properties. For each protein residue, physicochemical and topological properties of its atoms and non-covalent interactions are modeled as a graph which, in turn, is encoded as a feature vector. A matrix of feature vectors representing a set of proteins is the input for GRaSP. Our method is not based on sequence alignment nor superimposition of structures and it is able to predict binding sites across multi-chain proteins.

Results show that our strategy presents comparable or superior predictions when compared to methods considered state-of-the-art. GRaSP achieved better results when compared with six other residue-centric methods. In addition, our method outperformed the RaptorX-Binding, the method from CAMEO independent assessment that resembles GRaSP. When compared with five state-of-the-art pocket-centric methods, GRaSP ranked second, which we consider a significant result, as it was not devised with the purpose of predicting pockets that are potential binding sites. Finally, our method proved scalable as it took 10–20 s on average to predict the binding site for a protein complex whereas the state-of-the-art residue-centric method takes 2–5 h on average.

As future work, we intend to make non-binary predictions (values between 0 and 1), which are more informative, as they represent how confident our method is about a prediction. In addition, we intend to make GRaSP available as a web-server (which will make it accessible for more general users), and also as web services, which allow users to include our method in their own research pipeline.

Funding

This study was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 23038.004007/2014-82 grant 051/2013, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and European Bioinformatics Institute (facilities and support).

Conflict of Interest: none declared.

References

- Brylinski, M. and Skolnick, J. (2008) A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. USA*, **105**, 129–134.
- Capra, J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
- Chen, K. *et al.* (2011) A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure*, **19**, 613–621.
- Cock, P.J. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
- Consortium, U. (2019) Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
- Ding, Y. *et al.* (2017) Identification of protein–ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.*, **57**, 3149–3161.
- Duan, L. *et al.* (2016) Interaction entropy: a new paradigm for highly efficient and reliable computation of protein–ligand binding free energy. *J. Am. Chem. Soc.*, **138**, 5722–5728.
- El-Gebali, S. *et al.* (2019) The pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
- Fassio, A.V. *et al.* (2017) Vermont: a multi-perspective visual interactive platform for mutational analysis. *BMC Bioinformatics*, **18**, 403.
- Fassio, A.V. *et al.* (2019) nAPOLL: a graph-based strategy to detect and visualize conserved protein–ligand interactions in large-scale. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 4, pp. 1317–1328, 1 July–Aug. 2020, doi: 10.1109/TCBB.2019.2892099.
- Fuller, J.C. *et al.* (2015) LigDig: a web server for querying ligand–protein interactions. *Bioinformatics*, **31**, 1147–1149.
- Gallo Cassarino, T. *et al.* (2014) Assessment of ligand binding site predictions in casp10. *Proteins Struct. Funct. Bioinf.*, **82**, 154–163.
- Geurts, P. *et al.* (2006) Extremely randomized trees. *Mach. Learn.*, **63**, 3–42.
- Gheris, D. and Sanchez, R. (2009) EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics*, **25**, 3185–3186.
- Haas, J. *et al.* (2018) Continuous automated model evaluation (CAMEO) complementing the critical assessment of structure prediction in casp12. *Proteins Struct. Funct. Bioinf.*, **86**, 387–398.
- Hartshorn, M.J. *et al.* (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *J. Med. Chem.*, **50**, 726–741.
- Hernandez, M. *et al.* (2009) SiteHound-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **37**, W413–W416.
- Hubbard, S.J. and Thornton, J.M. (1993) 'NACCESS', computer program. *Technical report*, Department of Biochemistry Molecular Biology, University College London.
- Izidoro, S.C. *et al.* (2015) GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics*, **31**, 864–870.
- Jendele, L. *et al.* (2019) PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Res.*, **47**, W345–W349.
- Jiménez, J. *et al.* (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, **33**, 3036–3042.
- Komiyama, Y. *et al.* (2016) Automatic generation of bioinformatics tools for predicting protein–ligand binding sites. *Bioinformatics*, **32**, 901–907.
- Krivák, R. and Hoksza, D. (2018) P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminf.*, **10**, 39.
- Le Guilloux, V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
- Liu, Y. *et al.* (2020) CB-Dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacol. Sin.*, **41**, 138–144.

- Lopez, G. et al. (2007) FireDB—a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, **35**, D219–D223.
- López, G. et al. (2007) firestar—prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.
- Lopez, G. et al. (2011) firestar—advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
- Macari, G. et al. (2019) Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies. *J. Comput. Aided Mol. Des.*, **33**, 887–903.
- Moraes, J.P. et al. (2017) GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Res.*, **45**, W315–W319.
- Nitsche, C. and Otting, G. (2018) NMR studies of ligand binding. *Curr. Opin. Struct. Biol.*, **48**, 16–22.
- Pedregosa, F. et al. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Porter, C.T. et al. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
- Roche, D. et al. (2015) Proteins and their interacting partners: an introduction to protein–ligand binding site prediction methods. *Int. J. Mol. Sci.*, **16**, 29829–29842.
- Roche, D.B. et al. (2010) The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics*, **26**, 2920–2921.
- Roche, D.B. et al. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.
- Roche, D.B. et al. (2013) The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Res.*, **41**, W303–W307.
- Roy, A. et al. (2012) Cofactor: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–W477.
- Schmidt, T. et al. (2011) Assessment of ligand-binding residue predictions in casp9. *Proteins Struct. Funct. Bioinf.*, **79**, 126–136.
- Schmidtke, P. et al. (2010) Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Research*, **38**, W582–W589.
- Schubert, E. et al. (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst. (TODS)*, **42**, 1–21.
- Schwede, T. (2013) Protein modeling: what happened to the protein structure gap? *Structure*, **21**, 1531–1540.
- Somody, J.C. et al. (2017) Structural coverage of the proteome for pharmaceutical applications. *Drug Discov. Today*, **22**, 1792–1799.
- Vajda, S. et al. (2018) Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.*, **44**, 1–8.
- Warshel, A. and Bora, R.P. (2016) Perspective: defining and quantifying the role of dynamics in enzyme catalysis. *J. Chem. Phys.*, **144**, 180901.
- Wu, Q. et al. (2018) COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.*, **46**, W438–W442.
- Yang, J. et al. (2012) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
- Yang, J. et al. (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.
- Zhang, Z. et al. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.

Apêndice B

Informações adicionais

B.1 Propriedades físico-químicas dos átomos

Res.	Atom	Properties	Res.	Atom	Properties	Res.	Atom	Properties
ALA	N	DON	HIS	N	DON	PRO	N	-
ALA	CA	-	HIS	CA	-	PRO	CA	-
ALA	C	-	HIS	C	-	PRO	C	-
ALA	O	ACP	HIS	O	ACP	PRO	O	ACP
ALA	CB	HPB	HIS	CB	HPB	PRO	CB	HPB
ARG	N	DON	HIS	CG	ARM	PRO	CG	HPB
ARG	CA	-	HIS	ND1	ARM, POS, DON, ACP	PRO	CD	-
ARG	C	-	HIS	CD2	ARM	SER	N	DON
ARG	O	ACP	HIS	CE1	ARM	SER	CA	-
ARG	CB	HPB	HIS	NE2	ARM, POS, DON, ACP	SER	C	-
ARG	CG	HPB	ILE	N	DON	SER	O	ACP
ARG	CD	-	ILE	CA	-	SER	CB	-
ARG	NE	POS, DON	ILE	C	-	SER	OG	DON, ACP
ARG	CZ	POS	ILE	O	ACP	THR	N	DON
ARG	NH1	POS, DON	ILE	CB	HPB	THR	CA	-
ARG	NH2	POS, DON	ILE	CG1	HPB	THR	C	-
ASN	N	DON	ILE	CG2	HPB	THR	O	ACP
ASN	CA	-	ILE	CD1	HPB	THR	CB	-
ASN	C	-	LEU	N	DON	THR	OG1	DON, ACP
ASN	O	ACP	LEU	CA	-	THR	CG2	HPB
ASN	CB	HPB	LEU	C	-	TRP	N	DON
ASN	CG	-	LEU	O	ACP	TRP	CA	-
ASN	OD1	ACP	LEU	CB	HPB	TRP	C	-
ASN	ND2	DON	LEU	CG	HPB	TRP	O	ACP
ASP	N	DON	LEU	CD1	HPB	TRP	CB	HPB
ASP	CA	-	LEU	CD2	HPB	TRP	CG	HPB, ARM
ASP	C	-	LYS	N	DON	TRP	CD1	ARM
ASP	O	ACP	LYS	CA	-	TRP	CD2	HPB, ARM
ASP	CB	HPB	LYS	C	-	TRP	NE1	ARM, DON
ASP	CG	-	LYS	O	ACP	TRP	CE2	ARM
ASP	OD1	NEG, ACP	LYS	CB	HPB	TRP	CE3	HPB, ARM
ASP	OD2	NEG, ACP	LYS	CG	HPB	TRP	CZ2	HPB, ARM
CYS	N	DON	LYS	CD	HPB	TRP	CZ3	HPB, ARM
CYS	CA	-	LYS	CE	-	TRP	CH2	HPB, ARM
CYS	C	-	LYS	NZ	POS, DON	TYR	N	DON
CYS	O	ACP	MET	N	DON	TYR	CA	-
CYS	CB	HPB	MET	CA	-	TYR	C	-
CYS	SG	DON, ACP, SSB	MET	C	-	TYR	O	ACP
GLN	N	DON	MET	O	ACP	TYR	CB	HPB
GLN	CA	-	MET	CB	HPB	TYR	CG	HPB, ARM
GLN	C	-	MET	CG	HPB	TYR	CD1	HPB, ARM
GLN	O	ACP	MET	SD	ACP	TYR	CD2	HPB, ARM
GLN	CB	HPB	MET	CE	HPB	TYR	CE1	HPB, ARM
GLN	CG	HPB	PHE	N	DON	TYR	CE2	HPB, ARM
GLN	CD	-	PHE	CA	-	TYR	CZ	ARM
GLN	OE1	ACP	PHE	C	-	TYR	OH	DON, ACP
GLN	NE2	DON	PHE	O	ACP	VAL	N	DON
GLU	N	DON	PHE	CB	HPB	VAL	CA	-
GLU	CA	-	PHE	CG	HPB, ARM	VAL	C	-
GLU	C	-	PHE	CD1	HPB, ARM	VAL	O	ACP
GLU	O	ACP	PHE	CD2	HPB, ARM	VAL	CB	HPB
GLU	CB	HPB	PHE	CE1	HPB, ARM	VAL	CG1	HPB
GLU	CG	HPB	PHE	CE2	HPB, ARM	VAL	CG2	HPB
GLU	CD	-	PHE	CZ	HPB, ARM			
GLU	OE1	NEG, ACP						
GLU	OE2	NEG, ACP						
GLY	N	DON						
GLY	CA	-						
GLY	C	-						
GLY	O	ACP						

HPB: Hydrophobic, **DON:** Donor, **ACP:** Acceptor, **ARM:** Aromatic
POS: Positive, **NEG:** Negative, **SSB:** Sulfide Bridge

B.2 Resultados: base de dados COACH

Resultados do GRASP para cada proteína do conjunto de dados COACH. (**TN** = verdadeiro negativo; **FP** = falso positivo; **FN** = falso negativo; **TP** = verdadeiro positivo; **prec.** = precisão; **rec.** = revocação; **len** = tamanho da cadeia)

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1e4gT	362	4	5	9	0,65	0,69	0,64	0,98	0,65	380	11,58
2fsgA	664	5	2	8	0,7	0,62	0,8	0,99	0,62	679	24,28
118gA	281	1	1	14	0,93	0,93	0,93	0,99	0,94	297	12,44
2j72B	92	1	8	2	0,33	0,67	0,2	0,91	0,2	103	8,93
2dveA	211	2	2	20	0,9	0,91	0,91	0,98	0,91	235	8,33
1pmnA	322	6	3	13	0,73	0,68	0,81	0,97	0,7	344	36,7
3f1kA	228	0	8	16	0,8	1	0,67	0,97	0,67	252	22,07
3kjgA	238	2	11	3	0,34	0,6	0,21	0,95	0,22	254	7,56
1ogkB	218	4	3	14	0,78	0,78	0,82	0,97	0,78	239	8,38
1y30A	132	2	6	3	0,42	0,6	0,33	0,94	0,34	143	5,68
2e0nB	218	2	12	6	0,48	0,75	0,33	0,94	0,34	238	8,56
1ogoX	561	2	9	0	-0,01	0	0	0,98	0	572	17,98
1gsaA	275	3	30	6	0,3	0,67	0,17	0,89	0,17	314	8,65
1kv1A	311	10	1	9	0,64	0,47	0,9	0,97	0,49	331	28,81
1hwwA	988	14	0	12	0,67	0,46	1	0,99	0,47	1014	34,46
1eveA	516	8	1	9	0,68	0,53	0,9	0,98	0,55	534	36,59
3kakA	430	3	6	4	0,47	0,57	0,4	0,98	0,4	443	12,08
1gpkA	506	9	3	11	0,65	0,55	0,79	0,98	0,56	529	41
2g97A	441	9	5	8	0,52	0,47	0,62	0,97	0,48	463	23,22
1br6A	254	2	2	10	0,83	0,83	0,83	0,99	0,84	268	11,41
1um0A	348	2	10	5	0,47	0,71	0,33	0,97	0,34	365	11,49
1tkyA	203	7	8	6	0,41	0,46	0,43	0,93	0,45	224	7,65
3ct5A	148	0	11	0	0	0	0	0,93	0	159	10,98
1if7A	242	6	1	9	0,72	0,6	0,9	0,97	0,61	258	19,31
1nhvA	532	17	3	6	0,4	0,26	0,67	0,96	0,27	558	36,23
1ydrE	304	20	0	12	0,59	0,38	1	0,94	0,39	336	37,03
1xwqA	379	15	3	7	0,45	0,32	0,7	0,96	0,33	404	18,65
1f8gA	361	4	3	14	0,79	0,78	0,82	0,98	0,79	382	11,41
2z09A	106	0	7	11	0,76	1	0,61	0,94	0,61	124	3,72
2e6uX	120	0	7	15	0,8	1	0,68	0,95	0,68	142	4,56
1jylA	208	2	16	2	0,21	0,5	0,11	0,92	0,12	228	10,89
3egvA	235	1	0	18	0,97	0,95	1	1	0,95	254	12
2yw9D	221	3	12	11	0,58	0,79	0,48	0,94	0,48	247	28,8
2chtF	107	1	1	9	0,89	0,9	0,9	0,98	0,9	118	6,34
1oq5A	237	8	1	10	0,7	0,56	0,91	0,96	0,56	256	13,16
2r7aA	236	1	18	0	-0,02	0	0	0,93	0	255	10,9
1k22H	220	5	10	16	0,65	0,76	0,62	0,94	0,62	251	19,98
3ts1A	296	2	7	12	0,72	0,86	0,63	0,97	0,64	317	10,78
1p5rA	405	4	6	12	0,7	0,75	0,67	0,98	0,67	427	14,19
1a42A	239	4	3	10	0,73	0,71	0,77	0,97	0,72	256	19,5
1c1hA	287	1	8	10	0,7	0,91	0,56	0,97	0,56	306	13,87
2bsmA	193	1	2	12	0,88	0,92	0,86	0,99	0,86	208	10,52
1zajA	346	3	1	13	0,86	0,81	0,93	0,99	0,82	363	16,91
2i4nB	421	1	5	15	0,83	0,94	0,75	0,99	0,75	442	18,92
1hp0A	299	5	3	12	0,74	0,71	0,8	0,97	0,72	319	15,41
1q41A	322	5	1	11	0,79	0,69	0,92	0,98	0,7	339	36,38

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1e5qA	415	3	26	5	0,3	0,62	0,16	0,94	0,16	449	16,14
2ahwA	491	4	0	17	0,9	0,81	1	0,99	0,81	512	24,63
3nk7A	248	1	14	2	0,27	0,67	0,12	0,94	0,13	265	9,82
1sqfA	405	5	6	9	0,61	0,64	0,6	0,97	0,6	425	17,92
1goqA	279	1	8	14	0,76	0,93	0,64	0,97	0,64	302	14,79
1oweA	232	2	0	11	0,92	0,85	1	0,99	0,86	245	17,75
1lpzB	211	0	6	17	0,85	1	0,74	0,97	0,74	234	28,67
1jxmA	256	2	4	2	0,4	0,5	0,33	0,98	0,34	264	9,49
1xm6A	320	4	2	12	0,79	0,75	0,86	0,98	0,76	338	15
2qwiA	358	3	9	18	0,74	0,86	0,67	0,97	0,67	388	22
3h2kA	365	2	20	1	0,11	0,33	0,05	0,94	0,05	388	17,16
1mq6A	206	3	8	16	0,72	0,84	0,67	0,95	0,67	233	21,68
1a4kH	204	4	7	2	0,25	0,33	0,22	0,95	0,23	217	26,1
2vflA	197	2	1	5	0,76	0,71	0,83	0,99	0,73	205	9,77
1dcpA	90	2	2	5	0,69	0,71	0,71	0,96	0,72	99	4,06
2hblA	381	0	3	6	0,81	1	0,67	0,99	0,67	390	16,99
1svwA	169	2	15	4	0,34	0,67	0,21	0,91	0,22	190	10,51
3gh6A	197	1	3	8	0,79	0,89	0,73	0,98	0,73	209	12,34
1umlA	326	7	2	14	0,75	0,67	0,88	0,97	0,68	349	15,87
1gwmA	138	2	4	9	0,73	0,82	0,69	0,96	0,7	153	7,93
2e5mA	389	0	9	5	0,59	1	0,36	0,98	0,36	403	16,76
1ygcH	227	4	12	11	0,56	0,73	0,48	0,94	0,49	254	28,73
1f0rA	209	10	5	10	0,54	0,5	0,67	0,94	0,52	234	21,02
1l7fA	361	5	9	13	0,63	0,72	0,59	0,96	0,6	388	20,47
1einA	252	11	4	2	0,2	0,15	0,33	0,94	0,18	269	10,03
1ppcE	200	3	6	14	0,74	0,82	0,7	0,96	0,71	223	17,31
13gsA	193	0	4	13	0,87	1	0,76	0,98	0,76	210	13,6
1htfA	84	2	1	12	0,87	0,86	0,92	0,97	0,87	99	16,6
1m48A	108	3	2	10	0,78	0,77	0,83	0,96	0,78	123	3,97
5tlnA	285	3	17	11	0,53	0,79	0,39	0,94	0,4	316	12,32
1kaqA	162	4	11	9	0,52	0,69	0,45	0,92	0,46	186	8,2
2fv0A	353	9	4	11	0,62	0,55	0,73	0,97	0,56	377	14,09
1sqnA	230	5	2	11	0,75	0,69	0,85	0,97	0,7	248	19,68
2c6zA	255	5	1	12	0,8	0,71	0,92	0,98	0,72	273	9,38
1lqdB	209	3	7	15	0,73	0,83	0,68	0,96	0,69	234	26,15
1s7lA	157	2	10	8	0,57	0,8	0,44	0,93	0,45	177	6,49
1mehA	323	2	10	19	0,75	0,9	0,66	0,97	0,66	354	11,45
1jclB	218	22	2	9	0,45	0,29	0,82	0,9	0,32	251	9,61
2za1A	301	3	7	7	0,58	0,7	0,5	0,97	0,51	318	10,92
1b8uA	300	1	14	12	0,63	0,92	0,46	0,95	0,46	327	10,21
1fqoA	245	4	5	12	0,71	0,75	0,71	0,97	0,71	266	9,72
1ql9A	198	2	12	11	0,61	0,85	0,48	0,94	0,48	223	25,12
1u7zA	192	0	18	6	0,48	1	0,25	0,92	0,25	216	7,99
4dfrA	145	0	2	12	0,92	1	0,86	0,99	0,86	159	13,06
1k3yA	197	2	11	11	0,62	0,85	0,5	0,94	0,5	221	21,46
1nhuA	530	18	3	7	0,43	0,28	0,7	0,96	0,28	558	35,23
1x8xA	305	7	1	9	0,7	0,56	0,9	0,98	0,57	322	15,67
1sgcA	162	2	9	8	0,59	0,8	0,47	0,94	0,47	181	6,69
1lzsA	110	1	13	6	0,48	0,86	0,32	0,89	0,32	130	7,56
1sthA	124	1	2	9	0,85	0,9	0,82	0,98	0,82	136	4,75
1r15A	233	1	4	13	0,83	0,93	0,76	0,98	0,77	251	8,68
3ldkA	617	7	0	10	0,76	0,59	1	0,99	0,6	634	20,6
2zu3A	161	4	2	13	0,8	0,76	0,87	0,97	0,78	180	8,29

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
3gplA	500	2	11	3	0,35	0,6	0,21	0,97	0,22	516	14,99
1znzA	173	1	1	8	0,88	0,89	0,89	0,99	0,9	183	6,36
1swkA	100	2	2	13	0,85	0,87	0,87	0,97	0,87	117	7,7
1yvfA	539	11	1	13	0,7	0,54	0,93	0,98	0,56	564	24,32
3iiqA	196	3	13	5	0,38	0,62	0,28	0,93	0,29	217	13,66
2qo9A	258	9	2	8	0,6	0,47	0,8	0,96	0,49	277	26,25
2r68A	446	0	8	2	0,44	1	0,2	0,98	0,2	456	18,38
1jcgA	313	3	10	9	0,58	0,75	0,47	0,96	0,48	335	11,7
1jd0A	251	0	0	9	1	1	1	1	1	260	15,71
1ydtE	307	11	2	16	0,71	0,59	0,89	0,96	0,6	336	28,72
2v8lA	97	0	3	6	0,8	1	0,67	0,97	0,67	106	7,26
1llrD	91	0	3	9	0,85	1	0,75	0,97	0,75	103	4,92
1kv2A	304	4	4	13	0,75	0,76	0,76	0,98	0,77	325	29,47
1afkA	106	4	6	8	0,57	0,67	0,57	0,92	0,58	124	6,43
2j4eB	172	5	9	8	0,5	0,62	0,47	0,93	0,48	194	8,42
2csnA	276	5	3	9	0,68	0,64	0,75	0,97	0,66	293	26,17
1ytmA	494	4	3	16	0,81	0,8	0,84	0,99	0,81	517	20,55
1h9uA	184	3	7	7	0,57	0,7	0,5	0,95	0,51	201	28,66
3a0tA	137	0	14	1	0,25	1	0,07	0,91	0,07	152	7,19
1x2bA	295	9	0	9	0,7	0,5	1	0,97	0,51	313	16,09
2zgzA	300	2	6	12	0,74	0,86	0,67	0,98	0,67	320	9,84
1sz3A	139	2	9	5	0,47	0,71	0,36	0,93	0,36	155	5,47
1zu0A	503	3	19	4	0,3	0,57	0,17	0,96	0,18	529	18,05
3ku1A	200	1	13	0	-0,02	0	0	0,93	0	214	13,04
3iesA	406	6	6	13	0,67	0,68	0,68	0,97	0,69	431	22,53
2e1tA	421	0	19	0	0	0	0	0,96	0	440	16,58
3hitA	237	3	5	12	0,74	0,8	0,71	0,97	0,72	257	13,8
2j8yA	253	0	4	8	0,81	1	0,67	0,98	0,67	265	12,12
3cgyA	122	1	9	1	0,2	0,5	0,1	0,92	0,1	133	6,97
1khrD	186	1	4	15	0,85	0,94	0,79	0,98	0,79	206	9,26
2b3dA	172	0	16	5	0,47	1	0,24	0,92	0,24	193	10,84
1szjG	309	4	0	20	0,91	0,83	1	0,99	0,84	333	20,14
3b4yA	299	4	25	4	0,23	0,5	0,14	0,91	0,14	332	9,21
3hp8A	86	0	21	0	0	0	0	0,8	0	107	3,83
1mu0A	278	4	5	6	0,56	0,6	0,55	0,97	0,55	293	11,2
1ke5A	260	8	0	13	0,77	0,62	1	0,97	0,63	281	19,61
1kpeB	99	0	1	13	0,96	1	0,93	0,99	0,93	113	5,82
2f9wA	238	3	8	0	-0,02	0	0	0,96	0	249	9,01
1ri1A	233	1	11	7	0,56	0,88	0,39	0,95	0,39	252	12,74
2dttB	103	2	8	2	0,28	0,5	0,2	0,91	0,21	115	5,06
4stdA	144	4	4	12	0,72	0,75	0,75	0,95	0,76	164	8,51
2vaqA	287	4	5	9	0,65	0,69	0,64	0,97	0,65	305	17,05
1a7xA	93	9	2	3	0,34	0,25	0,6	0,9	0,26	107	4,5
1bxqA	293	8	6	16	0,67	0,67	0,73	0,96	0,68	323	24,12
1gkcA	121	1	28	9	0,41	0,9	0,24	0,82	0,24	159	6,59
1oytH	221	11	3	15	0,67	0,58	0,83	0,94	0,58	250	25,42
1j9kA	232	6	5	4	0,4	0,4	0,44	0,96	0,41	247	8,37
1jieA	102	1	11	8	0,57	0,89	0,42	0,9	0,42	122	5,31
3hpiA	353	5	2	9	0,72	0,64	0,82	0,98	0,65	369	19,16
1k7fA	239	4	5	5	0,51	0,56	0,5	0,96	0,52	253	15,18
1tywA	522	3	4	14	0,79	0,82	0,78	0,99	0,78	543	16,26
3exsA	195	5	6	2	0,24	0,29	0,25	0,95	0,27	208	6,76
3efvA	433	2	14	10	0,57	0,83	0,42	0,97	0,42	459	27,01

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
2x60A	311	4	14	4	0,31	0,5	0,22	0,95	0,23	333	11,97
1u72A	150	0	14	22	0,75	1	0,61	0,92	0,61	186	13,78
2fk8A	263	1	7	10	0,72	0,91	0,59	0,97	0,59	281	10,23
1i7zA	206	4	6	3	0,35	0,43	0,33	0,95	0,35	219	27,33
1l2sA	341	6	0	8	0,75	0,57	1	0,98	0,59	355	22,95
1pphE	201	3	5	14	0,76	0,82	0,74	0,96	0,74	223	21,12
1cetA	278	18	2	7	0,44	0,28	0,78	0,93	0,29	305	20,7
1p77A	247	6	8	4	0,34	0,4	0,33	0,95	0,34	265	10,35
1bq3A	214	6	8	6	0,43	0,5	0,43	0,94	0,44	234	9,54
3h72A	457	2	1	14	0,9	0,88	0,93	0,99	0,88	474	27,79
3a2sX	328	6	3	4	0,47	0,4	0,57	0,97	0,4	341	15,75
3aaqA	333	1	12	3	0,38	0,75	0,2	0,96	0,2	349	15,4
1o26B	198	0	6	17	0,85	1	0,74	0,97	0,74	221	11,76
1ucdA	174	0	6	10	0,78	1	0,62	0,97	0,62	190	9,04
1y6bA	251	0	2	13	0,93	1	0,87	0,99	0,87	266	30,51
3gqkA	147	6	3	7	0,59	0,54	0,7	0,94	0,56	163	9,41
2dyaB	143	1	1	10	0,9	0,91	0,91	0,99	0,91	155	9,98
1fcvA	304	8	11	1	0,07	0,11	0,08	0,94	0,09	324	17,04
1lbyA	231	2	10	9	0,6	0,82	0,47	0,95	0,48	252	10,11
3du4A	428	1	7	12	0,76	0,92	0,63	0,98	0,64	448	22,13
1zt9A	89	2	5	5	0,56	0,71	0,5	0,93	0,51	101	4,31
148lE	147	2	13	0	-0,03	0	0	0,91	0	162	9,12
3f47A	316	4	4	20	0,82	0,83	0,83	0,98	0,84	344	15,85
1hnjA	295	4	5	13	0,73	0,76	0,72	0,97	0,73	317	14,76
2hzqA	155	3	7	1	0,15	0,25	0,12	0,94	0,14	166	8,06
1v0pA	258	2	6	11	0,73	0,85	0,65	0,97	0,65	277	32,77
2qv7A	286	0	16	1	0,24	1	0,06	0,95	0,06	303	10,73
2c96A	224	0	5	8	0,78	1	0,62	0,98	0,62	237	8,39
1f4fA	242	11	0	9	0,66	0,45	1	0,96	0,47	262	17,75
3hvlA	268	1	4	14	0,84	0,93	0,78	0,98	0,78	287	28,58
3cxiA	104	2	9	6	0,51	0,75	0,4	0,91	0,41	121	6,88
3dsrA	406	3	9	1	0,15	0,25	0,1	0,97	0,1	419	23,95
1ddtA	499	8	4	12	0,66	0,6	0,75	0,98	0,61	523	18,5
3i6cA	97	4	7	5	0,43	0,56	0,42	0,9	0,42	113	5,15
1z95A	215	6	3	14	0,74	0,7	0,82	0,96	0,71	238	16,17
3cpaA	289	6	3	9	0,66	0,6	0,75	0,97	0,62	307	12,23
1bqoB	142	4	19	8	0,38	0,67	0,3	0,87	0,3	173	9,12
1f4eA	242	10	5	7	0,46	0,41	0,58	0,94	0,42	264	21,36
3crrA	227	2	0	7	0,88	0,78	1	0,99	0,78	236	10,07
2ateA	149	0	4	9	0,82	1	0,69	0,98	0,69	162	10,03
2qzzA	279	1	10	20	0,78	0,95	0,67	0,96	0,67	310	9,26
3kdiA	168	4	1	8	0,76	0,67	0,89	0,97	0,68	181	6,83
1qhiA	284	3	1	12	0,85	0,8	0,92	0,99	0,81	300	9,65
3kp5A	125	2	2	7	0,76	0,78	0,78	0,97	0,78	136	6,76
1k7eA	245	4	2	10	0,76	0,71	0,83	0,98	0,73	261	9,19
1e3vA	115	2	5	6	0,61	0,75	0,55	0,95	0,55	128	5,52
1arcA	252	0	10	1	0,3	1	0,09	0,96	0,09	263	10,59
2varA	280	3	21	7	0,39	0,7	0,25	0,92	0,25	311	10,05
2q6vA	350	5	14	1	0,08	0,17	0,07	0,95	0,07	370	16,36
966cA	125	0	23	9	0,49	1	0,28	0,85	0,28	157	6,54
3gdlA	241	0	5	11	0,82	1	0,69	0,98	0,69	257	9,92
1jjeA	204	2	2	12	0,85	0,86	0,86	0,98	0,86	220	8,43
2z1sA	425	7	6	7	0,5	0,5	0,54	0,97	0,52	445	29,29

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
3c2fA	255	6	4	2	0,27	0,25	0,33	0,96	0,27	267	9,4
1h1pA	279	8	0	10	0,73	0,56	1	0,97	0,57	297	34,26
1qpeA	257	3	0	11	0,88	0,79	1	0,99	0,8	271	25,59
1wopA	331	3	13	15	0,65	0,83	0,54	0,96	0,54	362	12,6
1ryaA	142	6	2	10	0,7	0,62	0,83	0,95	0,64	160	5,38
1vcuA	354	3	1	12	0,85	0,8	0,92	0,99	0,81	370	12,91
1n46A	230	1	5	12	0,8	0,92	0,71	0,98	0,71	248	25,25
7estE	213	1	17	9	0,53	0,9	0,35	0,92	0,35	240	22,34
3bynA	418	5	1	16	0,84	0,76	0,94	0,99	0,77	440	19,4
1sj0A	224	3	6	10	0,67	0,77	0,62	0,96	0,63	243	25,97
1exaA	215	5	3	13	0,75	0,72	0,81	0,97	0,73	236	24,43
2rfhA	289	7	0	11	0,77	0,61	1	0,98	0,63	307	11,05
3dwrA	264	4	16	2	0,16	0,33	0,11	0,93	0,12	286	18,66
1elaA	219	4	9	8	0,53	0,67	0,47	0,95	0,48	240	26,79
1n2vA	350	9	4	9	0,57	0,5	0,69	0,97	0,51	372	18,38
1uvsH	220	7	2	12	0,72	0,63	0,86	0,96	0,65	241	18,11
1f4gA	240	8	0	14	0,78	0,64	1	0,97	0,65	262	16,89
2dkcA	511	7	4	14	0,71	0,67	0,78	0,98	0,68	536	19,03
1jlaA	530	4	2	11	0,78	0,73	0,85	0,99	0,75	547	22,61
1fthA	108	6	3	0	-0,04	0	0	0,92	0	117	4,6
2cx8A	204	4	15	2	0,16	0,33	0,12	0,92	0,12	225	6,56
2uyqA	265	2	7	0	-0,01	0	0	0,97	0	274	16,36
1s19A	230	6	0	17	0,85	0,74	1	0,98	0,74	253	22,48
1d3pB	229	4	0	17	0,89	0,81	1	0,98	0,81	250	22,13
1bsvA	289	7	8	13	0,61	0,65	0,62	0,95	0,63	317	10,74
3i0dA	244	13	2	5	0,42	0,28	0,71	0,94	0,28	264	15,27
1cenA	317	6	6	5	0,44	0,45	0,45	0,96	0,47	334	12,97
2dptA	209	0	15	8	0,57	1	0,35	0,94	0,35	232	10,47
1xqpA	233	4	13	2	0,18	0,33	0,13	0,93	0,14	252	14,18
2ywcA	457	9	4	5	0,43	0,36	0,56	0,97	0,37	475	15,27
1foaA	308	14	8	12	0,49	0,46	0,6	0,94	0,47	342	11,63
1h72C	268	2	4	22	0,87	0,92	0,85	0,98	0,85	296	10,9
2x4oA	252	1	8	14	0,76	0,93	0,64	0,97	0,64	275	19,21
1y2vA	126	7	1	8	0,66	0,53	0,89	0,94	0,54	142	5,19
1teiA	216	0	12	9	0,64	1	0,43	0,95	0,43	237	10,6
1bnwA	241	5	0	10	0,81	0,67	1	0,98	0,67	256	12,03
1hfcA	134	0	12	11	0,66	1	0,48	0,92	0,48	157	9,77
1ltzA	258	6	6	4	0,38	0,4	0,4	0,96	0,43	274	9,93
2yw2A	401	1	15	6	0,48	0,86	0,29	0,96	0,29	423	15,76
1xpyC	354	3	1	12	0,85	0,8	0,92	0,99	0,81	370	23,44
2yyuB	214	7	4	4	0,4	0,36	0,5	0,95	0,38	229	12,25
1i8zA	242	2	2	12	0,85	0,86	0,86	0,98	0,86	258	14,94
1v48A	244	2	2	13	0,86	0,87	0,87	0,98	0,87	261	15,98
1r1hA	675	6	1	14	0,8	0,7	0,93	0,99	0,71	696	21,13
2bbwA	198	0	17	5	0,46	1	0,23	0,92	0,23	220	9,75
2rk2A	54	0	1	3	0,86	1	0,75	0,98	0,75	58	3,89
1rsdA	109	4	3	5	0,56	0,56	0,62	0,94	0,57	121	4,75
1eyrA	209	5	6	5	0,45	0,5	0,45	0,95	0,47	225	7,35
1vh3A	192	0	9	10	0,71	1	0,53	0,96	0,53	211	11,29
1mncA	134	1	13	10	0,59	0,91	0,43	0,91	0,44	158	8,09
1xoqA	305	3	3	15	0,82	0,83	0,83	0,98	0,84	326	16,89
1xvtA	380	3	5	14	0,77	0,82	0,74	0,98	0,74	402	14,55
1ixnA	222	4	3	13	0,77	0,76	0,81	0,97	0,78	242	8,83

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1nz7A	260	4	6	12	0,69	0,75	0,67	0,96	0,68	282	11,16
1wnzA	167	1	9	3	0,41	0,75	0,25	0,94	0,25	180	9,79
3h18A	277	0	12	4	0,49	1	0,25	0,96	0,25	293	20,38
2aibA	85	0	8	5	0,59	1	0,38	0,92	0,38	98	2,9
1c3jA	315	5	0	13	0,84	0,72	1	0,98	0,73	333	11,68
1fk6A	80	2	7	4	0,45	0,67	0,36	0,9	0,37	93	4,35
1ezqA	211	1	6	16	0,81	0,94	0,73	0,97	0,73	234	21,91
3lbzA	109	0	13	0	0	0	0	0,89	0	122	7,91
1ch8A	396	4	2	29	0,9	0,88	0,94	0,99	0,89	431	24,43
1nmkA	151	1	5	8	0,72	0,89	0,62	0,96	0,62	165	9,73
7dfrA	125	1	19	14	0,58	0,93	0,42	0,87	0,43	159	11,41
3ergA	196	1	6	5	0,6	0,83	0,45	0,97	0,46	208	11,63
3fwrA	131	2	11	1	0,13	0,33	0,08	0,91	0,09	145	6,34
2ioaA	549	6	20	14	0,52	0,7	0,41	0,96	0,41	589	17,46
3bazA	289	1	9	12	0,71	0,92	0,57	0,97	0,57	311	11,69
2oecB	224	4	4	7	0,62	0,64	0,64	0,97	0,65	239	12,61
1f0tA	199	5	6	13	0,68	0,72	0,68	0,95	0,69	223	27,25
1jkkA	257	2	4	14	0,81	0,88	0,78	0,98	0,79	277	29,08
1jsvA	268	9	0	10	0,71	0,53	1	0,97	0,54	287	24,44
1g97A	417	7	10	12	0,57	0,63	0,55	0,96	0,55	446	19,83
1f17A	273	6	1	13	0,79	0,68	0,93	0,98	0,7	293	15,34
1r9oA	414	8	10	23	0,7	0,74	0,7	0,96	0,7	455	30,85
2gz3A	323	3	13	18	0,68	0,86	0,58	0,96	0,58	357	11,82
1h1sA	277	6	0	14	0,83	0,7	1	0,98	0,71	297	24,98
1ohrA	85	4	1	9	0,76	0,69	0,9	0,95	0,7	99	7,94
2zcpA	257	1	4	22	0,89	0,96	0,85	0,98	0,85	284	10,07
1v3sA	78	1	11	7	0,53	0,88	0,39	0,88	0,39	97	3,92
1blcA	242	2	5	8	0,69	0,8	0,62	0,97	0,62	257	11,57
1ncoB	94	1	13	5	0,44	0,83	0,28	0,88	0,28	113	3,46
2hxmA	208	2	3	10	0,79	0,83	0,77	0,98	0,78	223	9,5
1g6cA	202	0	1	23	0,98	1	0,96	1	0,96	226	8,44
1xozA	308	2	4	12	0,79	0,86	0,75	0,98	0,76	326	23,49
1rn8A	132	0	0	8	1	1	1	1	1	140	6,17
1bs1A	197	4	4	19	0,81	0,83	0,83	0,96	0,83	224	7,71
1onhA	350	5	1	7	0,71	0,58	0,88	0,98	0,6	363	20,81
1mmbA	129	1	17	11	0,56	0,92	0,39	0,89	0,4	158	6,6
2ovdA	149	6	7	0	-0,04	0	0	0,92	0,02	162	5,43
3g1xA	200	1	3	9	0,81	0,9	0,75	0,98	0,75	213	7,92
1thlA	285	4	15	12	0,55	0,75	0,44	0,94	0,45	316	14,01
1h0sA	124	1	2	10	0,86	0,91	0,83	0,98	0,84	137	7,37
3h39A	394	6	3	12	0,72	0,67	0,8	0,98	0,67	415	15,1
1tjwA	433	8	2	6	0,56	0,43	0,75	0,98	0,44	449	31,16
1ettH	235	11	1	12	0,67	0,52	0,92	0,95	0,54	259	34,49
3kv8A	116	4	2	11	0,76	0,73	0,85	0,95	0,75	133	7,14
1x7pA	248	1	14	2	0,27	0,67	0,12	0,94	0,13	265	7,51
3ertA	229	4	4	9	0,68	0,69	0,69	0,97	0,71	246	17,43
2fa0A	426	2	13	9	0,56	0,82	0,41	0,97	0,41	450	18,15
1ex8A	136	2	6	14	0,76	0,88	0,7	0,95	0,7	158	5,01
2gfxA	388	7	4	12	0,67	0,63	0,75	0,97	0,64	411	15,35
2zgmB	151	0	2	6	0,86	1	0,75	0,99	0,75	159	7,13
1oxvA	340	2	2	9	0,81	0,82	0,82	0,99	0,82	353	25,05
1ywrA	317	8	2	11	0,69	0,58	0,85	0,97	0,6	338	23,7
1a26A	336	7	5	3	0,32	0,3	0,38	0,97	0,32	351	14,05

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1a2kC	180	2	4	10	0,76	0,83	0,71	0,97	0,72	196	14,99
2jbtA	377	0	15	8	0,58	1	0,35	0,96	0,35	400	22,36
1u4dA	243	4	2	9	0,74	0,69	0,82	0,98	0,71	258	26,88
3lzzA	153	1	6	1	0,25	0,5	0,14	0,96	0,14	161	6,52
1kgzB	315	2	9	4	0,44	0,67	0,31	0,97	0,31	330	11,78
1vpsB	274	4	5	6	0,56	0,6	0,55	0,97	0,56	289	17,44
1goyA	93	3	8	4	0,39	0,57	0,33	0,9	0,34	108	6,04
830cA	138	0	13	13	0,68	1	0,5	0,92	0,5	164	10,28
1k54A	230	5	1	8	0,73	0,62	0,89	0,98	0,63	244	15,37
1r58A	351	5	2	11	0,75	0,69	0,85	0,98	0,7	369	13,83
1lloA	254	3	4	12	0,76	0,8	0,75	0,97	0,76	273	13,62
2ixlA	182	3	0	11	0,88	0,79	1	0,98	0,79	196	7,57
1bzyA	194	1	5	14	0,82	0,93	0,74	0,97	0,74	214	7,46
3in1A	296	7	6	3	0,29	0,3	0,33	0,96	0,31	312	14,53
3kbnA	372	4	0	12	0,86	0,75	1	0,99	0,75	388	14,91
3cagA	63	0	3	11	0,87	1	0,79	0,96	0,79	77	3,85
1ow3B	160	0	3	16	0,91	1	0,84	0,98	0,84	179	13,59
1cimA	239	4	4	9	0,68	0,69	0,69	0,97	0,7	256	14,38
1towA	114	7	0	10	0,74	0,59	1	0,95	0,6	131	6,98
1x55A	410	4	1	19	0,88	0,83	0,95	0,99	0,83	434	16,22
2z0xA	143	0	12	2	0,36	1	0,14	0,92	0,14	157	8,39
1q8jA	518	28	7	6	0,26	0,18	0,46	0,94	0,18	559	17,79
2wvaA	541	3	5	16	0,79	0,84	0,76	0,99	0,77	565	20,66
1pw1A	328	2	1	14	0,9	0,88	0,93	0,99	0,88	345	19,62
2chzA	817	8	3	11	0,67	0,58	0,79	0,99	0,59	839	31,96
3ll3A	467	3	20	2	0,17	0,4	0,09	0,95	0,09	492	17,78
1k3uA	251	1	3	13	0,86	0,93	0,81	0,99	0,81	268	12,89
3maqA	735	29	2	10	0,45	0,26	0,83	0,96	0,26	776	22,15
3erkA	333	6	3	8	0,63	0,57	0,73	0,97	0,59	350	29,21
1mkaA	153	4	10	4	0,34	0,5	0,29	0,92	0,3	171	8,29
1cqfB	54	2	5	8	0,64	0,8	0,62	0,9	0,62	69	3,49
3adpA	292	1	13	4	0,42	0,8	0,24	0,95	0,24	310	17,65
3ftfA	227	1	14	4	0,4	0,8	0,22	0,94	0,23	246	11,13
1p2yA	378	7	3	19	0,78	0,73	0,86	0,98	0,74	407	23,89
2d29A	367	7	4	8	0,58	0,53	0,67	0,97	0,54	386	17,9
3i8xA	241	4	5	8	0,62	0,67	0,62	0,97	0,62	258	12,1
2ggaA	425	2	2	16	0,88	0,89	0,89	0,99	0,89	445	26,05
2qttA	239	2	1	6	0,8	0,75	0,86	0,99	0,76	248	11,23
2zasA	216	0	1	10	0,95	1	0,91	1	0,91	227	24,59
1mq5A	205	3	11	14	0,65	0,82	0,56	0,94	0,57	233	22,4
2rkmA	492	13	5	7	0,44	0,35	0,58	0,97	0,36	517	26,55
1atlA	181	2	7	10	0,68	0,83	0,59	0,96	0,59	200	7,37
1ybuA	159	0	6	1	0,37	1	0,14	0,96	0,14	166	8,07
3gd9A	347	2	13	0	-0,01	0	0	0,96	0	362	19,88
3idoA	145	2	4	7	0,68	0,78	0,64	0,96	0,64	158	8,57
1etrH	238	6	2	13	0,75	0,68	0,87	0,97	0,7	259	32,78
1tvpB	277	8	4	2	0,24	0,2	0,33	0,96	0,21	291	12,64
1mxiA	140	0	13	3	0,41	1	0,19	0,92	0,19	156	5,12
3jynA	295	3	19	8	0,44	0,73	0,3	0,93	0,3	325	16,3
1w1pA	477	11	0	9	0,66	0,45	1	0,98	0,46	497	23,44
1i1hA	195	0	9	5	0,58	1	0,36	0,96	0,36	209	7,65
1qbuA	84	3	1	11	0,83	0,79	0,92	0,96	0,79	99	9,57
2hl0A	129	1	2	11	0,87	0,92	0,85	0,98	0,85	143	4,79

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1ffqA	516	6	2	16	0,8	0,73	0,89	0,99	0,74	540	28,39
1q51A	254	1	11	5	0,49	0,83	0,31	0,96	0,31	271	11,7
2bawA	219	5	17	9	0,43	0,64	0,35	0,91	0,35	250	29,05
2vfcA	247	4	15	6	0,38	0,6	0,29	0,93	0,3	272	10,86
1qjiA	185	1	11	3	0,38	0,75	0,21	0,94	0,22	200	13,34
1s3vA	168	3	3	12	0,78	0,8	0,8	0,97	0,81	186	16,73
3b6aA	199	1	13	0	-0,02	0	0	0,93	0	213	6,89
4f10A	335	3	4	11	0,75	0,79	0,73	0,98	0,74	353	12,4
1g4oA	243	7	0	8	0,72	0,53	1	0,97	0,54	258	11,73
1og1A	207	2	5	9	0,71	0,82	0,64	0,97	0,65	223	8,47
2qyqA	174	2	1	9	0,85	0,82	0,9	0,98	0,82	186	7,21
2vkmA	360	4	1	24	0,9	0,86	0,96	0,99	0,86	389	33,97
3hvoA	546	6	3	4	0,47	0,4	0,57	0,98	0,4	559	19,91
1fjsA	211	3	6	14	0,74	0,82	0,7	0,96	0,71	234	22,79
2zhzB	162	0	8	4	0,56	1	0,33	0,95	0,33	174	7,6
2hobA	283	2	1	20	0,93	0,91	0,95	0,99	0,91	306	10,59
1s3fA	153	2	2	8	0,79	0,8	0,8	0,98	0,81	165	7,47
1jq3A	266	5	9	15	0,66	0,75	0,62	0,95	0,63	295	11,16
2br1A	254	6	0	12	0,81	0,67	1	0,98	0,68	272	28,72
2pknA	305	6	11	1	0,08	0,14	0,08	0,95	0,09	323	12,59
2oalB	494	10	8	16	0,62	0,62	0,67	0,97	0,62	528	18,15
2artA	222	2	9	14	0,71	0,88	0,61	0,96	0,61	247	10,41
1v0yA	480	4	2	7	0,7	0,64	0,78	0,99	0,66	493	19,79
1efyA	340	2	1	7	0,82	0,78	0,88	0,99	0,79	350	17,03
2zjaA	689	1	5	5	0,64	0,83	0,5	0,99	0,5	700	22,78
1k0nA	220	0	6	0	0	0	0	0,97	0	226	7,61
1azmA	245	5	0	8	0,78	0,62	1	0,98	0,63	258	11,84
2ed4A	124	0	21	4	0,37	1	0,16	0,86	0,16	149	4,81
1uvtH	224	9	1	13	0,72	0,59	0,93	0,96	0,61	247	17,75
2q71A	335	1	3	17	0,89	0,94	0,85	0,99	0,85	356	11,25
1uyyA	112	3	7	9	0,61	0,75	0,56	0,92	0,57	131	6,81
1hdqA	291	6	0	10	0,78	0,62	1	0,98	0,65	307	17,06
1uouA	410	15	5	8	0,44	0,35	0,62	0,95	0,37	438	13,87
3duwA	204	1	11	4	0,44	0,8	0,27	0,95	0,27	220	9,79
3stdA	141	7	1	13	0,75	0,65	0,93	0,95	0,66	162	4,89
1x6uA	257	2	8	5	0,51	0,71	0,38	0,96	0,39	272	14,1
2wzmA	247	9	14	4	0,22	0,31	0,22	0,92	0,23	274	15,97
1kwcB	276	2	1	9	0,85	0,82	0,9	0,99	0,83	288	12,78
1ig3A	243	2	2	7	0,77	0,78	0,78	0,98	0,79	254	9,24
1un1A	274	4	1	13	0,83	0,76	0,93	0,98	0,78	292	21,19
2zdqA	289	0	12	18	0,76	1	0,6	0,96	0,6	319	19,87
1hvpA	84	5	1	9	0,73	0,64	0,9	0,94	0,66	99	12,58
1wq1G	316	2	2	0	-0,01	0	0	0,99	0	320	10,23
1eswA	474	5	15	6	0,38	0,55	0,29	0,96	0,29	500	28,61
1qtiA	508	8	0	11	0,75	0,58	1	0,98	0,59	527	35,25
3btsB	371	3	16	0	-0,02	0	0	0,95	0	390	19,9
3d4pA	287	1	4	15	0,85	0,94	0,79	0,98	0,79	307	14,05
2gj5A	134	7	16	4	0,2	0,36	0,2	0,86	0,21	161	8,99
1uamA	229	0	6	15	0,83	1	0,71	0,98	0,71	250	11,88
2ihzA	353	1	8	17	0,79	0,94	0,68	0,98	0,68	379	19,45
3cwkA	122	3	5	7	0,61	0,7	0,58	0,94	0,59	137	6,16
3b3fA	312	5	5	15	0,73	0,75	0,75	0,97	0,76	337	10,37
3dzlB	147	0	0	11	1	1	1	1	1	158	7,7

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
2g25A	804	3	6	18	0,8	0,86	0,75	0,99	0,76	831	28,72
1pfkA	283	2	26	9	0,43	0,82	0,26	0,91	0,26	320	12,91
2w1aC	68	0	10	0	0	0	0	0,87	0	78	3,29
3gidA	437	8	12	3	0,21	0,27	0,2	0,96	0,2	460	20,02
1oxmA	184	4	1	7	0,73	0,64	0,88	0,97	0,65	196	8,43
1of8B	327	2	3	12	0,82	0,86	0,8	0,99	0,8	344	16,35
1of1A	291	4	0	12	0,86	0,75	1	0,99	0,76	307	11,06
1txcA	138	2	16	1	0,1	0,33	0,06	0,89	0,06	157	4,95
1yqyA	496	4	3	11	0,75	0,73	0,79	0,99	0,74	514	24,01
2gteA	102	1	12	9	0,58	0,9	0,43	0,9	0,43	124	5,6
3e3sA	191	2	12	1	0,14	0,33	0,08	0,93	0,08	206	12,94
1j8rA	188	0	8	0	0	0	0	0,96	0	196	17,95
3d1gA	350	5	8	3	0,3	0,38	0,27	0,96	0,28	366	13,8
2dzbA	236	2	4	9	0,74	0,82	0,69	0,98	0,7	251	9,01
1f7bA	274	4	1	14	0,84	0,78	0,93	0,98	0,79	293	11,62
1lxmA	778	4	4	8	0,66	0,67	0,67	0,99	0,67	794	27,93
1ndiA	573	5	10	8	0,51	0,62	0,44	0,97	0,45	596	24,41
2b99A	129	0	17	0	0	0	0	0,88	0	146	6,67
2vu9A	413	0	9	3	0,49	1	0,25	0,98	0,25	425	27,04
1uf8A	288	2	2	11	0,84	0,85	0,85	0,99	0,85	303	11,84
3ej0A	159	4	10	0	-0,04	0	0	0,92	0	173	9,55
3manA	286	3	0	8	0,85	0,73	1	0,99	0,73	297	14,24
2hixA	558	5	8	6	0,47	0,55	0,43	0,98	0,43	577	22,6
1xtbA	543	1	0	12	0,96	0,92	1	1	0,93	556	18,32
1r091	260	4	3	6	0,62	0,6	0,67	0,97	0,61	273	14,2
1theA	233	4	8	8	0,55	0,67	0,5	0,95	0,51	253	10,26
1r55A	184	0	12	7	0,59	1	0,37	0,94	0,37	203	7,39
2wn7A	375	7	1	9	0,7	0,56	0,9	0,98	0,57	392	12,36
1n07A	126	1	16	11	0,57	0,92	0,41	0,89	0,41	154	4,98
3b6rB	352	2	6	16	0,79	0,89	0,73	0,98	0,73	376	14,74
1i71A	289	3	12	5	0,41	0,62	0,29	0,95	0,3	309	16,7
1k1jA	200	1	9	13	0,72	0,93	0,59	0,96	0,59	223	26,37
1gwxA	247	3	9	11	0,64	0,79	0,55	0,96	0,56	270	22,61
1c5iA	167	2	3	13	0,82	0,87	0,81	0,97	0,81	185	9,07
2jgvD	288	10	4	8	0,52	0,44	0,67	0,95	0,46	310	10,06
2e9zA	457	8	2	9	0,65	0,53	0,82	0,98	0,54	476	18,3
1ytjA	85	5	2	7	0,64	0,58	0,78	0,93	0,59	99	11,85
1navA	234	1	6	12	0,77	0,92	0,67	0,97	0,67	253	26,53
1p4nA	299	12	23	1	0	0,08	0,04	0,9	0,06	335	11,61
2royA	111	4	0	6	0,76	0,6	1	0,97	0,61	121	6,36
1wxiA	235	8	5	13	0,64	0,62	0,72	0,95	0,63	261	8,63
2qehA	131	1	2	11	0,87	0,92	0,85	0,98	0,85	145	5,81
3m4eA	284	7	2	0	-0,01	0	0	0,97	0	293	11,71
3ek5A	223	6	7	5	0,41	0,45	0,42	0,95	0,42	241	9,13
1ydsE	302	20	0	14	0,62	0,41	1	0,94	0,42	336	36,66
5galB	124	2	0	6	0,86	0,75	1	0,98	0,76	132	8,35
2rjcA	299	0	7	7	0,7	1	0,5	0,98	0,5	313	17,52
2cwhB	311	0	7	19	0,85	1	0,73	0,98	0,73	337	13,38
2i56A	406	5	0	10	0,81	0,67	1	0,99	0,67	421	16,46
2vcjA	191	2	4	11	0,77	0,85	0,73	0,97	0,74	208	10,44
1zdfA	242	0	2	14	0,93	1	0,88	0,99	0,88	258	8,82
3gpoA	139	4	11	8	0,48	0,67	0,42	0,91	0,43	162	4,67
1pnfA	304	1	9	0	-0,01	0	0	0,97	0	314	14,37

Continua na próxima página

Tabela B.1 – *Continuação da página anterior*

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
2gwhA	257	0	8	20	0,83	1	0,71	0,97	0,71	285	11,99
2bogX	265	1	2	12	0,88	0,92	0,86	0,99	0,86	280	9,82
1xdhA	307	2	6	14	0,77	0,88	0,7	0,98	0,7	329	32,48
1ojzA	194	1	12	5	0,47	0,83	0,29	0,94	0,3	212	9,14
2fzsB	179	1	8	4	0,5	0,8	0,33	0,95	0,34	192	10,67
1rzuA	466	4	6	1	0,16	0,2	0,14	0,98	0,15	477	18,26
1y8eA	234	3	5	2	0,32	0,4	0,29	0,97	0,3	244	9,96
1pkkB	168	5	0	4	0,66	0,44	1	0,97	0,46	177	6,65
1mbzA	471	1	7	17	0,81	0,94	0,71	0,98	0,71	496	21,67
2fhjA	267	3	11	15	0,67	0,83	0,58	0,95	0,58	296	11,88
1o86A	541	12	5	16	0,65	0,57	0,76	0,97	0,59	574	20,4
2simA	368	2	0	11	0,92	0,85	1	0,99	0,85	381	15,6
2ntyC	159	3	2	9	0,77	0,75	0,82	0,97	0,76	173	9,83
3hiyA	359	6	3	14	0,75	0,7	0,82	0,98	0,71	382	13,94
1y3iA	218	0	8	5	0,61	1	0,38	0,97	0,38	231	8,38
1d4pB	229	8	0	13	0,77	0,62	1	0,97	0,63	250	20,68
5stdA	136	11	6	11	0,51	0,5	0,65	0,9	0,52	164	5,18
1fkgA	93	5	0	9	0,78	0,64	1	0,95	0,66	107	5,92
1sbyA	226	2	3	23	0,89	0,92	0,88	0,98	0,89	254	10,99
2e3rA	205	5	3	14	0,76	0,74	0,82	0,96	0,74	227	8,74
1a8tA	214	1	8	7	0,62	0,88	0,47	0,96	0,47	230	9,8
2hk1A	276	1	5	7	0,7	0,88	0,58	0,98	0,59	289	13,88
1yfrA	421	17	3	7	0,43	0,29	0,7	0,96	0,31	448	21,9
3a5rA	357	11	4	4	0,35	0,27	0,5	0,96	0,29	376	12,36
1lspA	171	1	9	4	0,48	0,8	0,31	0,95	0,31	185	8,15
1xz8A	149	1	10	2	0,31	0,67	0,17	0,93	0,17	162	5,89
2irxA	266	3	4	10	0,73	0,77	0,71	0,98	0,73	283	9,35
1vcjA	365	10	1	13	0,71	0,57	0,93	0,97	0,57	389	21,9
1xnyA	491	4	24	2	0,14	0,33	0,08	0,95	0,08	521	24,15
Média					0,61	0,69	0,61	0,96	0,53	286,95	14,73

B.3 Resultados: base de dados B44

Resultados do GRaSP para cada proteína do conjunto de dados B44. (**TN** = verdadeiro negativo; **FP** = falso positivo; **FN** = falso negativo; **TP** = verdadeiro positivo; **prec.** = precisão; **rec.** = revocação; **acc.** = acurácia; **len** = tamanho da cadeia)

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1a6w	215	5	2	7	0,66	0,58	0,78	0,97	0,6	229	16.87
1ahc	236	3	0	7	0,83	0,70	1,00	0,99	0,71	246	10.2
1bbs	324	7	9	15	0,63	0,68	0,63	0,95	0,64	334	22.7
1bid	241	8	4	11	0,63	0,58	0,73	0,95	0,59	264	13.17
1brq	156	11	2	6	0,48	0,35	0,75	0,93	0,37	175	5.71
1bya	467	5	4	15	0,76	0,75	0,79	0,98	0,76	491	12.67
1cge	146	1	3	12	0,85	0,92	0,80	0,98	0,8	162	6.64
1chg	221	1	5	3	0,52	0,75	0,38	0,97	0,38	245	15.44
1djb	243	4	3	7	0,65	0,64	0,70	0,97	0,65	260	10.89
1esa	224	3	4	9	0,71	0,75	0,69	0,97	0,7	240	17.27
1gcg	295	5	3	6	0,59	0,55	0,67	0,97	0,56	309	9.15
1hel	118	1	6	4	0,54	0,80	0,40	0,95	0,4	129	6.3
1hsi	174	11	4	9	0,52	0,45	0,69	0,92	0,47	198	10.72
1hxf	307	4	3	12	0,76	0,75	0,80	0,98	0,76	288	19.29
1ifb	115	4	6	6	0,51	0,60	0,50	0,92	0,51	131	5.69
1ime	260	1	1	11	0,91	0,92	0,92	0,99	0,92	273	08.06
1krn	72	1	2	4	0,71	0,80	0,67	0,96	0,67	79	4.12
1l3f	301	9	0	6	0,62	0,40	1,00	0,97	0,41	316	11.95
1mtw	207	4	1	11	0,81	0,73	0,92	0,98	0,75	233	16.25
1npc	300	6	5	6	0,50	0,50	0,55	0,97	0,51	317	11.29
1okm	242	4	1	10	0,80	0,71	0,91	0,98	0,72	258	11.97
1pdy	421	8	0	4	0,57	0,33	1,00	0,98	0,35	433	11.44
1phc	375	5	4	21	0,81	0,81	0,84	0,98	0,81	405	24.49
1psn	309	6	3	8	0,63	0,57	0,73	0,97	0,59	326	20.36
1pts	206	1	4	20	0,88	0,95	0,83	0,98	0,83	242	7.21
1qif	510	14	2	6	0,46	0,30	0,75	0,97	0,32	532	25.04
1qpe	247	3	9	11	0,64	0,79	0,55	0,96	0,56	271	20.83
1stn	125	4	0	7	0,79	0,64	1,00	0,97	0,65	136	4.71
1swb	105	2	1	12	0,88	0,86	0,92	0,98	0,87	120	4.63
1ula	272	8	3	6	0,52	0,43	0,67	0,96	0,45	289	10.39
1ypi	236	4	1	6	0,71	0,60	0,86	0,98	0,61	247	9.35
2ctb	292	8	0	7	0,67	0,47	1,00	0,97	0,49	307	10.75
2ctv	226	4	1	6	0,71	0,60	0,86	0,98	0,61	237	10.85
2fbp	292	3	14	8	0,49	0,73	0,36	0,95	0,37	330	9.13
2h4n	245	4	0	8	0,81	0,67	1,00	0,98	0,68	258	11.87
2sil	368	3	1	9	0,82	0,75	0,90	0,99	0,76	381	14.82
3app	299	22	1	1	0,13	0,04	0,50	0,93	0,04	323	20.6
3p2p	103	0	7	9	0,73	1,00	0,56	0,94	0,56	124	5.69
3phv	176	10	2	10	0,62	0,50	0,83	0,94	0,51	198	11.2
3ptn	208	8	1	6	0,59	0,43	0,86	0,96	0,45	233	16.96
5cpa	289	3	4	11	0,75	0,79	0,73	0,98	0,74	307	10.85
5dfr	142	4	0	8	0,81	0,67	1,00	0,97	0,67	159	10.64
7rat	113	4	1	6	0,70	0,60	0,86	0,96	0,62	124	5.39
8rat	113	4	0	7	0,78	0,64	1,00	0,97	0,65	124	5.44
Média					0,67	0,64	0,77	0,97		256,43	12.17

B.4 Resultados: base de dados U44

Resultados do GRaSP para cada proteína do conjunto de dados U44. (**TN** = verdadeiro negativo; **FP** = falso positivo; **FN** = falso negativo; **TP** = verdadeiro positivo; **prec.** = precisão; **acc.** = acurácia; **rec.** = revocação; **len** = tamanho da cadeia)

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1a6u	243	4	0	8	0,81	0,67	1,00	0,98	0,68	228	17.98
1acj	275	5	3	6	0,59	0,55	0,67	0,97	0,56	532	24.18
1apu	211	8	2	7	0,58	0,47	0,78	0,96	0,49	323	20.1
1blh	241	6	3	7	0,60	0,54	0,70	0,96	0,56	260	10.91
1byb	123	5	1	6	0,66	0,55	0,86	0,96	0,56	491	12.4
1dwd	200	17	3	12	0,54	0,41	0,80	0,91	0,42	299	19.4
1fbp	104	4	0	13	0,86	0,76	1,00	0,97	0,78	330	8.87
1gca	302	19	0	2	0,30	0,10	1,00	0,94	0,1	309	8.94
1hew	112	5	0	7	0,75	0,58	1,00	0,96	0,6	129	6.14
1hfc	108	9	0	7	0,64	0,44	1,00	0,93	0,45	157	6.28
1hyt	328	3	4	11	0,75	0,79	0,73	0,98	0,74	316	11.15
1ida	115	4	7	5	0,44	0,56	0,42	0,92	0,43	199	11.62
1imb	290	5	6	16	0,73	0,76	0,73	0,97	0,74	273	8.41
1inc	171	15	2	10	0,54	0,40	0,83	0,91	0,41	241	17.67
1mrg	147	4	0	8	0,81	0,67	1,00	0,97	0,68	246	10.43
1pdz	245	5	9	11	0,59	0,69	0,55	0,95	0,56	433	10.89
1phd	312	6	7	17	0,70	0,74	0,71	0,96	0,72	405	24.74
1pso	209	7	1	6	0,61	0,46	0,86	0,96	0,49	330	21.83
1rbp	235	4	0	7	0,79	0,64	1,00	0,98	0,65	175	6.16
1rne	509	11	3	5	0,43	0,31	0,63	0,97	0,34	334	21.98
1rob	290	10	0	7	0,63	0,41	1,00	0,97	0,43	124	5.45
1snc	100	3	6	10	0,65	0,77	0,63	0,92	0,64	135	4.27
1srf	242	7	6	9	0,55	0,56	0,60	0,95	0,57	238	7.57
1stp	228	1	2	6	0,80	0,86	0,75	0,99	0,75	121	4.61
1ulb	244	3	1	10	0,83	0,77	0,91	0,98	0,77	289	9.95
2cba	316	2	8	3	0,39	0,60	0,27	0,97	0,28	259	12.18
2ctc	292	13	4	7	0,45	0,35	0,64	0,95	0,36	307	10.99
2ifb	421	8	0	4	0,57	0,33	1,00	0,98	0,34	131	5.36
2pk4	226	4	1	6	0,71	0,60	0,86	0,98	0,62	82	4.12
2sim	300	10	0	6	0,60	0,38	1,00	0,97	0,38	381	14.93
2tga	236	4	0	7	0,79	0,64	1,00	0,98	0,65	233	16.59
2tmn	263	1	6	6	0,64	0,86	0,50	0,97	0,51	316	11.51
2ypi	162	5	1	7	0,70	0,58	0,88	0,97	0,6	247	9.4
3gch	74	1	2	4	0,71	0,80	0,67	0,96	0,67	245	15.93
3lck	142	0	3	12	0,89	1,00	0,80	0,98	0,8	271	21.6
3ptb	463	9	1	18	0,79	0,67	0,95	0,98	0,68	233	15.82
3tms	288	4	3	12	0,76	0,75	0,80	0,98	0,76	264	13.57
4ca2	376	4	3	22	0,85	0,85	0,88	0,98	0,85	256	11.62
4dfr	295	5	1	8	0,73	0,62	0,89	0,98	0,63	159	10.82
4phv	179	11	1	8	0,59	0,42	0,89	0,94	0,43	198	11.51
5cna	209	2	4	8	0,72	0,80	0,67	0,97	0,67	237	10.78
5p2p	368	3	0	10	0,87	0,77	1,00	0,99	0,78	124	5.88
6rsa	255	6	1	11	0,76	0,65	0,92	0,97	0,66	124	5.49
7cpa	118	1	4	6	0,70	0,86	0,60	0,96	0,6	307	10.36
Média					0,67	0,61	0,80	0,96			12.20

B.5 Resultados do experimento CASP10

Resultados do GRaSP e demais métodos participantes do CASP10 (**TN** = verdadeiro negativo; **FP** = falso positivo; **FN** = falso negativo; **TP** = verdadeiro positivo).

Target	N	Group	TP	FP	FN	TN	MCC	Z-score	BDT score
T0652	1	SEOK-SERVER	11	0	2	219	0.916	1.295	0.910
	2	SEOK	11	0	2	219	0.916	1.295	0.910
	3	INTFOLD2	12	3	1	216	0.850	0.694	0.830
	4	MCGUFFIN	12	3	1	216	0.850	0.694	0.830
	5	ATOME2_CBS	13	5	0	214	0.840	0.599	0.720
	6	FIRESTAR	12	4	1	215	0.821	0.425	0.770
	7	3DLIGANDSITE2	12	4	1	215	0.821	0.425	0.770
	8	HHPREDA	12	4	1	215	0.821	0.425	0.770
	9	SP-ALIGN	11	3	2	216	0.804	0.267	0.840
	10	COFACTOR_HUMAN	11	4	2	215	0.774	-0.007	0.790
	11	COFACTOR	11	4	2	215	0.774	-0.007	0.790
	12	CNIO	12	8	1	211	0.726	-0.447	0.620
	13	CONPRED-UCL	11	7	2	212	0.700	-0.691	0.650
	14	CHUO-BINDING-SITES	9	4	4	215	0.674	-0.930	0.810
	15	BINDING_KIHARA	8	6	5	213	0.568	-1.907	0.650
	16	GraSP	4	0	9	219	0.544	-2.130	0.364
T0657	1	BINDING_KIHARA	4	0	0	150	1.000	1.949	1.000
	2	GraSP	4	0	0	150	1.000	1.949	1.000
	3	SP-ALIGN	4	1	0	149	0.891	1.668	0.800
	4	CNIO	4	14	0	136	0.449	0.524	0.220
	5	COFACTOR_HUMAN	4	15	0	135	0.435	0.489	0.210
	6	COFACTOR	4	15	0	135	0.435	0.489	0.210
	7	FIRESTAR	4	16	0	134	0.423	0.457	0.200
	8	CHUO-BINDING-SITES	4	44	0	106	0.243	-0.008	0.080
	9	SEOK-SERVER	0	7	4	143	-0.036	-0.728	0.020
	10	SEOK	0	8	4	142	-0.038	-0.735	0.020
	11	3DLIGANDSITE2	0	9	4	141	-0.041	-0.741	0.010
	12	3DLIGANDSITE	0	9	4	141	-0.041	-0.741	0.010
	13	INTFOLD2	0	9	4	141	-0.041	-0.741	0.010
	14	MCGUFFIN	0	9	4	141	-0.041	-0.741	0.010
	15	FNGUSHAK	0	13	4	137	-0.050	-0.764	0.010
	16	ATOME2_CBS	0	14	4	136	-0.052	-0.769	0.010
	17	HHPREDA	0	15	4	135	-0.054	-0.774	0.010
	18	CONPRED-UCL	0	17	4	133	-0.058	-0.784	0.010
T0659	1	SP-ALIGN	3	3	0	68	0.692	1.928	0.500
	2	HHPREDA	2	1	1	70	0.653	1.788	0.730
	3	FNGUSHAK	1	6	2	65	0.168	0.076	0.180
	4	CONPRED-UCL	0	0	3	71	0.000	-0.517	0.000
	5	FIRESTAR	0	0	3	71	0.000	-0.517	0.000
	6	BINDING_KIHARA	0	0	3	71	0.000	-0.517	0.000
	7	INTFOLD2	0	0	3	71	0.000	-0.517	0.000
	8	MCGUFFIN	0	0	3	71	0.000	-0.517	0.000
	9	CHUO-BINDING-SITES	0	0	3	71	0.000	-0.517	0.000

Continued on next page

Tabela B.4 – *Continued from previous page*

Target	N	Group	TP	FP	FN	TN	MCC	Z-score	BDT score
	10	COFACTOR	0	4	3	67	-0.049	-0.690	0.060
T0675	1	COFACTOR_ HUMAN	8	0	0	67	1.000	0.990	1.000
	2	SEOK	8	0	0	67	1.000	0.990	1.000
	3	CNIO	8	0	0	67	1.000	0.990	1.000
	4	GraSP	8	0	0	67	1.000	0.990	1.000
	5	BINDING_ KIHARA	8	1	0	66	0.936	0.775	0.890
	6	FIRESTAR	7	0	1	67	0.929	0.750	0.900
	7	SEOK-SERVER	6	1	2	66	0.780	0.253	0.820
	8	INTFOLD2	4	0	4	67	0.687	-0.059	0.510
	9	MCGUFFIN	4	0	4	67	0.687	-0.059	0.510
	10	SP-ALIGN	4	0	4	67	0.687	-0.059	0.510
	11	CONPRED-UCL	6	4	2	63	0.627	-0.260	0.660
	12	FNGUSHAK	3	1	5	66	0.495	-0.703	0.430
	13	COFACTOR	5	6	3	61	0.467	-0.795	0.500
	14	CHUO-BINDING-SITES	7	29	1	38	0.273	-1.444	0.210
	15	ATOME2_ CBS	0	0	8	67	0.000	-2.359	0.000
T0686	1	GraSP	3	0	0	251	1.000	1.421	1.000
	2	CONPRED-UCL	3	1	0	250	0.864	0.912	0.750
	3	3DLIGANDSITE	3	1	0	250	0.864	0.912	0.750
	4	MCGUFFIN	3	1	0	250	0.864	0.912	0.750
	5	FIRESTAR	3	2	0	249	0.772	0.565	0.600
	6	COFACTOR_ HUMAN	3	2	0	249	0.772	0.565	0.600
	7	INTFOLD2	3	2	0	249	0.772	0.565	0.600
	8	SEOK	3	2	0	249	0.772	0.565	0.600
	9	3DLIGANDSITE2	2	1	1	250	0.663	0.157	0.750
	10	SP-ALIGN	2	1	1	250	0.663	0.157	0.750
	11	HHPREDA	3	4	0	247	0.649	0.107	0.430
	12	SEOK-SERVER	2	2	1	249	0.572	-0.185	0.560
	13	CNIO	2	2	1	249	0.572	-0.185	0.560
	14	FNGUSHAK	3	6	0	245	0.570	-0.189	0.330
	15	COFACTOR	3	14	0	237	0.408	-0.797	0.180
	16	BINDING_ KIHARA	1	5	2	246	0.223	-1.492	0.240
	17	CHUO-BINDING-SITES	3	67	0	184	0.177	-1.663	0.040
	18	ATOME2_ CBS	0	0	3	251	0.000	-2.327	0.000
T0696	1	FIRESTAR	3	0	0	97	1.000	1.517	1.000
	2	BINDING_ KIHARA	3	0	0	97	1.000	1.517	1.000
	3	CNIO	3	1	0	96	0.862	1.021	0.750
	4	SP-ALIGN	2	0	1	97	0.812	0.845	0.700
	5	GraSP	2	0	1	97	0.812	0.845	0.700
	6	FNGUSHAK	3	3	0	94	0.696	0.428	0.500
	7	COFACTOR_ HUMAN	2	1	1	96	0.656	0.286	0.700
	8	COFACTOR	2	1	1	96	0.656	0.286	0.700
	9	SEOK-SERVER	2	1	1	96	0.656	0.286	0.700
	10	CONPRED-UCL	3	6	0	91	0.559	-0.062	0.330
	11	HHPREDA	3	11	0	86	0.436	-0.504	0.210
	12	SEOK	2	5	1	92	0.411	-0.592	0.310
	13	3DLIGANDSITE	1	2	2	95	0.313	-0.945	0.450
	14	INTFOLD2	1	2	2	95	0.313	-0.945	0.420
	15	MCGUFFIN	1	2	2	95	0.313	-0.945	0.420
	14	CHUO-BINDING-SITES	3	22	0	75	0.305	-0.974	0.120
	16	ATOME2_ CBS	0	0	3	97	0.000	-2.065	0.000
T0697	1	COFACTOR_ HUMAN	13	1	1	447	0.926	0.793	0.930
	2	COFACTOR	13	1	1	447	0.926	0.793	0.930

Continued on next page

Tabela B.4 – *Continued from previous page*

Target	N	Group	TP	FP	FN	TN	MCC	Z-score	BDT score
	3	ATOME2_CBS	12	1	2	447	0.886	0.626	0.890
	4	SEOK-SERVER	11	0	3	448	0.883	0.615	0.830
	5	INTFOLD2	12	2	2	446	0.853	0.487	0.870
	6	MCGUFFIN	12	2	2	446	0.853	0.487	0.870
	7	HPREDA	11	1	3	447	0.844	0.453	0.820
	8	SEOK	10	0	4	448	0.841	0.440	0.780
	9	FNGUSHAK	13	4	1	444	0.837	0.423	0.770
	10	FIRESTAR	12	3	2	445	0.823	0.362	0.820
	11	3DLIGANDSITE2	12	3	2	445	0.823	0.362	0.820
	12	3DLIGANDSITE	12	3	2	445	0.823	0.362	0.820
	13	SP-ALIGN	12	4	2	444	0.795	0.248	0.780
	14	CNIO	13	6	1	442	0.790	0.226	0.690
	15	GraSP	6	2	8	446	0.557	-0.742	0.473
	16	CHUO-BINDING-SITES	14	44	0	404	0.467	-1.119	0.240
	17	BINDING_KIHARA	2	6	12	442	0.170	-2.352	0.260
	18	CONPRED-UCL	1	2	13	446	0.143	-2.465	0.170
T0706	1	HPREDA	5	1	1	197	0.828	0.829	0.880
	2	COFACTOR	4	0	2	198	0.812	0.771	0.770
	3	SP-ALIGN	4	0	2	198	0.812	0.771	0.770
	4	CONPRED-UCL	4	1	2	197	0.723	0.444	0.770
	5	FIRESTAR	4	1	2	197	0.723	0.444	0.770
	6	SEOK-SERVER	4	1	2	197	0.723	0.444	0.770
	7	SEOK	4	1	2	197	0.723	0.444	0.770
	8	CNIO	4	1	2	197	0.723	0.444	0.770
	9	COFACTOR_HUMAN	5	3	1	195	0.712	0.404	0.670
	10	INTFOLD2	4	2	2	196	0.657	0.200	0.770
	11	MCGUFFIN	4	2	2	196	0.657	0.200	0.770
	12	FNGUSHAK	3	1	3	197	0.603	0.005	0.570
	13	CHUO-BINDING-SITES	5	23	1	175	0.352	-0.914	0.190
	14	ATOME2_CBS	0	0	6	198	0.000	-2.204	0.000
	15	BINDING_KIHARA	0	3	6	195	-0.021	-2.281	0.020
T0720	1	HPREDA	9	1	7	185	0.694	1.692	0.620
	2	CNIO	8	0	8	186	0.692	1.687	0.630
	3	GraSP	6	0	8	188	0.641	1.448	0.428
	4	SP-ALIGN	5	0	11	186	0.543	0.992	0.430
	5	FIRESTAR	4	0	12	186	0.485	0.719	0.270
	6	SEOK	4	1	12	185	0.425	0.443	0.270
	7	MCGUFFIN	3	3	13	183	0.273	-0.267	0.240
	8	BINDING_KIHARA	2	1	14	185	0.267	-0.293	0.260
	9	FNGUSHAK	4	7	12	179	0.253	-0.360	0.290
	10	CHUO-BINDING-SITES	5	13	11	173	0.230	-0.466	0.310
	11	INTFOLD2	3	5	13	181	0.222	-0.501	0.250
	12	3DLIGANDSITE2	2	2	14	184	0.221	-0.505	0.310
	13	COFACTOR_HUMAN	2	2	14	184	0.221	-0.505	0.190
	14	COFACTOR	2	2	14	184	0.221	-0.505	0.190
	15	SEOK-SERVER	2	2	14	184	0.221	-0.505	0.180
	16	CONPRED-UCL	0	0	16	186	0.000	-1.536	0.000
	17	ATOME2_CBS	0	0	16	186	0.000	-1.536	0.000
T0721	1	HPREDA	29	5	2	263	0.880	1.356	0.860
	2	INTFOLD2	25	4	6	264	0.815	0.887	0.850
	3	CNIO	27	8	4	260	0.798	0.759	0.790
	4	FNGUSHAK	26	7	5	261	0.791	0.709	0.810
	5	SP-ALIGN	28	11	3	257	0.780	0.636	0.740

Continued on next page

Tabela B.4 – *Continued from previous page*

Target	N	Group	TP	FP	FN	TN	MCC	Z-score	BDT score
	6	COFACTOR_HUMAN	23	5	8	263	0.757	0.466	0.810
	7	COFACTOR	23	5	8	263	0.757	0.466	0.810
	8	MCGUFFIN	21	3	10	265	0.747	0.399	0.750
	9	3DLIGANDSITE	29	16	2	252	0.747	0.393	0.650
	10	FIRESTAR	23	7	8	261	0.726	0.246	0.810
	11	ATOME2_CBS	22	8	9	260	0.690	-0.017	0.780
	12	SEOK	22	8	9	260	0.690	-0.017	0.790
	13	3DLIGANDSITE2	26	16	5	252	0.683	-0.063	0.660
	14	SEOK-SERVER	21	7	10	261	0.681	-0.077	0.770
	15	CONPRED-UCL	20	11	11	257	0.604	-0.634	0.710
	16	CHUO-BINDING-SITES	31	46	0	222	0.577	-0.826	0.400
	17	GraSP	6	1	25	267	0.383	-2.229	0.200
	18	BINDING_KIHARA	6	2	25	266	0.352	-2.454	0.310
T0726	1	FIRESTAR	3	0	0	584	1.000	1.467	1.000
	2	SEOK-SERVER	3	0	0	584	1.000	1.467	1.000
	3	MCGUFFIN	3	0	0	584	1.000	1.467	1.000
	4	INTFOLD2	3	2	0	582	0.773	0.645	0.600
	5	SEOK	3	2	0	582	0.773	0.645	0.600
	6	GraSP	3	2	0	582	0.773	0.645	0.600
	7	3DLIGANDSITE	3	3	0	581	0.705	0.398	0.500
	8	CONPRED-UCL	3	5	0	579	0.610	0.052	0.380
	9	3DLIGANDSITE2	3	5	0	579	0.610	0.052	0.380
	10	COFACTOR_HUMAN	3	7	0	577	0.544	-0.185	0.300
	11	COFACTOR	3	7	0	577	0.544	-0.185	0.300
	12	HHPREDA	3	8	0	576	0.519	-0.279	0.270
	13	CNIO	3	8	0	576	0.519	-0.279	0.270
	14	FNGUSHAK	3	10	0	574	0.476	-0.432	0.230
	15	SP-ALIGN	3	18	0	566	0.372	-0.810	0.140
	16	ATOME2_CBS	3	21	0	563	0.347	-0.901	0.130
	17	CHUO-BINDING-SITES	3	97	0	487	0.158	-1.586	0.030
	18	BINDING_KIHARA	0	3	3	581	-0.005	-2.178	0.040
T0737	1	3DLIGANDSITE	20	2	2	229	0.900	1.320	0.920
	2	MCGUFFIN	18	1	4	230	0.870	1.079	0.860
	3	INTFOLD2	18	2	4	229	0.845	0.882	0.860
	4	CNIO	18	2	4	229	0.845	0.882	0.860
	5	SEOK-SERVER	16	1	6	230	0.814	0.630	0.790
	6	SEOK	16	1	6	230	0.814	0.630	0.790
	7	3DLIGANDSITE2	17	4	5	227	0.772	0.295	0.830
	8	COFACTOR_HUMAN	17	4	5	227	0.772	0.295	0.830
	9	COFACTOR	17	4	5	227	0.772	0.295	0.830
	10	FIRESTAR	16	3	6	228	0.764	0.233	0.770
	11	ATOME2_CBS	20	10	2	221	0.755	0.160	0.680
	12	HHPREDA	16	4	6	227	0.741	0.055	0.800
	13	FNGUSHAK	18	9	4	222	0.711	-0.185	0.710
	14	SP-ALIGN	16	7	6	224	0.683	-0.408	0.770
	15	CONPRED-UCL	18	14	4	217	0.642	-0.733	0.580
	16	GraSP	11	3	11	228	0.600	-1.067	0.537
	17	BINDING_KIHARA	5	0	17	231	0.460	-2.182	0.370
	18	CHUO-BINDING-SITES	19	40	3	191	0.460	-2.182	0.340
T0744	1	FIRESTAR	15	2	4	306	0.825	1.579	0.830
	2	CNIO	15	4	4	304	0.776	1.289	0.830
	3	FNGUSHAK	16	6	3	302	0.768	1.240	0.750
	4	3DLIGANDSITE2	16	9	3	299	0.716	0.926	0.660

Continued on next page

Tabela B.4 – *Continued from previous page*

Target	N	Group	TP	FP	FN	TN	MCC	Z-score	BDT score
	5	SP-ALIGN	16	13	3	295	0.658	0.583	0.570
	6	INTFOLD2	10	2	9	306	0.647	0.515	0.580
	7	MCGUFFIN	9	1	10	307	0.639	0.470	0.550
	8	3DLIGANDSITE	10	3	9	305	0.619	0.347	0.590
	9	COFACTOR_HUMAN	12	7	7	301	0.609	0.289	0.710
	10	COFACTOR	12	7	7	301	0.609	0.289	0.710
	11	CONPRED-UCL	13	15	6	293	0.531	-0.174	0.510
	12	SEOK	9	7	10	301	0.489	-0.426	0.540
	13	HHPREDA	9	8	10	300	0.472	-0.529	0.580
	14	ATOME2_CBS	9	10	10	298	0.441	-0.711	0.540
	15	CHUO-BINDING-SITES	14	38	5	270	0.392	-1.002	0.290
	16	SEOK-SERVER	7	13	12	295	0.318	-1.443	0.430
	17	GraSP	3	2	16	306	0.289	-1.621	0.320
	18	BINDING_KIHARA	3	2	16	306	0.289	-1.621	0.320

B.6 Resultados: base de dados ASTEX

Resultados do GRASP para cada proteína do conjunto de dados ASTEX. (**TN** = verdadeiro negativo; **FP** = falso positivo; **FN** = falso negativo; **TP** = verdadeiro positivo; **prec.** = precisão; **rec.** = revocação; **len** = tamanho da cadeia)

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1k3u	613	14	8	27	0,70	0,66	0,77	0,97	0,67	662	32,13
1n1m	680	8	20	18	0,55	0,69	0,47	0,96	0,48	726	42,79
1hnn	232	2	16	11	0,56	0,85	0,41	0,93	0,41	261	10,77
1ywr	318	7	1	12	0,75	0,63	0,92	0,98	0,65	348	34,87
1jd0	250	1	0	9	0,95	0,90	1,00	1,00	0,90	265	19,97
1oyt	300	3	16	11	0,54	0,79	0,41	0,94	0,41	285	27,37
1m2z	255	5	4	12	0,71	0,71	0,75	0,97	0,72	276	27,24
1pmn	323	5	3	13	0,75	0,72	0,81	0,98	0,74	356	32,30
1xoq	307	4	1	14	0,84	0,78	0,93	0,98	0,79	326	23,54
1l2s	673	13	10	17	0,58	0,57	0,63	0,97	0,58	716	31,32
1sq5	276	5	10	11	0,57	0,69	0,52	0,95	0,53	308	11,02
1p62	202	1	12	14	0,68	0,93	0,54	0,94	0,54	241	9,32
1of6	333	5	6	6	0,51	0,55	0,50	0,97	0,51	351	13,41
1s19	231	5	0	17	0,87	0,77	1,00	0,98	0,78	304	31,83
1vcj	365	10	1	13	0,71	0,57	0,93	0,97	0,57	389	21,62
1n2j	250	10	15	12	0,45	0,55	0,44	0,91	0,46	287	11,18
1y6b	251	0	2	13	0,93	1,00	0,87	0,99	0,87	350	28,98
1r1h	668	6	8	14	0,66	0,70	0,64	0,98	0,65	696	33,93
2br1	247	6	7	12	0,62	0,67	0,63	0,95	0,65	275	38,39
1t9b	517	14	12	40	0,73	0,74	0,77	0,96	0,75	603	30,16
1x8x	302	5	5	10	0,65	0,67	0,67	0,97	0,68	322	18,07
1r58	351	5	1	12	0,80	0,71	0,92	0,98	0,71	369	18,58
1unl	423	5	1	13	0,81	0,72	0,93	0,99	0,73	442	34,28
1yqy	495	5	3	11	0,73	0,69	0,79	0,98	0,70	514	20,42
1q41	322	5	0	12	0,83	0,71	1,00	0,99	0,72	352	33,68
1r9o	414	6	13	22	0,68	0,79	0,63	0,96	0,64	467	47,26
1s3v	159	3	10	14	0,66	0,82	0,58	0,93	0,59	186	16,37
1v48	243	2	3	13	0,83	0,87	0,81	0,98	0,82	283	15,77
1t46	269	3	12	13	0,63	0,81	0,52	0,95	0,53	369	34,60
1w2g	180	1	7	9	0,69	0,90	0,56	0,96	0,57	209	11,12
1xoz	308	2	4	12	0,79	0,86	0,75	0,98	0,76	326	30,79
1q4g	473	7	47	26	0,49	0,79	0,36	0,90	0,36	553	25,82
1hwi	363	16	1	13	0,63	0,45	0,93	0,96	0,45	401	14,50
1tow	116	5	0	10	0,80	0,67	1,00	0,96	0,68	131	6,32
1jje	200	2	6	12	0,74	0,86	0,67	0,96	0,67	220	9,48
1hvy	261	0	9	17	0,80	1,00	0,65	0,97	0,65	288	17,81
1tt1	238	1	4	8	0,76	0,89	0,67	0,98	0,67	251	19,96
1hq2	126	2	11	19	0,71	0,90	0,63	0,92	0,64	158	6,54
1gpk	499	10	9	11	0,52	0,52	0,55	0,96	0,54	532	37,68
1w1p	443	7	35	13	0,38	0,65	0,27	0,92	0,28	498	23,16
1opk	414	5	14	16	0,62	0,76	0,53	0,96	0,54	449	33,62
1tz8	422	9	15	15	0,53	0,62	0,50	0,95	0,51	461	16,96
1uml	325	8	2	14	0,73	0,64	0,88	0,97	0,65	349	18,97
1uou	417	8	5	8	0,54	0,50	0,62	0,97	0,52	448	15,02
1sqn	229	6	2	11	0,72	0,65	0,85	0,97	0,66	251	25,50
1q1g	228	1	2	12	0,88	0,92	0,86	0,99	0,86	243	13,59

Continua na próxima página

Tabela B.5 – Continuação da página anterior

Target	TN	FP	FN	TP	MCC	prec.	rec.	acc.	BDT	len	time(s)
1v0p	258	2	7	10	0,68	0,83	0,59	0,97	0,59	286	34,35
1n46	228	3	6	11	0,69	0,79	0,65	0,96	0,65	251	30,89
1yv3	664	4	11	24	0,76	0,86	0,69	0,98	0,69	746	39,60
1sg0	210	11	5	4	0,31	0,27	0,44	0,93	0,27	230	13,89
117f	345	4	27	12	0,45	0,75	0,31	0,92	0,31	390	18,15
1jla	926	8	10	10	0,52	0,56	0,50	0,98	0,51	988	48,32
1mmv	380	5	5	17	0,76	0,77	0,77	0,98	0,78	418	30,32
1v4s	411	5	14	18	0,64	0,78	0,56	0,96	0,57	448	17,41
1sj0	225	2	6	10	0,71	0,83	0,62	0,97	0,63	245	28,44
1yvf	524	12	13	15	0,52	0,56	0,54	0,96	0,55	564	31,85
1owe	225	2	7	11	0,70	0,85	0,61	0,96	0,62	258	25,49
1mzc	684	6	17	13	0,53	0,68	0,43	0,97	0,44	720	38,57
1u1c	231	4	3	15	0,80	0,79	0,83	0,97	0,80	253	18,53
1ig3	450	1	39	7	0,35	0,88	0,15	0,92	0,15	497	23,42
1of1	287	3	5	13	0,75	0,81	0,72	0,97	0,73	330	15,67
1lpz	263	1	6	17	0,82	0,94	0,74	0,98	0,74	289	26,68
1hp0	300	5	3	11	0,72	0,69	0,79	0,97	0,70	328	12,18
1r55	183	0	11	9	0,65	1,00	0,45	0,95	0,45	203	9,56
1gm8	729	19	8	8	0,37	0,30	0,50	0,96	0,31	764	28,70
1oq5	237	8	2	9	0,64	0,53	0,82	0,96	0,54	257	16,95
1t40	278	1	21	16	0,61	0,94	0,43	0,93	0,43	316	22,10
1lrh	144	4	7	5	0,45	0,56	0,42	0,93	0,43	160	10,14
1ygc	272	2	37	14	0,45	0,88	0,27	0,88	0,28	318	33,49
1j3j	505	2	14	26	0,76	0,93	0,65	0,97	0,65	557	44,94
2bm2	1074	36	9	21	0,49	0,37	0,70	0,96	0,38	992	50,90
1nav	232	1	9	11	0,69	0,92	0,55	0,96	0,55	263	29,52
1p2y	379	6	3	19	0,80	0,76	0,86	0,98	0,77	407	37,81
1meh	320	2	11	20	0,75	0,91	0,65	0,96	0,65	482	19,10
1z95	215	2	6	15	0,78	0,88	0,71	0,97	0,72	246	26,48
1u4d	243	4	2	9	0,74	0,69	0,82	0,98	0,71	273	27,60
1g9v	475	10	41	48	0,62	0,83	0,54	0,91	0,54	574	35,35
1hww	986	10	3	15	0,70	0,60	0,83	0,99	0,61	1014	52,65
1gkc	120	1	28	10	0,43	0,91	0,26	0,82	0,26	334	10,07
1ia1	162	2	15	13	0,59	0,87	0,46	0,91	0,47	192	15,39
1xm6	291	9	26	9	0,31	0,50	0,26	0,90	0,27	351	19,59
1n2v	353	6	3	10	0,68	0,62	0,77	0,98	0,64	372	20,10
2bsm	193	2	2	12	0,85	0,86	0,86	0,98	0,86	209	12,71
1kzk	142	3	28	25	0,57	0,89	0,47	0,84	0,47	198	14,33
1ke5	260	8	0	13	0,77	0,62	1,00	0,97	0,63	298	35,43
Média					0,66	0,74	0,65	0,96	0,59	389,62	24,76