UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

José Geraldo de Carvalho Pereira

"CARACTERIZAÇÃO DOS AMINOÁCIDOS DA INTERFACE PROTEÍNA-PROTEÍNA COM MAIOR CONTRIBUIÇÃO NA ENERGIA DE LIGAÇÃO E SUA PREDIÇÃO A PARTIR DOS DADOS ESTRUTURAIS"

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a) JOSÉ GERALDO DE CARVALHO PEREIRA e aprovada pela Comissão Julgadora.

Dissertação apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Genética de Biologia Molecular, na área de Bioinformática.

Orientador: Prof. Dr. Goran Neshich
Co-orientador: João Alexandre Ribeiro Gonçalves Barbosa

Campinas, 2012

FICHA CATALOGRÁFICA ELABORADA POR
ROBERTA CRISTINA DAL' EVEDOVE TARTAROTTI – CRB8/7430
BIBLIOTECA DO INSTITUTO DE BIOLOGIA - UNICAMP

P414c

Pereira, José Geraldo de Carvalho, 1984-
Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir de dados estruturais / José Geraldo de Carvalho Pereira. – Campinas, SP: [s.n.], 2012.

Orientador: Goran Neshich.
Coorientador: João Alexandre Ribeiro Gonçalves Barbosa.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Mapeamento de interação de proteínas. 2. *Hot spots*. 3. Varredura por alalinas. I. Neshich, Goran. II. Barbosa, João Alexandre Ribeiro Gonçalves. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

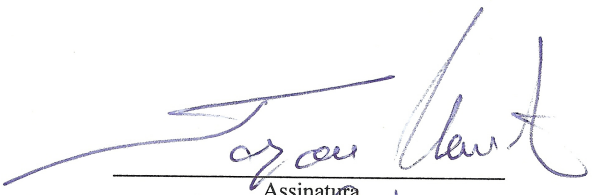
Informações para Biblioteca Digital

Título em Inglês: Characterization of amino acids from protein-protein interface with the highest contribution to the binding energy and its prediction from structural data
Palavras-chave em Inglês:
Protein interaction mapping
Hot spots
Alanine scanning
Área de concentração: Bioinformática
Titulação: Mestre em Genética e Biologia Molecular
Banca examinadora:
Goran Neshich [Orientador]
André Luis Berteli Ambrósio
Ricardo Aparicio
Data da defesa: 28-02-2012
Programa de Pós Graduação: Genética e Biologia Molecular

Campinas, 28 de fevereiro de 2012.

BANCA EXAMINADORA

Prof. Dr. Goran Neshich (Orientador)

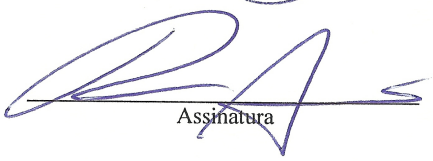


Assinatura

Prof. Dr. André Luis Berteli Ambrósio

Assinatura

Prof. Dr. Ricardo Aparicio



Assinatura

Prof. Dr. Renato Vicentini dos Santos

Assinatura

Prof. Dr. Fernando José Von Zuben

Assinatura

RESUMO

A propriedade das proteínas de se ligarem umas as outras de forma altamente específica, formando complexos estáveis, é uma característica fundamental para todos os processos biológicos. Uma melhor compreensão da formação do complexo abre perspectivas para muitas aplicações práticas, entre elas o design racional de novos fármacos. Trabalhos anteriores demonstraram, através de experimentos de varredura por alaninas, que um pequeno número de resíduos das interfaces proteicas contribui com a maior parte da energia de ligação e por isso foram chamados de *hot spots*. Devido à importância desses resíduos para as interações proteína-proteína, diversos métodos computacionais têm sido propostos para predizer os *hot spots* complementando assim o procedimento experimental. Entre esses, estão métodos *physics-based* como dinâmica molecular, e também métodos *knowledge-based,* onde dados experimentais são utilizados para treinar métodos computacionais que aprendem as regras para classificar corretamente os *hot spots* e usados posteriormente para classificar novos casos em estruturas de complexos proteicos. Entre os algoritmos de aprendizado computacional mais utilizados estão árvores de decisão, redes neurais, máquinas de vetor de suporte. Nesse trabalho, desenvolvemos métodos de predição de *hot spots* utilizando máquinas de vetor de suporte, que foram abastecidas na entrada com um conjunto de 186 descritores estruturais extraídos do banco de dados STING_DB e também com 112 novos descritores propostos neste trabalho. Os métodos propostos nesse trabalho apresentaram desempenho superior aos métodos de predição de *hot spots* mais conhecidos da literatura, como KFC, Minerva, Rosetta e FOLDEF. Além disso, a análise estatística dos descritores e também a seleção dos descritores mais eficientes na tarefa de classificar *hot spots* permitiu que observássemos diversas características que são distintas entre resíduos que são *hot spots* e os que não são. Entre estas características, a entalpia de hidratação ao redor do resíduo sugere que essa região é mais hidrofílica em *hot spots*. Essa região, que para *hot spots* é denominada de anel-O, tem a função de impedir o contato do solvente com o *hot spot* e por isso, alguns autores acreditavam tratar-se de uma região hidrofóbica, algo que os resultados deste trabalho não confirmaram. Futuramente, os novos descritores propostos neste trabalho serão agregados ao STING_DB e o método de predição de *hot spots* será integrado ao STING permitindo a predição de *hot spots* de todos os complexos proteicos depositados no Protein Data Bank (PDB) assim como de complexos proteicos fornecidos pelo usuário.

ABSTRACT

The property of the proteins to bind each other in a highly specific way, forming stable complexes, is a key feature for all biological processes. A better understanding of the formation of protein complexes provides many practical applications, including the rational design of new drugs. Through experiments of alanine scanning, it was shown that a small number of residues belonging to protein interfaces contribute decisively to the binding energy and so were called hot spots. Because of the importance of these residues for protein-protein interactions many computational methods have been proposed to predict the hot spots and thus complement the experimental procedure. These include physics-based methods such as molecular dynamics and also knowledge-based methods where experimental data are used to train computational methods that learn the rules for correctly classifying the hot spots and are then used to classify new cases in structures of protein complexes. Among the computational learning algorithms most frequently used are decision trees, neural networks, support vector machines, among others. In this work, we developed methods to predict hot spots using support vector machines, using at the input 186 structural descriptors extracted from the STING_DB and 112 new descriptors proposed in this work. The methods proposed here showed superior performance to methods of predicting hot spots best known from the literature, such as KFC, Minerva, Rosetta and FOLDEF.. In addition, statistical analysis of the descriptors and also the selection of the descriptors more efficient in the task of classifying hot spots allowed us to observe several characteristics that are distinct for residues that are hot spots. Among these features, the enthalpy of hydration suggests that the region around hot spots is more hydrophilic. This region, which for hot spots is called O-ring, serves to prevent the contact of the solvent with the hot spot and therefore some authors believe that this was a hydrophobic region whereas results presented here show otherwise. In future, the new descriptors described in this work will be added to the STING_DB and the method of prediction of hot spots will be integrated with STING allowing the prediction of hot spots of all protein complexes deposited in the Protein Data Bank (PDB) as well as protein complexes supplied by the user.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— Donald E. Knuth (1974)

AGRADECIMENTOS

Tenho a impressão que essa seja uma das partes mais complicadas para os autores. De certa forma, tantas pessoas influenciam a nossa vida e inevitavelmente nos lembramos de todas elas nos agradecimentos. Infelizmente, ou felizmente, tudo isso tem que ser resumido em poucas frases.

Obviamente, começarei pela minha família. Qualquer coisa que penso em escrever para agradecer a meus pais me parece pouco para descrever o quanto sou grato a eles, por isso vou resumir em uma frase curta, mas cheia de sentimentos. Amo muito vocês dois. Tenho que agradecer também minhas avós e meus avôs, assim como minhas tias e tios, primas e primo. Obrigado a todos vocês.

Agradeço também aos meus amigos que moraram comigo enquanto estive em Campinas, Renan (Johnny), Picelli, Gága, Jacó, Yuzo, Korea, foi divertido. Agradeço ao Paulo, com quem aprendi boa parte do que sei sobre computação e com quem adoro discutir os mais variados temas, apesar de quase nunca concordarmos um com o outro. Bem, eu quase nunca, você nunca!

Ao pessoal da Embrapa. Goran, obrigado pelo convite de me juntar ao seu grupo no momento em que estava muito preocupado com o mestrado, e pela compreensão nos momentos difíceis. Sou muito grato a você por isso e por tudo o que tive a oportunidade de aprender no lab. Ivan, Inácio, Izabella, Zé, Jardine obrigado por toda a ajuda que me deram. Fábio, foi bom sentar ao seu lado durante esse tempo, acho que aprendi muito com nossas conversas e também obrigado por toda a ajuda. Não sei onde estaremos daqui um tempo, mas espero que achemos uma forma de continuar com nossas conversas. Agradeço também ao pessoal com quem convivi no LNLS, João, Nádia, Tati, Camila, Germa, Carol, Daniel, e outros que não mencionei para não estender ainda mais esse agradecimento.

E por fim, agradeço a minha namorada. Obrigado por me enterder quando eu estava estressado e sempre me apoiar quando precisei. Sei que tenho muito mais a agradecer a você, mas isso eu farei pessoalmente. Amo você.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 INTERAÇÃO PROTEÍNA-PROTEÍNA

Proteínas são polímeros constituídos de uma sequência, de tamanho variado, de 20 aminoácidos quimicamente e estruturalmente diferentes, que se enovelam em uma estrutura tridimensional altamente específica. Esta complexidade permite que as proteínas interajam com quase todos os tipos de moléculas, como íons metálicos, açucares, sais inorgânicos, pequenas moléculas orgânicas, nucleotídeos e outras proteínas. Dentre estas, as interações proteína-proteína, apesar de serem estudadas a bastante tempo, tem recebido grande atenção nos anos recentes devido aos avanços na biologia de sistemas, área que busca mapear as interações entre as proteínas e estudar o funcionamento, a topologia, a dinâmica e a função fisiológica das redes formadas (Cusick et al., 2005). As proteínas interagem com outras proteínas com uma afinidade que varia de baixo milimolar até alto fentomolar, mas apesar dessa variação todas elas mantém um alto grau de especificidade por seus parceiros (Reichmann et al., 2007). Muitas proteínas possuem ainda a habilidade de interagir com múltiplos parceiros, tanto simultaneamente quanto separadamente, aumentando a complexidade e robustez das redes de interação (Tsai et al., 2009).

As interações que as proteínas fazem são fundamentais para quase todos os processos biológicos, sendo essenciais em todos os estágios de desenvolvimento e homeostase. Não surpreendentemente, muitas doenças podem ser atribuídas a interações proteína-proteína incorretas, ou impossibilitadas, tanto em relação à perda de uma interação atômica essencial ou a formação de um complexo em locais ou momentos inadequados. Entre os exemplos encontrados na literatura estão doenças degenerativas como a doença de Huntigton, leucemia e câncer cervical (Ryan e Matthews, 2005).

Estudos em larga escala revelam redes de interação proteína-proteína na célula através da identificação de prováveis parceiros de interação. No entanto, para uma melhor compreensão dessas interações, e para uma possível manipulação dessas interações, torna-se necessário identificar (1) as regiões da superfície proteica que irão interagir formando o complexo, e (2) os resíduos que mais contribuem para a ligação e estabilidade do complexo. A hipótese mais aceita atualmente indica que apenas poucos resíduos da interface proteína-proteína são realmente essenciais para a interação, por exemplo, em interfaces entre 1200 e 2000 $Å^2$, que é o tamanho padrão para as interfaces de acordo com Conte e colaboradores (1999), menos de 5% dos resíduos

A predição dos resíduos que compõem interfaces proteicas também é uma uma área de pesquisa do Grupo de Pesquisa em Biologia Computacional, CNPTIA, Embrapa. Esse trabalho é atualmente desenvolvido pelo aluno de doutorado Fábio R. de Moraes.

da interface contribuem reduzindo a energia livre em mais de 2 Kcal/-mol (Bogan e Thorn, 1998). Identificar quais são esses resíduos da interface essenciais para interação, no sentido da importante contribuição energética, e quais características os tornam tão importantes, além de permitir uma melhor compreensão dos mecanismos de interação proteína-proteína, pode ser essencial para indicar alvos de fármacos que poderiam regular uma função, impedindo a interação entre dois parceiros (Arkin e Wells, 2004), ou ainda propor soluções que aumentem ou diminuam dependendo do objetivo da pesquisa, tanto a afinidade quanto a especificidade da interação proteína-proteína de interesse. Está última aplicação é um dos casos estudados na recente área de *design* computacional de proteínas (Butterfoss e Kuhlman, 2006; Suárez e Jaramillo, 2009) e diversos artigos estão sendo publicados abordando especificamente as interações proteína-proteína (Kortemme et al., 2004; Humphris e Kortemme, 2007; Potapov et al., 2008; Karanicolas e Kuhlman, 2009; Fromer e Shifman, 2009).

1.2 OS *hot spots* DO SÍTIO DE LIGAÇÃO PROTEÍNA-PROTEÍNA

Clackson e Wells (1995) foram os pesquisadores que introduziram o conceito de *hot spots* de ligação através da observação que a maior parte da energia de ligação provém da interação de alguns poucos resíduos da interface (Reichmann et al., 2007). Este trabalho foi estendido posteriormente por Bogan e Thorn (1998), que utilizaram a biologia computacional para analisar diversos complexos e notaram que os *hot spots* são rodeados por resíduos de menor importância energética, formando um anel, denominado por eles de anel-O (*O-ring*), que provavelmente tem a função de proteger os resíduos com maior contribuição energética do contato com o solvente. Essa exclusão do solvente diminui a constante dielétrica local e aumenta a força das interações eletrostáticas e das ligações de hidrogênio (Deremble e Lavery, 2005). Consequentemente, para um resíduo ter uma maior contribuição energética na interação, ele deve estar menos sujeito ao contato com o solvente. No entanto, somente a ausência de contato com o solvente não é suficiente para o resíduo ter uma contribuição energética significativa, pois esta depende de outros fatores como o tipo e o número de interações. A Figura 1, extraída do trabalho de Bogan e Thorn (1998), sustenta a hipótese de que a ausência de contato com o solvente é necessária, mas não suficiente, para aumentar a contribuição energética de interação de um resíduo.

A oclusão do solvente favorecida pelo anel-O é uma hipótese comumente aceita e considerada como uma característica necessária à presença de um *hot spot*, no entanto essa hipótese ainda carece de evidências conclusivas (Moreira et al., 2007b), pois acredita-se que resíduos à margem da interface, isto é, próximos ao solvente, quando substituídos por alanina poderiam ter suas interações substituídas

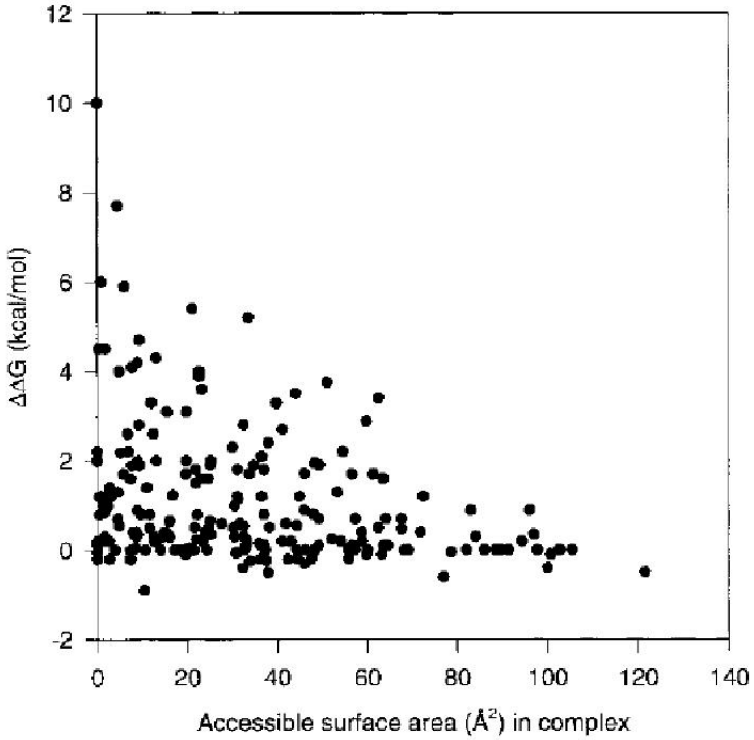


Figura 1: Correlação entre a variação da energia livre de ligação e a superfície acessível ao solvente em complexo dos resíduos da interface mostrando os dados que levaram a hipótese do anel-O (figura extraída de Bogan e Thorn (1998))

por moléculas de água que fariam uma ponte entre as duas proteínas (Janin, 1999). Devido a essa possível substituição das cadeias laterais dos resíduos por moléculas do solvente, mesmo que um resíduo apresentasse uma grande contribuição para a energia de ligação, ele seria classificado como não *hot spot* em um experimento de varredura por alaninas, pois uma baixa variação da energia de ligação seria observada no experimento (DeLano, 2002).

Bogan e Thorn (1998) também notaram que os hot spots apresentam preferência por alguns resíduos (Tabela 1), entre os mais frequentes estão o triptofano, presente em 21% dos *hot spots*, a arginina, em 13,3%, e a tirosina em 12,3%, enquanto outros como leucina, metionina, serina, treonina e valina raramente são *hot spots* apresentando frequência menor que 3% no banco de dados analisado. A grande frequência observada de resíduos de triptofano deve-se provavelmente ao seu grande tamanho e sua natureza aromática, apresentando característica hidrofóbica em uma boa parte da superfície de sua cadeia lateral, além da possibilidade de fazer uma ligação de hidrogênio e interações π (Bogan e Thorn, 1998). A tirosina, assim como o triptofano, tem regiões hidrofóbicas que podem estar expostas na superfície da proteína, e também a capacidade de fazer intera-

ções π devido ao anel aromático e ligações de hidrogênio através do grupo hidroxila. Está ultima característica provavelmente explica o fato da tirosina ser três vezes mais frequentemente classificada como *hot spots* que a fenilalanina (Bogan e Thorn, 1998). Já a alta frequência da arginina deve-se a sua capacidade de realizar muitas ligações de hidrogênio, e da interação eletrostática (Bogan e Thorn, 1998). A maior capacidade de realizar ligações de hidrogênio e a característica pseudoaromática do grupo guanidina, o qual permite interações π , são possivelmente os responsáveis pela arginina ser duas vezes mais frequentemente encontrada como hot spot que a lisina (Bogan e Thorn, 1998; Bahadur e Zacharias, 2008). Outras diferenças interessantes ocorrem entre o resíduo aspartato em relação ao glutamato, e a asparagina em relação a glutamina. Em ambos os casos, os resíduos aspartato e asparagina, que possuem uma cadeia lateral menor, são mais frequentes e segundo Bogan e Thorn (1998), isso ocorre provavelmente devido a menor entropia conformacional das cadeias laterais, pois após a formação do complexo haveria uma menor perda de entropia.

1.3 MÉTODOS EXPERIMENTAIS PARA DETERMINAR *hot spots*

Por definição, um resíduo é um hot spot quando, após ser substituído por alanina, ocorre um aumento na energia livre de ligação em pelo menos 2.0 Kcal/mol. Até onde conhecemos, essa definição foi proposta inicialmente no trabalho de Bogan e Thorn (1998) e não apresenta uma explicação detalhada do motivo da escolha desse valor específico. Entretanto, tal definição foi adotada pela maioria dos trabalhos sobre *hot spots.*

Normalmente, a medida da variação da energia livre de ligação é feita através de experimentos de varredura por alaninas onde resíduos são substituídos um a um por uma alanina. No entanto, esse experimento é muito demorado e trabalhoso, uma vez que cada proteína com um aminoácido substituído por alanina precisa ser expressa, purificada e analisada separadamente através de experimentos *in vitro* (Morrison e Weiss, 2001). Há ainda outros métodos como *shotgun scanning* que utiliza bibliotecas de *phage-display* e o método de substituição binomial que permite a modificação dos aminoácidos aspartato, glutamato, glicina, prolina, serina, valina e treonina para uma alanina através da mutação sítio específica de apenas um nucleotídeo do códon (Tabela 2), mas todos são baseados no princípio de modificar um determinado resíduo e calcular a variação de energia livre de ligação (Moreira et al., 2007b).

A mutação por alanina equivale à remoção dos átomos que estão além do carbono β em outros aminoácidos. A princípio, a glicina também poderia ser usada, no entanto, a sua maior flexibilidade estrutural, poderia causar modificações estruturais e energéticas no monô-

Tabela 1: Preferência dos aminoácidos em *hot spots* (Tabela adaptada de Bogan e Thorn 1998).

Resíduo	Banco de dados ASEdb		*Hot spots*		Enriquecimento
	número	%	número	%	
ARG	218	9,38	29	13,30	2,47
ASN	99	4,26	5	5,05	0,93
ASP	177	7,61	16	9,04	1,67
CYS	3	0,13	0	0	0
GLN	160	6,88	5	3,13	0,58
GLU	220	9,46	8	3,64	0,68
GLY	28	1,20	1	3,57	0,45
HIS	50	2,15	4	8,00	1,49
ILE	104	4,47	10	9,62	1,79
LEU	242	10,41	2	0,83	0,01
LYS	143	6,15	9	6,29	1,17
MET	69	2,97	2	2,90	0,54
PHE	166	7,14	5	3,01	0,56
PRO	89	3,83	6	6,74	1,25
SER	178	7,66	2	1,12	0,21
THR	131	5,63	2	1,53	0,28
TRP	19	0,82	4	21,05	3,91
TYR	122	5,25	15	12,30	2,29
VAL	107	4,60	0	0	0

mero dificultando a análise da variação de energia livre de ligação do complexo, e por isso, a mutação por alanina é a mais utilizada (Morrison e Weiss, 2001).

1.4 MÉTODOS COMPUTACIONAIS PARA PREDIZER *hot spots*

Devido à demora e ao alto custo envolvido nos métodos experimentais de determinação dos hot spots, diversos métodos computacionais tem sido propostos com o objetivo de predizer precisamente os resultados obtidos através do método de varredura por alaninas. Entre esses métodos computacionais encontramos alguns que utilizam apenas a sequência de aminoácidos de uma proteína, sem considerar qualquer informação do parceiro de interação e da região de interação (Ofran e Rost, 2007). No entanto, a grande maioria dos métodos computacionais utilizam dados estruturais do complexo para predizer os *hot spots*. Esses métodos podem ser classificados em dois tipos principais: (a) métodos baseados no conhecimento e análise de dados

Tabela 2: Tabela com alguns aminoácidos comumente utilizados no método de varredura por alaninas e seus códons de DNA.

Aminoácido			Códon de DNA
Nome	Códigos		
Alanina	A	ALA	GCT
			GCC
			GCA
			GCG
Aspartato	D	ASP	GAC
			GAT
Glutamato	E	GLU	GAA
			GAG
Glicina	G	GLY	GGT
			GGC
			GGA
			GGG
Prolina	P	PRO	CCT
			CCC
			CCA
			CCG
Serina	S	SER	TCT
			TCC
			TCA
			TCG
			AGT
			AGC
Valina	V	VAL	GGT
			GTC
			GTA
			GTG
Treonina	T	THR	ACT
			ACC
			ACA
			ACG

experimentais (*knowledge-based*) e (b) métodos baseados em princípios físicos (*physics-based*).

Os métodos baseados no conhecimento englobam métodos físicos simplificados (Kortemme e Baker, 2002; Guerois et al., 2002; Kortemme e Baker, 2004) assim como métodos que utilizam diversos parâmetros ou descritores estruturais de complexos com estrutura resolvida, como por exemplo a superfície exposta ao solvente e a hidrofobicidade, em conjunto com métodos de aprendizado de máquina (*machine learning*) que após serem treinados com um conjunto de dados contendo dados experimentais do $\Delta\Delta G_{\text{ligação}}$ de resíduos e alguns descritores para o mesmo, usam as regras aprendidas para predizer os *hot spots* em outras estruturas de complexos proteicos (Gao et al., 2004; Darnell et al., 2007; Cho et al., 2009; Tuncbag et al., 2009).

Os métodos baseados em princípios físicos utilizam simulações no nível atômico para calcular o $\Delta\Delta G_{\text{ligação}}$ causado pela mutação de um resíduo por uma alanina. Entre os procedimentos utilizados estão: dinâmica molecular (Chong et al., 2006), *molecular mechanics/Poisson-Boltzmann surface area* (MM/PBSA) (Massova e Kollman, 1999), *molecular mechanics/generalized Born surface area* (MM/GBSA) (Moreira et al., 2007a) e mecânica quântica semiempírica (Diller et al., 2010). Esses métodos estão mais próximos de reproduzir os resultados experimentais, no entanto, seu alto custo computacional e o conhecimento exigido pelo usuário para utilizá-los impedem que ele seja aplicado em larga escala utilizando-se *desktops* e servidores padrão.

A seguir, veremos alguns métodos computacionais do tipo baseados em conhecimento que foram publicados na literatura com o objetivo de entender como cada um deles funciona, quais as vantagens e desvantagens de cada um deles e também qual o nível de precisão obtido por eles na identificação dos resíduos energeticamente mais importantes na interação proteína-proteína.

1.4.1 *Método computacional para predição de* hot spots *que utiliza apenas a sequência proteica*

1.4.1.1 *ISIS*

O programa ISIS[1] foi originalmente desenvolvido com a finalidade de prever, a partir da sequência de aminoácidos de uma proteína, quais os resíduos que compõem regiões de interação proteína-proteína (Ofran e Rost, 2007). Por utilizar a sequência de aminoácido de apenas uma das proteínas envolvidas na interação, o ISIS não utiliza nenhuma informação do complexo ao qual a interface pertence, nem mesmo qual ou quais parceiros interagem com um respectivo resíduos predito como constituinte da interface.

O método de predição utilizado pelo ISIS é composto de diversos descritores disponíveis para uma sequência proteica sendo eles:

1 Atualmente o programa está disponível em www.predictprotein.org com o nome PROFisis.

- Ambiente do resíduo - considera qual é o resíduo alvo, o qual será classificado como pertencente ou não a uma interface, juntamente com os oito vizinhos sequencialmente mais próximos, sendo quatro deles no sentido N-terminal e quatro no sentido C-terminal.
- Perfil evolutivo - considera o perfil evolutivo de nove resíduos selecionados do modo mencionado no item anterior.
- Acessibilidade ao solvente - acessibilidade ao solvente predita para o resíduo alvo e para seus dois vizinhos, sendo um de cada lado da sequência.
- Estrutura secundária - estrutura secundária predita do resíduo alvo e de seus dois vizinhos, um de cada lado.
- Score evolutivo - medida da conservação evolutiva do resíduo alvo.

Os descritores de cada resíduo alvo foram então utilizados como dados de entrada de uma rede neural que, após ser treinada, pode ser aplicada em sequências proteicas e assim gerar uma resposta indicando se o respectivo resíduo foi ou não predito como constituinte de uma interface. Para o treinamento, os autores utilizaram dois terços dos 59559 resíduos provenientes de 333 complexos proteicos inicialmente selecionados. O terço restante dos dados foi utilizado para teste, onde foram calculados a precisão (Equação 1) e a sensibilidade (Equação 2) do método.

$$\text{Precisão} = \frac{\text{número de verdadeiros positivos}}{\text{número de verdadeiros positivos} + \text{número de falsos positivos}} \quad (1)$$

$$\text{Sensibilidade} = \frac{\text{número de verdadeiros positivos}}{\text{número de verdadeiros positivos} + \text{número de falsos negativos}} \quad (2)$$

Quando o ISIS foi aplicado para predizer resíduos que compõem uma interface proteína-proteína, a precisão obtida foi de 90%. No entanto, com essa precisão o método identificou somente 5% do total de resíduos pertencentes a interfaces. A sensibilidade baixa, de apenas 5%, levou os autores à hipótese de que o método estaria identificando principalmente os resíduos essenciais a interação, os quais mais contribuem para a energia de ligação, em outras palavras, o método desenvolvido estaria identificando os *hot spots* das interações proteína-proteína (Ofran e Rost, 2007).

Para testar essa hipótese Ofran e Rost (2007) utilizaram o ISIS para predizer 296 resíduos com dados disponíveis de experimentos de varredura por alaninas pertencentes a 80 cadeias proteicas depositadas

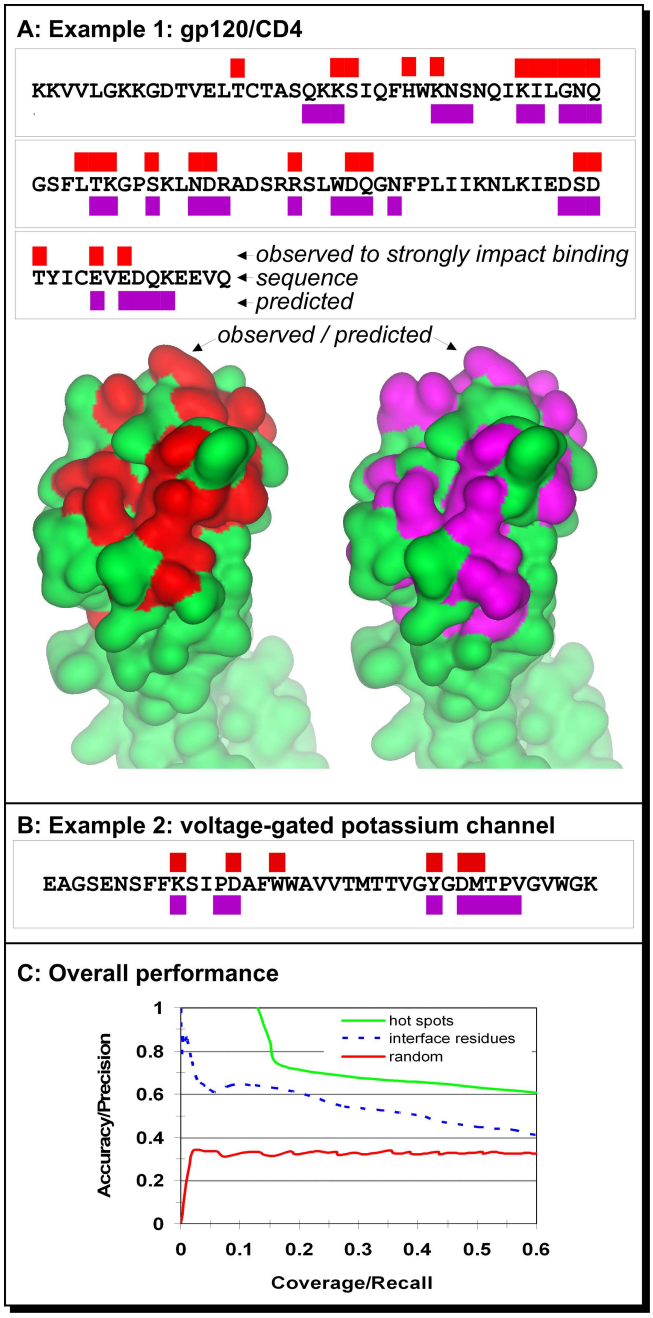


Figura 2: Figura extraída de Ofran e Rost (2007). (A) e (B) são exemplos de duas proteínas com os resíduos identificados experimentalmente como *hot spots* marcados com um retângulo vermelho sobre a sequência e os resíduos preditos pelo ISIS como *hot spots* marcados com um retângulo lilás sob a sequência. (C) Gráfico da precisão pela sensibilidade, mostrando a melhor performance do método na predição dos *hot spots* do que dos resíduos da interface proteína-proteína.

no ASEdb (Thorn e Bogan, 2001). A aplicação do ISIS para predizer *hot spots* demonstrou uma precisão de 70% e sensibilidade de 20% (Figura 2) e foram considerados hot spots resíduos cujo $\Delta\Delta G_{\text{ligação}}$ observado foi maior que 2,5 Kcal/mol quando substituído por alanina. É importante ressaltar o ISIS não foi treinado para predizer hot spots, mas para predizer resíduos que constituem uma interface proteína-proteína, e os autores não realizaram um novo treinamento para aplicá-lo a *hot spots*. Segundo eles, esse resultado obtido sugere que os hot spots possuem características mais marcantes que o restante dos resíduos da interface proteína-proteína (Ofran e Rost, 2007).

1.4.2 *Métodos computacionais para predição de* hot spots *que utilizam a estrutura do complexo*

1.4.2.1 *Rosetta Alanine Scanning*

Um dos métodos mais utilizados para a predição de *hot spots* é conhecido como Rosetta Alanine Scanning (Kortemme e Baker, 2002; Kortemme et al., 2004). Rosetta Alanine Scanning é um método construído a partir do programa de modelagem molecular computacional, Rosetta, que além da predição de *hot spots*, possui a capacidade de (1) modelar a estrutura terciária de proteínas, utilizando ou não proteínas homólogas; (2) modelagem de alças (*loops*) da estrutura terciária; (3) acoplamento (*docking*) proteína-proteína com a possibilidade de permitir uma flexibilidade parcial das mesmas; (4) acoplamento de pequenas moléculas com flexibilidade do ligante e parcial da proteína; (5) enovelamento de RNAs; (5) design de proteínas; (6) design de interfaces proteína-proteína e (7) design de interfaces proteína-DNA (Das e Baker, 2008; Kaufmann et al., 2010).

Rosetta Alanine Scanning realiza a predição de *hot spots* através do cálculo do $\Delta\Delta G_{\text{ligação}}$ para cada resíduo presente na interface. O cálculo é efetuado a partir do $\Delta G^{\text{MUT}}_{\text{ligação}}$, que corresponde a energia livre de ligação do complexo com um de seus resíduos da interface substituído por uma alanina, e do $\Delta G^{\text{WT}}_{\text{ligação}}$, correspondente a energia livre de ligação do complexo tipo selvagem (Equação 3).

$$\Delta\Delta G_{\text{ligação}} = \Delta G^{\text{MUT}}_{\text{ligação}} - \Delta G^{\text{WT}}_{\text{ligação}} \tag{3}$$

Onde o $\Delta G^{\text{MUT}}_{\text{ligação}}$ e o $\Delta G^{\text{WT}}_{\text{ligação}}$ são calculados a partir da energia livre do complexo e da energia livre de cada um dos parceiros do complexo isoladamente (Equação 4).

$$\begin{aligned}\Delta\Delta G_{\text{ligação}} \quad = \quad & (\Delta G^{\text{MUT}}_{\text{complexo}} - \Delta G^{\text{MUT}}_{\text{parceiro A}} - \Delta G^{\text{MUT}}_{\text{parceiro B}}) \\ & - (\Delta G^{\text{WT}}_{\text{complexo}} - \Delta G^{\text{WT}}_{\text{parceiro A}} - \Delta G^{\text{WT}}_{\text{parceiro B}})\end{aligned} \tag{4}$$

A energia livre é estimada pelo Rosetta Alanine Scanning através da seguinte equação linear:

$$\begin{aligned}\Delta G \;=\; & W_{\text{atração}} E_{\text{LJatração}} + W_{\text{repulsão}} E_{\text{LJrepulsão}} + W_{\text{HB(sc-bb)}} E_{\text{HB(sc-bb)}} \\ & + W_{\text{HB(sc-sc)}} E_{\text{HB(sc-sc)}} + W_{\text{Coul}} E_{\text{Coul}} + W_{\text{Sol}} G_{\text{Sol}} \\ & + W_{\Phi/\Psi} E_{\Phi/\Psi(\text{aa})} + \sum_{\text{aa}=1}^{20} n_{\text{aa}} E_{\text{aa}}^{\text{ref}} \end{aligned} \tag{5}$$

Onde $E_{\text{LJ atração}}$ e $E_{\text{LJ repulsão}}$ são respectivamente a parte atrativa e repulsiva do potencial de Lennard-Jones; $E_{\text{HB (sc-bb)}}$é o potencial de ligações de hidrogênio entre a cadeia lateral e a cadeia principal dos aminoácidos e $E_{\text{HB (sc-sc)}}$ é o potencial de ligação de hidrogênio entre as cadeias laterais dos aminoácidos, sendo ambos dependentes da orientação em que se encontram na estrutura. E_{Coul} é a energia eletrostática de Coulomb e G_{sol} corresponde a um modelo de solvatação implícito proposto por Lazaridis e Karplus(1999). O termo $E_{\phi/\psi}(aa)$ representa a propensão dos ângulos de torção da cadeia principal para um determinado tipo de aminoácido (aa) e o termo E_{aa}^{ref} é a energia de referência de um determinado tipo de aminoácido e representa as interações que esse aminoácido faz quando a proteína está no estado desenovelado.

As constantes W representam pesos aplicados a cada um dos diferentes termos de energia da equação, sendo que o peso (W) para ligações de hidrogênio apresenta três valores diferentes de acordo com o ambiente onde ocorre a ligação de hidrogênio: enterrado, parcialmente enterrado, ou exposto ao solvente. Os valores dos pesos foram obtidos através da parametrização da equação de forma a minimizar o erro entre a varição da energia livre calculada e a variação da energia livre observada em 743 mutações do tipo X $\rightarrow$ Ala, onde X significa qualquer aminoácido. Os valores experimentais para as variações da energia livre entre a proteína tipo selvagem e a proteína mutante, com uma mutação tipo X $\rightarrow$ Ala, foram obtidas do banco de dados ProTherm (Bava et al., 2004) sendo todos os dados utilizados, provenientes de proteínas monoméricas.

A parametrização da equação de variação da energia livre apresentou um coeficiente de correlação de 0,75 entre o valor predito e o observado e um erro médio absoluto de 0,81 Kcal/mol para as mutações X $\rightarrow$ Ala. Korteme e Baker (2002) também testaram a mesma equação, com pesos obtidos para mutações X $\rightarrow$ Ala, para prever a variação da energia livre entre a proteína tipo selvagem e proteína com mutação X $\rightarrow$ Y, onde Y significa qualquer aminoácido diferente de X. Nesse caso, o coeficiente de correlação entre os valores preditos e observados em 1584 mutações X $\rightarrow$ Y foi 0,70, fato que segundo os autores, pode indicar que o método poderia ser útil na classifica-

ção de mutações causadas por alguns tipos de polimorfismos de um único nucleotídeo (SNPs).

O teste do método na predição de hot spots foi realizado com a mesma equação de variação da energia livre parametrizada para mutações em proteínas monoméricas. Isto significa que os pesos aplicados aos termos energéticos da função não foram alterados. No entanto, os pesos para ligações de hidrogênio foram ajustados de forma que ligações de hidrogênio não expostas ao solvente e com ângulos ideais apresentassem um valor próximo ao estimado experimentalmente (-4,5 Kcal/mol). Esta modificação foi introduzida devido à grande importância das ligações de hidrogênio não expostas ao solvente nas interfaces proteína-proteína, e ao fato delas serem sub-representadas em proteínas monoméricas.

O método, esquematizado na figura 3, foi aplicado à 380 mutações, sendo 139 *hot spots* e 241 não *hot spots*, pertencentes à 19 complexos proteicos. Nesse trabalho, foram considerados *hot spots* resíduos que apresentam $\Delta\Delta G_{\text{ligação}}$ maior ou igual a 1,0 Kcal/mol quando substituídos por alanina. Rosetta conseguiu classificar corretamente 0,69 dos *hot spots* e 0,84 dos não *hot spots*, o que corresponde a uma precisão de 71% e sensibilidade de 69%. Os autores analisaram também a importância de cada termo da equação na predição de *hot spots* e notaram que os termos correspondentes as ligações de hidrogênio são os que mais influenciam na sensibilidade do método, enquanto o modelo implícito de solvatação tem a maior influência na precisão (Kortemme e Baker, 2002).

1.4.2.2 *FOLDEF*

O método de predição de *hot spots* usado pelo programa FOLDEF (Guerois et al., 2002) é baseado nos mesmos princípios utilizados pelo programa Rosetta Alanine Scanning e da mesma forma, os resíduos não são simplesmente classificados em *hot spots* ou não *hot spots*, mas ao invés disso, o valor da $\Delta\Delta G_{\text{ligação}}$ é calculado através das equações 3 e 4, as mesma utilizadas pelo Rosetta Alanine Scanning .

A variação da energia livre utilizada nas equações 3 e 4 é calculada através da função de energia FOLD-X, definida abaixo:

$$\begin{aligned}\Delta G \;=\; & W_{\text{vdw}}\Delta G_{\text{vdw}} + W_{\text{solvH}}\Delta G_{\text{solvH}} + W_{\text{solvP}}\Delta G_{\text{solvP}} \\ & + \Delta G_{\text{wb}} + \Delta G_{\text{hbond}} + \Delta G_{\text{el}} + W_{\text{mc}} T\Delta S_{\text{mc}} \\ & + W_{\text{sc}} T\Delta S_{sc} \end{aligned} \tag{6}$$

Onde ΔG_{vdw} é soma da contribuição de interações de van der Waals de todos os átomos da proteína. Os parâmetros ΔG_{solvH} e ΔG_{solvP} são, respectivamente, a variação da solvatação dos grupos hidrofóbicos e dos grupos polares quando mudam do estado desenovelado para o

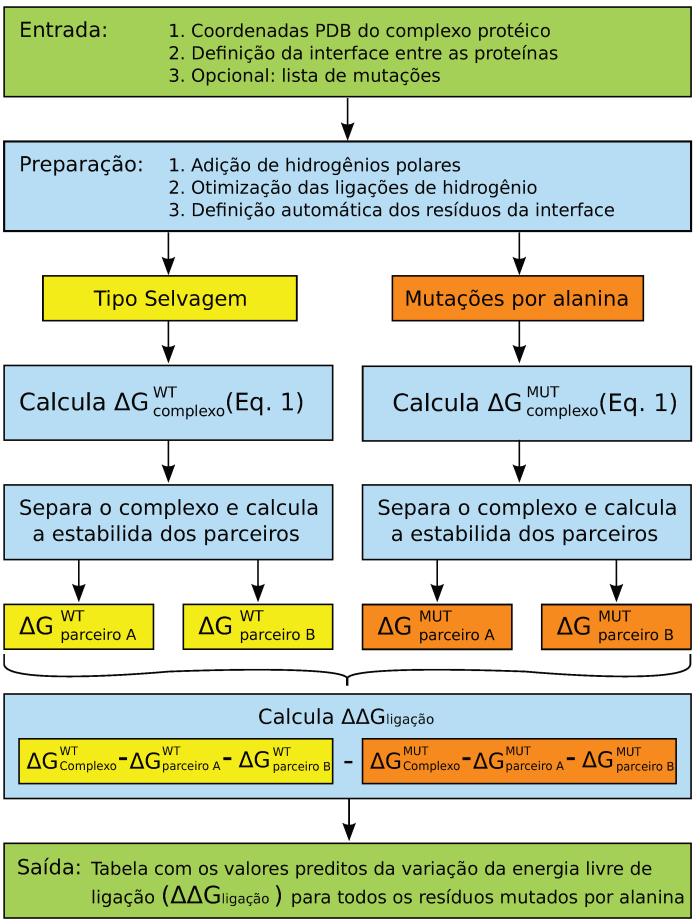


Figura 3: Diagrama do método Rosetta Alanine Scanning mostrando os passo executados para o cálculo da $\Delta\Delta G_{\text{ligação}}$.

estado enovelado. ΔG_{hbond} é a variação de energia livre entre uma ligação de hidrogênio intramolecular comparado com uma ligação de hidrogênio intermolecular, proteína-solvente. ΔG_{wb} corresponde a energia de moléculas de água que realizam mais de uma ligação de hidrogênio com a proteína. O termo ΔG_{el} representa a contribuição eletrostática dos grupos carregados. ΔS_{mc} e ΔS_{sc} representam o custo entrópico de manter fixa, respectivamente, a cadeia principal e a cadeia lateral.

Os pesos W, aplicados a alguns termos da função, foram obtidos ajustando a equação para representar, com o menor erro, o valor da variação de energia livre observado experimentalmente em 339 mutações únicas de proteínas monoméricas. A função parametrizada apresentou, para as mesmas 339 mutações, um coeficiente de correlação de 0,7 e um desvio padrão de 0,97 Kcal/mol. Quando aplicada a outras 625 mutações de proteínas monoméricas, não utilizadas no ajuste da equação, a função apresentou um coeficiente de correlação de 0,73 e um desvio padrão de 1,02 Kcal/mol.

Apesar das semelhanças entre o FOLDEF e o método do Rosetta, descrito anteriormente, o objetivo principal de ambos são, aparentemente, diferentes. O Rosetta foca principalmente na predição de *hot spots*, incluindo no artigo publicado os valores de *hot spots* e de não *hot spots* preditos corretamente, mas enfatiza também a possibilidade de ser utilizado na predição da variação da energia livre de mutações únicas em proteínas monoméricas. Já o FOLDEF, aparentemente tem como principal objetivo a predição da variação da energia livre em mutações únicas de proteínas monoméricas ao invés de *hot spots*. Este fato fica evidenciado pelo reduzido número de mutações analisadas que compõem regiões de interface proteína-proteína, que totalizam 82 extraídas de apenas quatro complexos. Além disso, nenhuma análise referente a precisão ou a sensibilidade do método na predição de *hot spots* foi realizada, sendo o coeficiente de correlação, igual a 0,64, e o desvio padrão, de 0,88 Kcal/mol, as principais informações sobre a performance do método quando aplicado a predição de *hot spots*.

1.4.2.3 *KFC*

O KFC (*Knowledge-based FADE/CON*) proposto por Darnell, Page e Mitchell (2007) apesar de utilizar informações estruturais, difere dos dois métodos anteriores por não utilizar uma equação que resultará em um valor estimado do $\Delta\Delta G_{\text{ligação}}$. Neste método, alguns atributos obtidos da estrutura do complexo como complementariedade geométrica da interface e tipos de contatos interproteicos são utilizados em conjunto com um método de aprendizado de máquina, neste caso árvores de decisão. Como consequência da metodologia utilizada, mais especificamente de atributos qualitativos e do uso de árvores de decisão, o KFC atua apenas como um classificador de resíduos da interface, classificando entre os tipos *hot spots* e não *hot spots*, não sendo possível predições quantitativas como a predição do valor do $\Delta\Delta G_{\text{ligação}}$ de um resíduo após ser substituído por uma alanina.

Os atributos utilizados pelo KFC são:

- Tamanho do resíduo - divididos em pequenos, médios e grandes.
- Natureza química do resíduo - separados em apolares, polares e carregados.
- Complementariedade geométrica da interface - a complementariedade geométrica da interface próxima a um resíduo específico é estimada através da densidade atômica na região do resíduo. A densidade atômica é calculada utilizando o programa FADE (*Fast Atomic Density Evaluation*) e são calculadas a densidade para dez esferas de raio variando de 1 até 10Å, centradas no centro de massa do resíduo. Resíduos localizados em regiões com maior complementariedade geométrica da interface

tendem a apresentar maior densidade atômica local (Darnell et al., 2007).

- Interações atômicas não covalentes - interações não covalentes são identificadas pelo programa WHAT IF (Vriend, 1990) e podem ser dos seguintes tipos: (1) contatos atômicos, que podem ser classificados em polares ou apolares; (2) ligações de hidrogênio, que recebem um pontuação de 0,0 à 1,0, onde 0,0 indica que não há uma ligação de hidrogênio e 1,0 representa um ligação de hidrogênio ideal; (3) pontes salinas.

Além do KFC, que utiliza todos os atributos acima, Darnell e colaboradores (2007) testaram um classificador apenas com atributos correspondentes à densidade atômica próxima a região do resíduo, denominado K-FADE, e outro classificador apenas com atributos referentes aos contatos, chamado K-CON. Todos os classificadores utilizaram árvores de decisão e foram comparados ao Rosetta Alanine Scanning (Tabela 3). A comparação dos métodos demonstrou que, apesar de K-FADE e K-CON terem um desempenho inferior ou próximo ao obtido pelo Rosetta alanine scanning segundo o parâmetro F1-score[2]. No entanto, a combinação de ambos os métodos utilizada pelo KFC apresentou uma performance um pouco superior ao Rosetta Alanine Scanning (ΔF1 $= +0,04$), ainda que não estatisticamente significante ($p\,valor = 0,22$).

Tabela 3: Comparação de desempenho entre os métodos K-FADE, K-CON, o método combinado KFC e o Rosetta Alanine Scanning.

Método	Precisão	Sensibilidade	F1 score	ΔF1	*p valor*
Rosetta	0,51	0,47	0,49	-	-
K-FADE	0,47	0,37	0,41	-0,08	0,88
K-CON	0,52	0,45	0,48	-0,01	0,54
KFC	0,49	0,58	0,53	+0,04	0,22

1.4.2.4 *Minerva*

Minerva (*MINE Residue VAlue*) é outro método de classificação de *hot spots* que utiliza algoritmos de aprendizado de máquina e estruturas de complexos proteicos. Nesse método, desenvolvido por Cho e colaboradores (2009) também foi usado o banco de dados ASEdb como fonte principal dos dados de treinamento e teste, além de diversos descritores estruturais dos resíduos como densidade atômica, hidrofobicidade e superfície exposta ao solvente. No entanto, uma das diferenças observadas nesse trabalho de Cho, Kim e Lee (2009) em relação ao trabalho de Darnell, Page e Mitchell(2007), que desenvolveram o KFC, é o emprego de uma etapa de seleção dos melhores descritores

2 F1-score $= \frac{2 \times \text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}$

com os objetivos de otimizar o classificador e também de ajudar a compreender quais os descritores mais importantes na predição dos *hot spots*.

Ao todo, esse trabalho contem 54 descritores sendo 53 derivados de dados estruturais e apenas um derivado da sequência proteica. A seguir estão listados os principais descritores.

- Densidade atômica - calculada utilizando uma esfera com raio de 5 Å centrada no centro de massa do resíduo;
- Densidade atômica ponderada - calculada da mesma forma que o descritor anterior, mas multiplicada por um fator de peso $F_w(i) = \frac{\Delta ASA_i}{\sum_{j=1} \Delta ASA_j}$, $j =$ resíduos da interface, onde ΔASA_i é a variação da superfície acessível ao solvente do resíduo entre os estados isolados e em complexo da proteína e ΔASA_j é a variação da superfície acessível ao solvente dos demais resíduos da interface entre os estados isolados e em complexo da proteína;
- Hidrofobicidade - calculada através do índice de hidrofobicidade dos resíduos proposto por Fauchere e Pliska (1983);
- Hidrofobicidade ponderada - calculada como o item anterior e multiplicada pelo fator de peso $F_w(i)$;
- Interações moleculares com resíduos da interface - analisa 18 tipos diferentes de interações;
- Interações moleculares no microambiente do resíduo - analisa 18 tipos diferentes de interações que ocorrem dentro de uma esfera com raio igual a 5 Å e centrada no centro de massa do aminoácido;
- Área superficial acessível ao solvente;
- Variação da área superficial acessível ao solvente - calculada como a diferença entre área acessível do resíduo quando a proteína está isolada e quando está em complexo;
- Conservação do resíduo - único descritor baseado em informações da sequência proteica.

A seleção dos melhores descritores foi realizada utilizando árvores de decisão e o método de máquinas de vetor de suportes (*SVM*) foi escolhido para o classificador. O resultado obtido com essa estratégia e esses descritores superou os métodos Rosetta Alanine Scanning , FOLDEF e KFC como mostra a tabela a seguir (Tabela 4):

Tabela 4: Comparação do desempenho entre o método Minerva e os métodos KFC, Rosetta Alanine Scanning e FOLDEF (Tabela adaptada de Cho e colaboradores (2009)).

Método	Precisão	Sensibilidade	Especificidade	F1-score
Minerva	0,73	0,58	0,89	0,65
KFC	0,58	0,55	0,85	0,56
Rosetta	0,62	0,49	0,90	0,55
FOLDEF	0,59	0,32	0,93	0,41

A seleção dos descritores, juntamente com uma análise estatística dos mesmos, possibilitou observar-se quais os descritores mais importantes para a tarefa de classificação entre *hot spots* ou não *hot spots*, ou os que mais diferem entre as duas classes. O resultado da análise demonstrou que a densidade atômica ponderada, a variação da área acessível ao solvente e a hidrofobicidade ponderada foram os descritores mais úteis na classificação dos *hot spots* (Cho et al., 2009).

2 DESENVOLVIMENTO

Esse trabalho foi desenvolvido em cinco etapas: (1) seleção dos conjuntos, um de dados para treinamento e validação, e outro para teste; (2) obtenção dos descritores dos resíduos, como por exemplo a acessibilidade ao solvente; (3) análise estatística dos descritores para observar as diferenças entre os resíduos que são *hot spots* e o que não são; (4) treinamento e validação dos classificadores juntamente com seleção dos melhores descritores, e por fim, (5) a aplicação dos classificadores no conjunto de teste para compará-los com outros classificadores da literatura. A seguir veremos os procedimentos executados em cada etapa.

2.1 SELEÇÃO DO CONJUNTO DE DADOS

2.1.1 *Dados para treinamento e validação*

Com o objetivo de criar um classificador de resíduos das interfaces proteína-proteína entre *hot spots* e não *hot spots*, inicialmente necessitamos de resultados experimentais, por exemplo os obtidos através de experimentos como o de varredura por alaninas, que identifiquem quais resíduos da interface são ou não são *hot spots*. Além dos dados que indicam quais resíduos da interface são *hot spots* ou não, como nosso trabalho utiliza descritores estruturais para a predição, é fundamental que a estrutura do complexo das mesmas proteínas utilizadas no experimento de varredura por alaninas tenham sido resolvidas previamente pelo método de cristalografia de proteínas por difração de raios-X.

A princípio, as estruturas das proteínas do complexo que fossem resolvidas individualmente poderiam ser acopladas por métodos computacionais para formar o complexo e então, poderíamos obter os descritores estruturais do complexo para serem utilizados na predição dos *hot spots*. No entanto, os métodos computacionais de acoplamento de proteínas ainda não são precisos o suficiente para garantir que o complexo resultante será igual ou próximo a estrutura real do complexo, considerando a estrutura real como sendo a resolvida por difração de raios-X. Tal abordagem poderia resultar em valores incorretos para os descritores estruturais dos resíduos da interface e por isso, o método de acoplamento computacional não foi utilizado para gerar estruturas do complexo com a finalidade de obtermos tais descritores. Por motivos semelhantes, também não utilizamos estruturas proteicas modeladas por homologia, uma vez que mesmo pequenos

erros na estrutura modelada poderiam resultar em valores incorretos dos descritores estruturais. Nenhum dos complexos estudados utilizando varredura por alaninas tiveram sua estrutura resolvida por ressonância magnética nuclear, o que evitou que tivéssemos que optar por utilizá-las ou não.

De acordo com os princípios descritos nos parágrafos anteriores, decidimos utilizar o banco de dados ASEdb (*Alanine Scanning Energetics database*) (Thorn e Bogan, 2001) que armazena resultados experimentais de varredura por alaninas disponíveis na literatura e o código de identificação da estrutura do complexo proteico estudado, quando o mesmo encontra-se depositado no banco de dados PDB (Berman et al., 2000). Por esse motivo, os complexos utilizados nesse trabalho são semelhantes aos utilizados em trabalhos publicados sobre predição de *hot spots* a partir de dados estruturais do complexo como em Kortemme e Baker (2002; 2004); Darnell, Page e Mitchell (2007) e Cho, Kim e Lee(2009). A tabela 5 contem uma descrição resumida dos complexos proteicos que compõem o conjunto de dados.

Os 14 complexos proteicos selecionados do ASEdb possuem um total de 283 resíduos analisados em experimentos de varreduras por alaninas, sendo que 53 desses resíduos são classificados como *hot spots* devido ao $\Delta\Delta G_{\text{ligação}} \geq 2,00$ Kcal/mol observado quando esses resíduos foram substituídos por alanina. Dos 230 classificados como não *hot spots*, 57 deles possuem $1,00 \leq \Delta\Delta G_{\text{ligação}} < 2,00$ Kcal/mol, 129 possuem $0,00 \leq \Delta\Delta G_{\text{ligação}} < 1,00$ Kcal/mol e 44 deles possuem $\Delta\Delta G_{\text{ligação}} < 0,00$ Kcal/mol o que indica que em alguns casos, a mutação dos respectivos resíduo na proteína original por uma alanina favoreceu energeticamente a formação do complexo. A figura 4 demonstra a distribuição dos resíduos de acordo com sua contribuição energética para a formação do complexo.

Os resíduos selecionados do ASEdb para comporem o conjunto de dados estão desigualmente distribuídos entre os complexos (Tabela 6). Além da desigualdade numérica de resíduos entre os complexos, há também diferenças na distribuição dos valores do $\Delta\Delta G_{\text{ligação}}$ dos resíduos de acordo com o complexo (Figura 5).

Ao comparamos o subconjunto de resíduos selecionados do ASEdb para compor os dados utilizados nesse trabalho com o conjunto total de dados disponíveis no ASEdb observamos que 18,66% dos resíduos do subconjunto selecionado são *hot spots* enquanto que no ASEdb são apenas 7,46% dos resíduos. Para comparar estatisticamente as duas distribuições utilizamos o teste de Kolmogorov-Smirnov. O teste apresentou um *p valor* de $6,79 \times 10^{-45}$ o que indica que há diferença entre as distribuições da $\Delta\Delta G_{\text{ligação}}$ dos resíduos do ASEdb e do subconjunto de resíduos selecionados para esse trabalho.

Ainda comparando o conjunto total de dados do ASEdb com o subconjunto utilizado nesse trabalho, notamos que a provável diferença entre os dois esteja na maior porcentagem de resíduos com $\Delta\Delta G_{\text{ligação}}$

Tabela 5: Tabela dos complexos proteicos que possuem dados de experimentos de varreduras por alaninas depositados no ASEdb assim como a estrutura do complexo resolvida e depositada no *Protein Data Bank*.

PDB ID	Proteína 1	Proteína 2	Resolução (Å)
1A4Y	Angiogenina	Inibidor de ribonuclease	2,00
1AHW	Imunoglobulina Fab 5G9	Fator tecidual	3,00
1BRS	Barnase	Barstar	2,00
1BXI	Proteína de imunidade à colicina E9	Domínio DNase da colicina E9	2,05
1CBW	Quimotripsina	BPTI	2,60
1DFJ	Inibidor de ribonuclease	Ribonuclease A	2,50
1DVF	Anticorpo D1.3	Anticorpo anti-idiotípico D1.3	1,90
1FC2	Fragmento Fc	Fragmento B da proteína A	2,80
1GC1	Proteína do envelope GP120	CD4	2,50
1JCK	Receptor de antígeno da célula T	Enterotoxina C3	3,50
1VFB	Anticorpo monoclonal de camundongo D1.3	Lisozima	1,80
2PTC	Tripsina	BPTI	1,90
3HFM	Anticorpo HY HEL-10	Lisozima	3,00
3HHR	Hormônio do crescimento humano	Receptor de hormônio do crescimento humano	2,80

no intervalo de 0,0 Kcal/mol à 1.0 Kcal/mol (Tabela 7). Essa diferença pode ser mais facilmente visualizada observando os histogramas de ambas as distribuições: subconjunto selecionado para ser utilizado neste trabalho (Figura 4) e do conjunto de dados do ASEdb (Figura 6).

Observamos também que além da diferença na distribuição relativa ao $\Delta\Delta G_{\text{ligação}}$ as diferenças nas frequências dos aminoácidos entre o subconjunto selecionado e o conjunto presente no ASEdb faz com que alguns aminoácidos, mais especificamente fenilalanina, metionina e prolina, não apresentem exemplos de *hot spots* no subconjunto selecionado (Tabela 8). Podemos observar também distribuições desiguais da $\Delta\Delta G_{\text{ligação}}$ para cada tipo de resíduo entre o subconjunto selecionado e o conjunto do ASEdb (Figura 7).

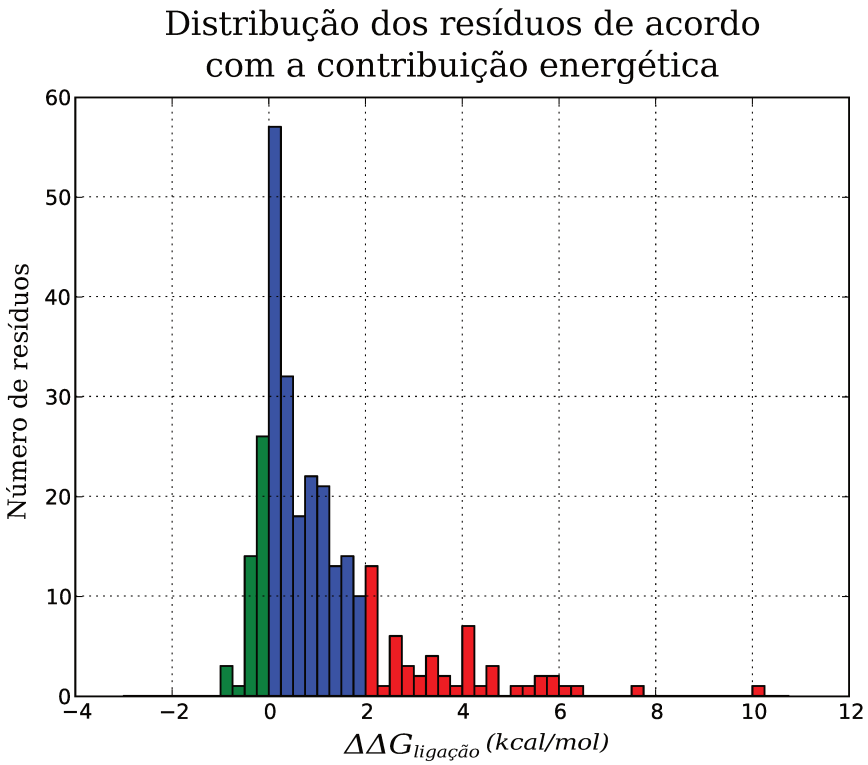


Figura 4: Distribuição dos resíduos selecionados do ASEdb para serem utilizados nesse trabalho em relação aos valores observados da $\Delta\Delta G_{ligação}$ em Kcal/mol.

2.1.2 *Dados para teste*

Assim como os dados para treinamento e validação, os dados para teste foram extraídos de um banco de dados que contém informação obtida experimentalmente da contribuição energética de resíduos substituídos por alaninas, e foram selecionados apenas resíduos pertencentes a complexos proteicos com estrutura resolvida. O banco de dados utilizado para montar o conjunto de dados de teste foi o BID (Fischer et al., 2003). Neste banco de dados os resíduos da interface substituídos por alanina são classificados em quatro categorias de acordo com a $\Delta\Delta G_{ligação}$ calculada: (1) forte (*strong*), que apresentou $\Delta\Delta G_{ligação}$ >= 2,0 Kcal/mol; (2) intermediário (*intermediate*), com 2,0 > $\Delta\Delta G_{ligação}$ >= 1,0 Kcal/mol; (3) fraco (*weak*), com 1,0 > $\Delta\Delta G_{ligação}$ >= 0,5 Kcal/mol; e insignificante (*insignificant*), que $\Delta\Delta G_{ligação}$ < 0,5 Kcal/mol quando substituídos por alanina.

No total, foram selecionados 105 resíduos de 13 complexos proteicos (Tabela 9), sendo 31 dos resíduos classificados como "fortes", 18 classificados como "intermediários", 21 como "fracos" e 35 como insignificantes. Apenas os resíduos classificados como "fortes" foram considerados *hot spots*, dessa forma, o conjunto de teste é composto de 31 *hot spots* e 74 não *hot spots*.

Tabela 6: Número de resíduos substituídos por alanina nos complexos proteicos, organizados separadamente de acordo com a proteína ligante a que pertencem e classificados entre *hot spots* e não *hot spots*.

PDB ID	Proteína 1		Proteína 2	
	Hot spots	Não hot spots	Hot spots	Não hot spots
1A4Y	1	13	2	12
1AHW	-	-	1	7
1BRS	6	2	3	3
1BXI	6	22	-	-
1CBW	-	-	1	8
1DFJ	4	10	-	-
1DVF	1/5[a]	5/5[a]	-/3[a]	1/5[a]
1FC2	-	-	1	2
1GC1	-	-	-	49
1JCK	3	16	4	6
1VFB	-/2[a]	7/8[a]	1	11
2PTC	-	-	1	-
3HFM	2	1	3	10
3HHR	3	28	-[b]	-[b]

a - Proteínas com duas cadeias distintas que interagem com a outra proteína ligante. Nesse caso, os resíduos substituídos estão separados também de acordo com a cadeia a que pertencem.

b - Proteína homodimérica, onde as duas cadeias interagem com regiões distintas da proteína ligante. Nesse caso, optamos por não utilizar esses resíduos.

Tabela 7: Número de resíduos em intervalos da $\Delta\Delta G_{\text{ligação}}$ no subconjunto de resíduos selecionados para treinamento e validação e no conjunto de resíduos do ASEdb.

Intervalo	Seleção (treinamento e validação)	ASEdb
$\Delta\Delta G_{\text{ligação}} \geq 2,00$ Kcal/mol	53	227
$1,00 \leq \Delta\Delta G_{\text{ligação}} < 2,00$ Kcal/mol	57	225
$0,00 \leq \Delta\Delta G_{\text{ligação}} < 1,00$ Kcal/mol	129	2341
$\Delta\Delta G_{\text{ligação}} < 0,00$ Kcal/mol	44	250
Total	283	3043

2.2 DESCRITORES ESTRUTURAIS

Descritores estruturais são dados obtidos através de análises realizadas nas estruturas proteicas. Nesse trabalho utilizamos 186 descritores, de 16 tipos, dos mais de 700 disponíveis no banco de dados STING_DB desenvolvido no Grupo de Pesquisa em Biologia Computacional da Embrapa (Neshich et al., 2006; Oliveira et al., 2007),

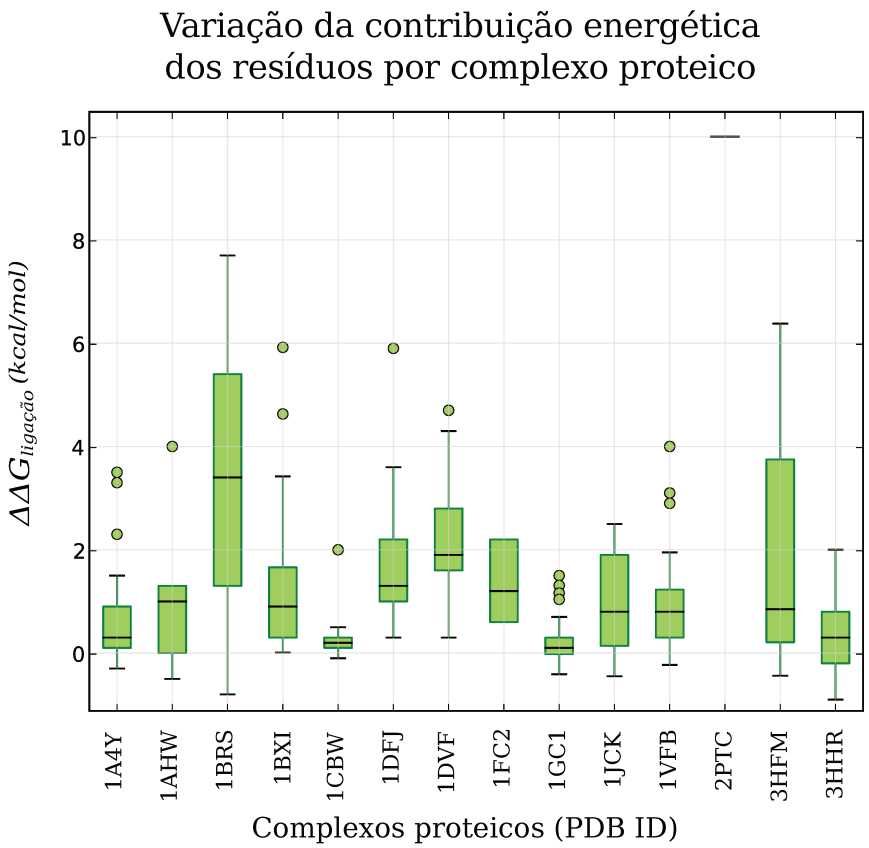


Figura 5: Distribuição dos valores da $\Delta\Delta G_{\text{ligação}}$ por complexo proteico.

além de 112 novos descritores, de 4 tipos, propostos nesse trabalho. A escolha de quais parâmetros do STING_DB seriam utilizados nesse trabalho foi feita de acordo com alguns critérios preestabelecidos: (1) apenas descritores numéricos são permitidos; (2) os descritores devem possuir um e somente um valor para cada resíduo da proteína; (3) não serão utilizados dados de conservação evolutiva.

O primeiro critério foi definido por ser um requisito do método de aprendizado de máquina utilizado nesse trabalho e como consequência excluiu descritores qualitativos como o de estrutura secundária. O segundo critério foi escolhido para evitar-se que alguns resíduos da proteína não possuíssem um valor definido para um determinado descritor ou ainda que o resíduo possuísse múltiplos valores. São exemplos de descritores excluídos pelo critério anterior os descritores *cross presence order* C_β e *cross link order* $C_\beta$que dependem da presença do C_β no resíduo o que impossibilita o cálculo para glicinas. O terceiro critério que exclui descritores referentes a conservação evolutiva foi utilizado devido ao objetivo do trabalho de analisar apenas os descritores estruturais e físico-químicos. Acreditamos que os descritores selecionados nesse trabalho possam representar as características determinantes para que um resíduo desempenhe a função de *hot spot*, enquanto que a conservação evolutiva apenas reflete a importância

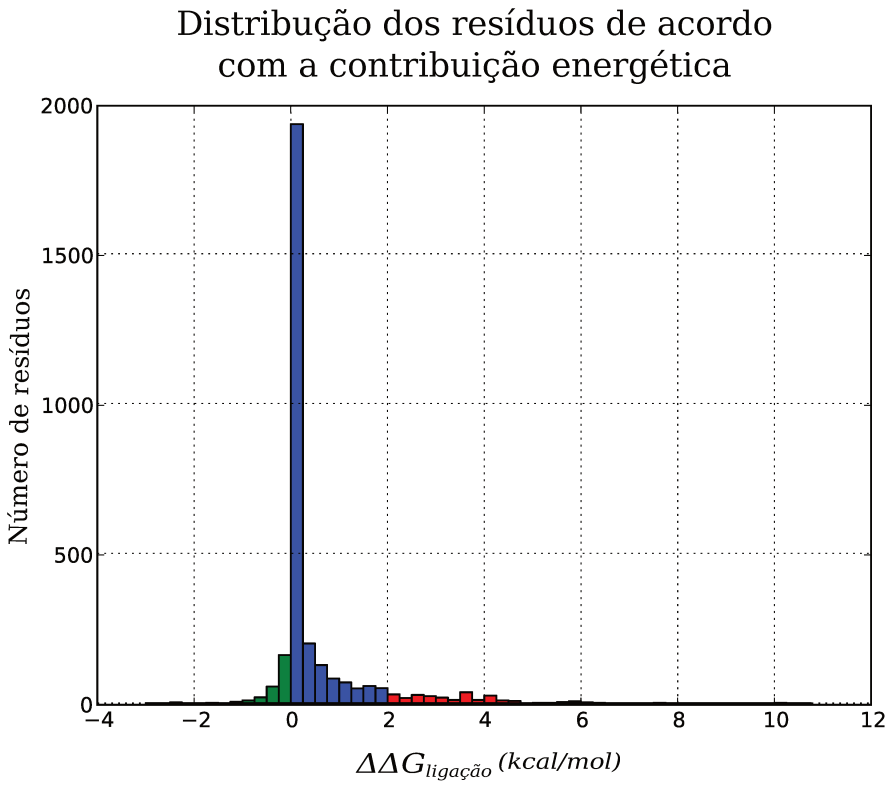


Figura 6: Distribuição dos resíduos do ASEdb em relação aos valores observados da $\Delta\Delta G_{\text{ligação}}$ em Kcal/mol.

do resíduo. Outro fator que contribuiu para a não utilização dos descritores de conservação evolutiva nesse trabalho foi a característica de nosso conjunto de dados, que apresenta um pequeno número de proteínas e a presença de diversos anticorpos os quais possuem grande variabilidade. No trabalho de Cho, Kim e Lee (2009), criadores do Minerva, os autores optaram por utilizar um descritor de conservação evolutiva, no entanto, após as análises estatísticas, eles observaram que esse descritor não apresentava diferenças significativas entre os resíduos que eram *hot spots* e os que não eram *hot spots*. Isso, segundo os próprios autores, ocorreu devido a presença de complexos com anticorpos nos dados utilizados por eles, os quais são muito semelhantes aos usados nesse trabalho. Para comprovar isso, eles fizeram uma análise estatística do descritor de conservação evolutiva retirando os complexos compostos por anticorpos e observaram que, nesse caso, há uma diferença significativa sendo os *hot spots* mais conservados que os não *hot spots*, esse fato está de acordo com outros trabalhos sobre conservação de *hot spots* (Hu et al., 2000; Guharoy e Chakrabarti, 2009).

Os 298 descritores selecionados para serem utilizados nesse trabalho são divididos nos 20 tipos descritos abaixo.

Tabela 8: Comparação do número de *hot spots* por resíduo entre os dados selecionados e o ASEdb.

Aminoácido	Seleção		ASEdb	
	Hot spots	Não hot spots	Hot spots	Não hot spots
C	-	1	-	4
D	9	18	19	206
E	3	28	13	271
F	-	3	14	204
G	2	3	5	39
H	1	11	5	84
I	2	6	13	114
K	6	22	23	183
L	1	6	5	267
M	-	1	3	76
N	3	21	16	146
P	-	5	8	98
Q	2	14	8	184
R	5	20	41	246
S	1	20	5	226
T	1	20	4	182
V	2	5	5	128
W	2	11	8	25
Y	13	16	32	133

2.2.1 *Acessibilidade ao solvente*

Este descritor, assim como os descritores abaixo podem ser obtidos para quaisquer proteínas através do Java Protein Dossier, *o qual permite exportar uma tabela com valores numéricos para todos eles.*

A área do resíduo acessível ao solvente é calculada utilizando o programa SurfV (Sridharan et al., 1992) e é armazenado no STING_DB em três descritores distintos.

- ACCC - Acessibilidade do resíduo ao solvente calculada com as proteínas unidas, formando o complexo;
- ACCI - Acessibilidade do resíduo ao solvente calculada para cada proteína isolada, isto é, sem estar em contato com outras proteínas do complexo;
- ACCR - Acessibilidade relativa do resíduo ao solvente que corresponde a $\frac{\text{ACCI}}{\text{ACC}_{\text{max}}}$, onde ACC_{max} corresponde a acessibilidade máxima do aminoácido.

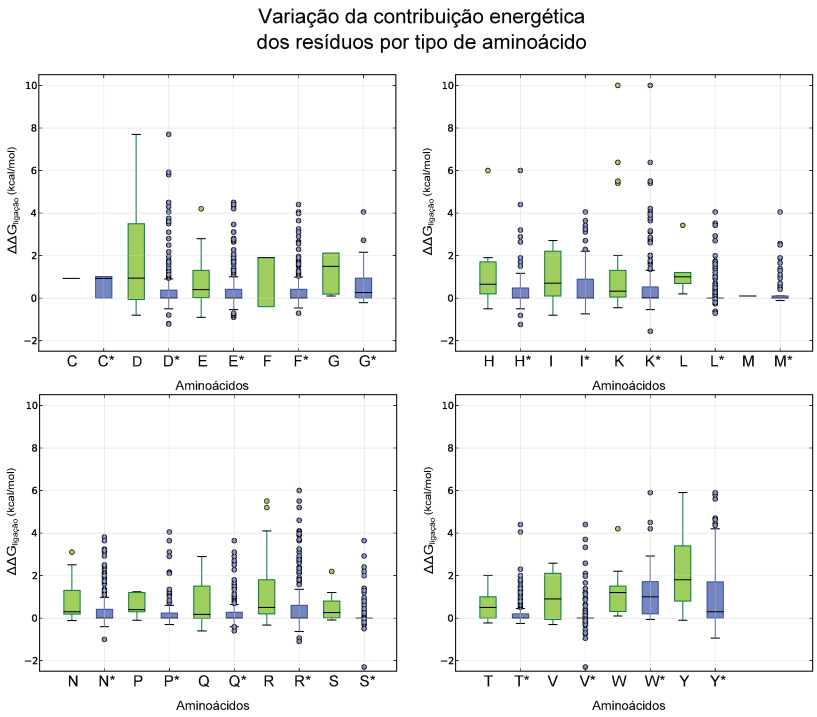


Figura 7: Comparação das distribuições da $\Delta\Delta G_{\text{ligação}}$ por aminoácido entre o conjunto de dados selecionados para treinamento e validação e os dados do ASEdb. Em verde são os dados do conjunto selecionado e em azul e marcado com * os dados do ASEdb.

2.2.2 *Cross link order*

O parâmetro *cross link order* indica o número de resíduos separados na sequência por pelo menos 15 resíduos, mas próximos estruturalmente, e que realizam interações hidrofóbicas, ligações de hidrogênio, ligações dissulfeto, interações eletrostáticas ou empilhamento aromático com o resíduo analisado. Este parâmetro está presente no STING_DB na forma de três descritores, onde somente são analisados resíduos que estejam dentro de uma esfera de raio igual a 3,5 Å.

- CloCA - Cross link order com esfera centrada no C_α;
- CloCB - Cross link order com esfera centrado no C_β e que não foi utilizado nesse trabalho;
- CloLHA - Cross link order com esfera centrado no último átomo da cadeia lateral com exceção do hidrogênio.

2.2.3 *Cross presence order*

Cross presence order indica o número de resíduos separados na sequência por pelo menos 15 resíduos, mas próximos estruturalmente e que diferentemente do parâmetro anterior, não realizam interações hidro-

Tabela 9: Tabela dos 13 complexos proteicos extraídos do BID do qual foram obtidos 105 resíduos para comporem o conjunto de dados de teste.

PDB ID	Proteína 1	Proteína 2	Resolução (Å)
1CDL	Calmodulina	CaMKII	2,00
1DVA	Fator de coagulação VIIA	Peptídeo E-76	3,00
1DX5	Trombina	Trombomodulina	2,30
1EBP	Receptor de EPO	Peptídeo mimético de EPO	2,80
1ES7	BMP-2	Receptor IA de BMP	2,90
1FAK	Fator de coagulação VIIA	Fator tecidual	2,10
1G3I	Protease HSLU	Protease HSLV	3,41
1IHB	P18-INK4C(INK6)	P18-INK4C(INK6)	1,95
1JAT	Ubc13	Mms2	1,60
1MQ8	ICAM-1	Integrina alpha-L	3,30
1NFI	NF-kappa-B	I-kappa-B-alpha	2,70
1NUN	FGF-10	FGFR-2	2,90
2HHB	Deoxihemoglobina	Deoxihemoglobina	1,74

fóbicas, ligações de hidrogênio, ligações dissulfeto, interações eletrostáticas ou empilhamento aromático com o resíduo analisado. Assim como o parâmetro anterior, este também está presente no STING_DB na forma de três descritores, onde somente são analisados resíduos que estejam dentro de uma esfera de raio igual a 3,5 Å.

- CpoCA - Cross presence order com esfera centrada no C_α;
- CpoCB - Cross presence order com esfera centrado no C_β, mas que não foi utilizado nesse trabalho;
- CpoLHA - Cross presence order com esfera centrado no último átomo da cadeia lateral com exceção do hidrogênio.

2.2.4 *Curvatura*

A curvatura do resíduo é calculada como a média da curvatura de cada um dos átomos do resíduo que estão na superfície, sendo calculada para as proteínas isoladas e em complexo. A curvatura des-

ses átomos é calculada utilizando-se o programa SurfRace[1] (Tsodikov et al., 2002). Valores negativos indicam uma superfície convexa e valores positivos uma superfície côncava.

- Curvature - Curvatura média do resíduo quando as proteínas estão em complexo;
- CurvatureIsolation - Curvatura média do resíduo quando as proteínas estão isoladas.

2.2.5 *Densidade*

A densidade é calculada como a soma da massa de todos os átomos da mesma cadeia presentes dentro de uma esfera de volume r, onde r pode ser 3, 4, 5 ou 6 Å, e dividida pelo volume da esfera. A esfera pode ser centrada tanto no $C_\alpha$quanto no último átomo da cadeia lateral com exceção do hidrogênio (LHA). A densidade também é calculada utilizando uma janela deslizante (*sliding window*) com n resíduos, onde n pode ser 3, 5, 7 ou 9 resíduos. Nesse caso, a densidade do resíduo é somada a densidade dos vizinhos ($\frac{n\text{-}1}{2}$) a direita e a esquerda e dividida por n. No total, há 40 descritores de densidade.

- DensityCA(r) - densidade calculada com esfera centrada no C_αe raio r, sendo r igual a 3, 4, 5 ou 6 Å;
- DensityCAsw(n, r) - densidade calculada com esfera centrada no C_α , raio r, onde r é igual a 3, 4, 5 ou 6 Å, e utilizando uma janela deslizante com n resíduos.
- DensityLHA(r) - densidade calculada com esfera centrada no LHA e raio r;
- DensityLHAsw(n, r) - densidade calculada com esfera centrada no LHA, raio r e janela deslizante.

2.2.6 *Densidade de energia*

A densidade de energia é a soma dos valores de energia para cada tipo de contato dentro de uma esfera de raio r, onde r pode ser 3, 4, 5 ou 6 Å, e dividida pelo volume da esfera. Os valores de energia para cada tipo de contato usados no cálculo estão na tabela 11.

1 Atualmente o programa foi renomeado Surface Racer e está disponível em http://apps.phar.umich.edu/tsodikovlab/index.htm

Tabela 10: Tabela com os valores energéticos para cada tipo de contato de acordo com o STING.

Tipo de Contato	Energia do contato (Kcal/mol)
van der Waals	0,08
Interação hidrofóbica	0,6
Empilhamento aromático	1,5
Ligação de hidrogênio	2,6
Ponte salina	10,0
Ponte dissulfeto	85,0

Nesse trabalho foram utilizados apenas os descritores de densidade de energia calculados usando uma janela deslizante com n resíduos, sendo n igual a 3, 5, 6 ou 9 resíduos. No total, usamos 32 descritores de densidade de energia, sendo 16 com esfera centrada no C_α e 16 centradas no último átomo da cadeia lateral não hidrogênio (LHA).

- EnergyDensityCAsw(n, r) - densidade de energia calculada com esfera centrada no C_α , raio r, onde r é igual a 3, 4, 5 ou 6 Å, e utilizando uma janela deslizante com n resíduos, sendo n igual a 3, 5, 7 ou 9 resíduos.
- EnergyDensityLHAsw(n, r) - densidade de energia calculada como no item acima, mas com esfera centrada no último átomo da cadeia lateral não hidrogênio .

2.2.7 *Potencial eletrostático*

O potencial eletrostático é calculado usando o programa DELPHI (Honig e Nicholls, 1995) com modificações feitas por Walter Rocchia e colaboradores(2002) e Neshich e colabores (2004). Utilizamos o potencial eletrostático calculado de seis modos.

- EPabsolute - é a soma do potencial eletrostático de todos os átomos do resíduo;
- EPaverage - é a média do potencial eletrostático de todos os átomos do resíduo;
- EPca - potencial eletrostático calculado no C_α;
- EPlha - potencial eletrostático calculado no último átomo da cadeia lateral não hidrogênio;
- EPsurface - potencial eletrostático calculado na superfície do resíduo.

2.2.8 *Hidrofobicidade*

A hidrofobicidade é calculada empregando dois índices de hidrofobidade dos resíduos, o proposto por Radzicka (Radzicka e Wolfenden, 1988) e o proposto por Kyte e Doolittle(Kyte e Doolittle, 1982), em ambos, quanto maior o valor para o resíduo, maior a hidrofobicidade. No total, cada resíduo possui seis descritores de hidrofobicidade sendo três deles calculados da mesma forma, mas com índices diferentes.

- HydroKD - valor do índice de hidrofobicidade do aminoácido proposto por Kyte e Doolittle;
- HydroKDC - calculado como $\text{HydroKDC} = \frac{\text{ACCC}}{\text{ACC}_{\text{max}}} \times \text{HydroKD}$, onde ACCC e ACC_{max} são calculados como no descritor 2.2.1;
- HydroKDI - calculado como $\text{HydroKDI} = \text{ACCR} \times \text{HydroKD}$, onde ACCR é calculado como descrito em 2.2.1;
- HydroR - valor do índice de hidrofobicidade do aminoácido proposto por Radzicka;
- HydroRC - calculado como $\text{HydroRC} = \frac{\text{ACCC}}{\text{ACC}_{\text{max}}} \times \text{HydroR}$, onde ACCC e ACC_{max} são calculados como no descritor 2.2.1;
- HydroRI - calculado como $\text{HydroRI} = \text{ACCR} \times \text{HydroR}$, onde ACCR é calculado como descrito em 2.2.1.

2.2.9 *Densidade na interface*

A densidade na interface é calculada como a soma da massa dos átomos dentro de uma esfera de raio r, e divido pelo volume da esfera. No entanto, diferentemente do descritor Densidade, esse descritor considera apenas átomos de outras cadeias ou proteínas. Nós utilizamos oito descritores de densidade na interface, são eles:

- IFRDensityCA(n) - densidades na interface calculadas com a esfera centrada no $C_\alpha$com raio r assumindo valores 3, 4, 5 e 6 Å;
- IFRDensityLHA(n) - densidades na interface calculadas com a esfera centrada no último átomo pesado da cadeia lateral não hidrogênio com raio r assumindo valores 3, 4, 5 e 6 Å.

2.2.10 *Esponjosidade na interface*

A esponjosidade na interface é calculada como a soma dos volumes de van der Waals dos átomos no interior de uma esfera de raio r, e dividido pelo volume da esfera. Assim como no descritor anterior são considerados apenas átomos de outras cadeias ou proteínas e

desde estejam dentro da esfera. Foram utilizados oito descritores de esponjosidade na interface:

- IFRSpongeCA(n) - esponjosidade na interface calculadas com a esfera centrada no $C_\alpha$com raio r assumindo valores 3, 4, 5 e 6 Å;
- IFRSpongeLHA(n) - esponjosidade na interface calculadas com a esfera centrada no último átomo pesado da cadeia lateral não hidrogênio com raio r assumindo valores 3, 4, 5 e 6 Å.

2.2.11 *Energia de contatos na interface*

A energia de contatos na interface é a soma dos valores de energia para cada tipo de contato que o resíduo faz com resíduos de outra cadeia ou de outra proteína. Ela é calculada de três formas:

O nome de todos os descritores do STING_DB foi mantido igual ao da tabela exportada pelo módulo Java Protein Dossier.

- InterfaceContactsEnergy(true,true) - soma da energia de todos os contatos do resíduo na interface, inclusive contatos com moléculas de água;
- InterfaceContactsEnergy(true,false) - soma da energia de todos os contatos do resíduo na interface, com exceção das moléculas de água;
- InterfaceContactsEnergy(false,true) - soma da energia apenas dos contatos distintos, isto é, caso um tipo de contato ocorra duas ou mais vezes com o mesmo resíduo, apenas um deles é somado. Também são considerados contatos com a água.

Os valores de energia para cada tipo de contatos são definidos pelo STING como na tabela 11.

Tabela 11: Valor da energia utilizado pelo STING para diferentes tipos de contatos.

Tipo de contato	Energia de contato do STING (Kcal/mol)
van der Walls	0,08
Interação hidrofóbica	0,6
Interação π	1,5
Ligação de hidrogênio	2,6
Interação eletrostática	10,0
Ponte dissulfeto	85,0

2.2.12 *Energia de contatos internos*

A energia de contatos internos é a soma dos valores de energia para cada tipo de contato (Tabela 11) que o resíduo faz com resíduos da mesma cadeia. Ela é calculada de três formas:

- InternalContactsEnergy(true,true) - soma da energia de todos os contatos internos do resíduo, inclusive contatos com moléculas de água;
- InternalContactsEnergy(true,false) - soma da energia de todos os contatos internos do resíduo, com exceção das moléculas de água;
- InternalContactsEnergy(false,true) - soma da energia apenas dos contatos distintos, isto é, caso um tipo de contato ocorra duas ou mais vezes com o mesmo resíduo, apenas um deles é somado. Também são considerados contatos com a água.

A energia de contados internos é calculada também usando uma janela deslizante com n resíduos, onde n pode ser 3, 5, 7 ou 9 resíduos.

- ContactsEnergyAllsw(true, n) - média da energia de todos os contatos do resíduo e $n-1$ vizinhos, considerando também contatos com moléculas de água;
- ContactsEnergyAllsw(false, n) - média da energia de todos os contatos do resíduo e $n-1$ vizinhos, não considerando contatos com moléculas de água;
- ContactsEnergyShortsw(true, n) - média da energia de contatos de curta distância do resíduo e $n-1$ vizinhos, considerando também contatos com moléculas de água;
- ContactsEnergyShortsw(false, n) - média da energia de contatos de curta distância, onde contatos de curta distância são, interações hidrofóbicas, ligações de hidrogênio ou interações π, do resíduo e $n-1$ vizinhos, com exceção de moléculas de água.

2.2.13 *Número de contatos na interface*

O número de contatos na interface é separado de acordo com o tipo de contato que o resíduo faz com resíduos de outras cadeias ou outras proteínas. Os tipos são:

- NumberofIFRContacts(1) - número de interações hidrofóbicas;
- NumberofIFRContacts(2) - número de interações eletrostáticas atrativas;
- NumberofIFRContacts(3) - número de interações eletrostáticas repulsivas;
- NumberofIFRContacts(4) - número de ligações de hidrogênio entre átomos da cadeia principal dos resíduos;

- NumberofIFRContacts(5) - número de ligações de hidrogênio entre um átomo da cadeia principal, uma molécula de água, e um átomo da cadeia principal de outro resíduo;
- NumberofIFRContacts(6) - número de ligações de hidrogênio entre um átomo da cadeia principal, duas moléculas de água, e um átomo da cadeia principal de outro resíduo;
- NumberofIFRContacts(7) - número de ligações de hidrogênio entre um átomo da cadeia principal e um da cadeia lateral;
- NumberofIFRContacts(8) - número de ligações de hidrogênio entre um átomo da cadeia principal, uma molécula de água, e um átomo da cadeia lateral;
- NumberofIFRContacts(9) - número de ligações de hidrogênio entre um átomo da cadeia principal, duas moléculas de água, e um átomo da cadeia lateral;
- NumberofIFRContacts(10) - número de ligações de hidrogênio entre átomos da cadeia lateral;
- NumberofIFRContacts(11) - número de ligações de hidrogênio entre um átomo da cadeia lateral, uma molécula de água, e outro átomo de uma cadeia lateral;
- NumberofIFRContacts(12) - número de ligações de hidrogênio entre um átomo da cadeia lateral, duas moléculas de água, e outro átomo de uma cadeia lateral;
- NumberofIFRContacts(13) - número de interações π;
- NumberofIFRContacts(14) - número de pontes dissulfeto.

2.2.14 *Número de contatos internos*

O número de contatos internos também é separado de acordo com o tipo de contato que o resíduo participa com outros resíduos da mesma cadeia. Os tipos de contatos são:

- NumberofINTContacts(1) - número de interações hidrofóbicas;
- NumberofINTContacts(2) - número de interações eletrostáticas atrativas;
- NumberofINTContacts(3) - número de interações eletrostáticas repulsivas;
- NumberofINTContacts(4) - número de ligações de hidrogênio entre átomos da cadeia principal dos resíduos;

- NumberofINTContacts(5) - número de ligações de hidrogênio entre um átomo da cadeia principal, uma molécula de água, e um átomo da cadeia principal de outro resíduo;
- NumberofINTContacts(6) - número de ligações de hidrogênio entre um átomo da cadeia principal, duas moléculas de água, e um átomo da cadeia principal de outro resíduo;
- NumberofINTContacts(7) - número de ligações de hidrogênio entre um átomo da cadeia principal e um da cadeia lateral;
- NumberofINTContacts(8) - número de ligações de hidrogênio entre um átomo da cadeia principal, uma molécula de água, e um átomo da cadeia lateral;
- NumberofINTContacts(9) - número de ligações de hidrogênio entre um átomo da cadeia principal, duas moléculas de água, e um átomo de uma cadeia lateral;
- NumberofINTContacts(10) - número de ligações de hidrogênio entre átomos da cadeia lateral;
- NumberofINTContacts(11) - número de ligações de hidrogênio entre um átomo da cadeia lateral, uma molécula de água, e outro átomo de uma cadeia lateral;
- NumberofINTContacts(12) - número de ligações de hidrogênio entre um átomo da cadeia lateral, duas moléculas de água, e outro átomo de uma cadeia lateral;
- NumberofINTContacts(13) - número de interações π;
- NumberofINTContacts(14) - número de pontes dissulfeto.

2.2.15 *Número de resíduos em contato na interface*

Número de resíduos que estão em outra cadeia ou proteína e que participam de um determinado tipo de interação com o resíduo sendo analisado. Esse descritor também é separado de acordo com os tipos de interações.

- NumberofIFResidues(1) - número de resíduos que estão participando de interações hidrofóbicas com o resíduo analisado;
- NumberofIFResidues(2) - número de resíduos estão realizando interações eletrostáticas atrativas com o resíduo analisado;
- NumberofIFResidues(3) - número de resíduos participando de interações eletrostáticas repulsivas;

- NumberofIFResidues(4) - número de resíduos participando de ligações de hidrogênio envolvendo seus átomos da cadeia principal e a cadeia principal do resíduo analisado;
- NumberofIFResidues(5) - número de resíduos que realizam ligações de hidrogênio do tipo cadeia principal, uma molécula de água, cadeia principal;
- NumberofIFResidues(6) - número de resíduos que realizam ligações de hidrogênio do tipo cadeia principal, duas moléculas de água, cadeia principal;
- NumberofIFResidues(7) - número de resíduos participando de ligações de hidrogênio envolvendo a cadeia principal de um resíduo com a cadeia lateral de outro resíduo;
- NumberofIFResidues(8) - número de resíduos participando de ligações de hidrogênio envolvendo um átomo de cadeia principal e um átomo da cadeia lateral, com uma molécula de água entre eles;
- NumberofIFResidues(9) - número de resíduos participando de ligações de hidrogênio envolvendo um átomo de cadeia principal e um átomo da cadeia lateral, com duas moléculas de água entre eles;
- NumberofIFResidues(10) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias laterais;
- NumberofIFResidues(11) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias laterais, mas com uma molécula de água entre eles intermediando a ligação;
- NumberofIFResidues(12) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias laterais, mas com duas molécula de água entre eles intermediando a ligação;
- NumberofIFResidues(13) - número de resíduos participando de interações π;
- NumberofIFResidues(14) - número de resíduos envolvidos em pontes dissulfeto.

2.2.16 *Número de resíduos internos em contato*

Número de resíduos da mesma cadeia que participam de um determinado tipo de interação com o resíduo sendo analisado. Esse descritor também é separado de acordo com os tipos de interações.

- NumberofINTResidues(1) - número de resíduos realizando interações hidrofóbicas;

- NumberofINTResidues(2) - número de resíduos participando de interações eletrostáticas atrativas;
- NumberofINTResidues(3) - número de resíduos participando de interações eletrostáticas repulsivas;
- NumberofINTResidues(4) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias principais;
- NumberofINTResidues(5) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias principais, mas com uma molécula de água entre eles intermediando a ligação;
- NumberofINTResidues(6) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias principais, mas com duas molécula de água entre eles intermediando a ligação;
- NumberofINTResidues(7) - número de resíduos participando de ligações de hidrogênio envolvendo a cadeia principal de um resíduo com a cadeia lateral de outro resíduo;
- NumberofINTResidues(8) - número de resíduos participando de ligações de hidrogênio envolvendo um átomo de cadeia principal e um átomo da cadeia lateral, com uma molécula de água entre eles;
- NumberofINTResidues(9) - número de resíduos participando de ligações de hidrogênio envolvendo um átomo de cadeia principal e um átomo da cadeia lateral, com duas moléculas de água entre eles;
- NumberofINTResidues(10) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias principais;
- NumberofINTResidues(11) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias principais, mas com uma molécula de água entre eles intermediando a ligação;
- NumberofINTResidues(12) - número de resíduos envolvidos em ligações de hidrogênio entre átomos das cadeias laterais, mas com duas molécula de água entre eles intermediando a ligação;
- NumberofINTResidues(13) - número de resíduos participando de interações π;
- NumberofINTResidues(14) - número de resíduos envolvidos em pontes dissulfeto.

2.2.17 *Número de átomos acessíveis ao solvente em região definida*

Este descritor e também os três próximos foram propostos neste trabalho e não fazem parte dos descritores do STING_DB.

O número de átomos acessíveis ao solvente em região definida é a contagem dos átomos localizados na superfície proteica separados por no máximo uma distância d de pelo menos um átomo do resíduo selecionado. Neste descritor, as proteínas são analisadas somente em isolamento, isto é, sem estarem ligadas à outras formando um complexo. Outra característica desse descritor é que somente são considerados átomos que formam uma superfície contínua a partir do resíduo analisado.

- NumAtomDist(d) - número de átomos há uma distância d, onde d pode ser 3, 4, 5, 6, 7, 8 ou 9 Å, de qualquer átomo do resíduos (são incluídos os átomos do resíduo);

Este descritor também calcula o número de átomos em coroas circulares ao redor do resíduo. As coroas circulares tem um raio maior R e um raio menor r. A contagem nesse caso é feita da seguinte forma:

- NumAtomCCirc(R, r) = NumAtomDist(R) − NumAtomDist(r), onde R assume os valores de 4, 5, 6, 7, 8 e 9 Å e r os valores de 3, 4, 5, 6, 7 e 8 Å.

Este, assim como os demais tipos de descritores propostos nesse trabalho, apresentam variações que descrevem especificamente a região ao redor do resíduo analisado. Estas variações foram propostas porque acreditamos que essas a região ao redor do resíduo, que assemelha-se a uma coroa circular, pode fornecer informações do anel-O, e devida a sua importância funcional, poderia apresentar diferenças entre resíduos que são *hot spots* e resíduos que não são e consequentemente auxiliar na predição dos *hot spots*. O fato de não haver uma definição precisa de quais resíduos ou da região específica que compõem o anel-O nos levou a utilizar diversos valores de distância (d) e raios (R, r). Nossa escolha de utilizar átomos da região ao redor dos *hot spots* assemelha-se a definição utilizada por Li e Li (2010) onde são considerados os resíduos imediatamente vizinhos dos *hot spots* como constituintes do anel-O. A figura 8 demonstram algumas dessas regiões analisadas pelo descritor.

2.2.18 *Área acessível ao solvente em região definida*

A área acessível ao solvente em região definida é a somatória das áreas atômicas acessíveis ao solvente calculada pelo SurfV (Sridharan et al., 1992) dos átomos selecionados no descritor anterior. A área é calculada tanto para a região circular quanto para a coroa circular.

- ÁreaAccSolv(d) - somatória das áreas acessíveis dos átomos a uma distância máxima d de qualquer átomo do resíduo;

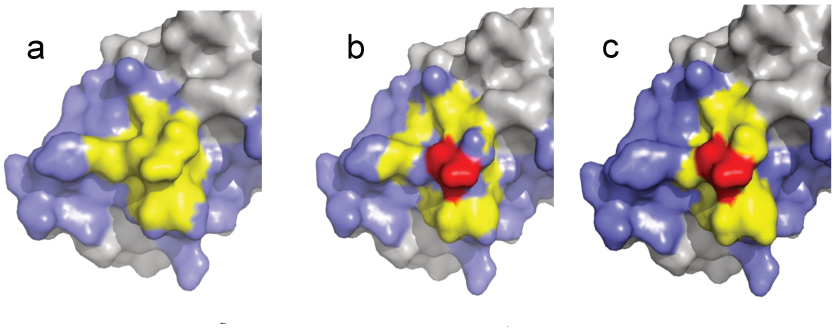


Figura 8: Estes são exemplos de regiões analisadas pelos descritores propostos neste trabalho. Na figura (a) a região em azul é a área da interface do complexo e em amarelo, os átomos distantes até 6 Å do resíduo analisado, incluindo os átomos do resíduo, neste caso um *hot spot*, e parte da região ao redor representando uma área do anel-O . Na figura (b), a região em amarelo são os átomos da coroa circular, que engloba parte do anel-O, distantes mais de 4 Å e menos de 8 Å dos átomos do resíduo analisado. Em vermelho estão os átomos do resíduo analisado, neste caso um *hot spot*. Na figura (c) podemos observar em vermelho o *hot spot* e em amarelo os resíduos que são vizinhos imediatos do *hot spot* e que por esse motivo, são considerados como constituintes do anel-O segundo Li e Li (2010).

- ÁreaAccSolv(R, r) - somatória das áreas acessíveis dos átomos na coroa circular de raio maior R e raio menor r.

2.2.19 *Entalpia de hidratação em região definida*

Uma das hipóteses mais aceitas como sendo uma das características necessárias para um resíduo ser um *hot spot* é a presença do anel-O. Esse anel seria constituído de resíduos mais hidrofóbicos que teriam a função de impedir o acesso do solvente aos hot spots. Para testar essa hipótese e adicionar descritores que possam fornecer essa informação, nós calculamos a entalpia de hidratação dos resíduos da superfície proteica como uma forma de medir a hidrofobicidade, permitindo identificar regiões onde moléculas de água seriam mais facilmente deslocadas. A entalpia para cada átomo exposto a superfície foi calculada utilizando-se valores medidos experimentalmente no trabalho de (Makhatadze e Privalov, 1993) para proteínas a 25 °C (Tabela 12) e utilizando a fórmula abaixo:

$$H_{\text{hidratação}} = ASA \times \sigma \tag{7}$$

Onde $H_{\text{hidratação}}$ é a entalpia de hidratação, ASA é a área do átomo exposto ao solvente, e σ é o valor calculado por (Makhatadze e Privalov, 1993) para o átomo.

Tabela 12: Parâmetros de entalpia de hidratação calculados por Makhatadze e Privalov (1993).

Superfície	σ (J.mol^{-1}.Å^{-1})
Alifática	-122
Aromática	-148
Parte polar:	
Arg	-827
Asn	-894
Asp	-715
Cys	-271
Gln	-703
Glu	-562
His	-1128
Lys	-714
Met	-473
Ser	-1045
Thr	-1287
Trp	-1161
Tyr	-854
-CONH-	-1702

A entalpia de hidratação numa região definida é a somatória da entalpia de hidratação dos átomos selecionados como nos dois descritores anteriores, ou seja, calculada para a região circular ao redor do resíduo e para a coroa circular.

- EntalpiaHidratação(d) - somatória da entalpia de hidratação dos átomos expostos ao solvente e separados no máximo por uma distância d de pelo menos um átomo do resíduo analisado;
- EntalpiaHidratação(R, r) - somatória da entalpia de hidratação dos átomos expostos ao solvente e na região da coroa circular de raio maior R e raio menor r.

2.2.20 *Entalpia de hidratação por área em região definida*

A entalpia de hidratação por área é uma modificação do descritor anterior para corrigir a influência da área exposta ao solvente na entalpia de hidratação. Este descritor é calculado das seguintes formas:

- $\text{EHporÁrea}(d) = \frac{\text{EntalpiaHidratação}(d)}{\text{ÁreaAccSolv}(d)}$;
- $\text{EHporÁrea}(R, r) = \frac{\text{EntalpiaHidratação}(R,r)}{\text{ÁreaAccSolv}(R,r)}$.

2.3 ANÁLISE ESTATÍSTICA DOS DESCRITORES

A análise estatística dos descritores foi realizada com o objetivo de identificar quais descritores diferem entre os subgrupos *hot spot* e não *hot spot*. Definido esse objetivo, optamos por comparar os valores médios apresentados pelos dois subgrupos para cada um dos descritores. Essa análise possibilita observar quais descritores diferem entre os dois subgrupos e assim inferir as possíveis causas e a importância de tal diferença como requisito para que um resíduo desempenhe o papel de *hot spot* na interação proteína-proteína.

O teste selecionado para esta comparação entre as médias foi o teste t para variâncias desiguais, também conhecido com teste t de Welch. Diferentemente do popular teste t de Student, que assume igualdade de variância entre os dois grupos comparados, o teste de Welch não assume tal igualdade e por isso, possibilita sua utilização tanto em casos onde a variância é igual como em casos onde ela seja desigual. Isso torna desnecessário qualquer teste preliminar para comparação das variâncias, permitindo que ele seja utilizado diretamente para a comparação das médias entre duas amostras. Diferentes trabalhos foram publicados comparando o teste t de Student e o teste t de Welch e indicam que há um consenso a favor da utilização do teste t de Welch para comparar a média de duas amostras independentemente da igualdade ou desigualdade de variância entre as amostras (Ruxton, 2006).

O teste t de Welch é calculado pela seguinte fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8}$$

onde $\bar{x}_1$ e $\bar{x}_2$ são as médias dos descritores, s_1^2 e s_2^2 são as variâncias e n_1 e n_2 correspondem ao número de elementos em cada uma das amostras. A variância das amostras é calculada pela fórmula:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{9}$$

O teste t de Welch foi calculado utilizando a função `welch_test` presente no programa OCTAVE[2] e retorna o *p valor* calculado indicando a significância estatística do teste, isto é, a probabilidade do teste apresentar o mesmo resultado caso a média entre a amostras sejam iguais (hipótese nula). Isso significa que quanto menor o *p valor*, mais significante estatisticamente é a diferença do valor médio entre as amostras.

2 Programa de código livre disponível em www.octave.org

2.4 MÁQUINAS DE VETORES DE SUPORTE

A máquina de vetor de suporte, abreviadamente SVM, é um algoritmo muito utilizado atualmente para problemas de classificação e de regressão. Este algoritmo, proposto inicialmente por Boser, Guyon e Vapnik(1992), foi o método de aprendizado de máquina utilizado nesse trabalho. Esse método faz parte dos métodos de aprendizado supervisionado, o qual exige que um conjunto de dados seja fornecido inicialmente para que o método de aprendizado de máquina, ou especificamente nesse trabalho a SVM, aprenda como classificar corretamente para então aplicar a novos dados. Esse conjunto de dados utilizado inicialmente para que o método aprenda a classificar é chamado de conjunto de treinamento e baseando-se nele, a SVM produzirá um modelo e o aplicará no conjunto de teste para medir o desempenho do modelo (Hsu et al., 2003).

As máquinas de vetores de suporte são normalmente utilizadas como classificadores de duas classes, onde o objetivo é distinguir com eficácia a qual das duas classes existentes cada elemento da amostra pertence. Essa classificação é, no caso das SVMs, realizada definindo-se um limite ou fronteira, chamado de hiperplano, entre as classes (Figura 9) de modo a separá-las da melhor forma possível, isto é, com a maior margem de separação entre as classes. As SVMs podem ser aplicadas tanto em dados linearmente separáveis quanto em dados não linearmente separáveis. No caso de dados não linearmente separáveis, utiliza-se funções de kernel que transformarão os dados de entrada aumentando as dimensões, tornando-os linearmente separáveis e possibilitando a busca de um hiperplano de separação das classes (Han et al., 2011).

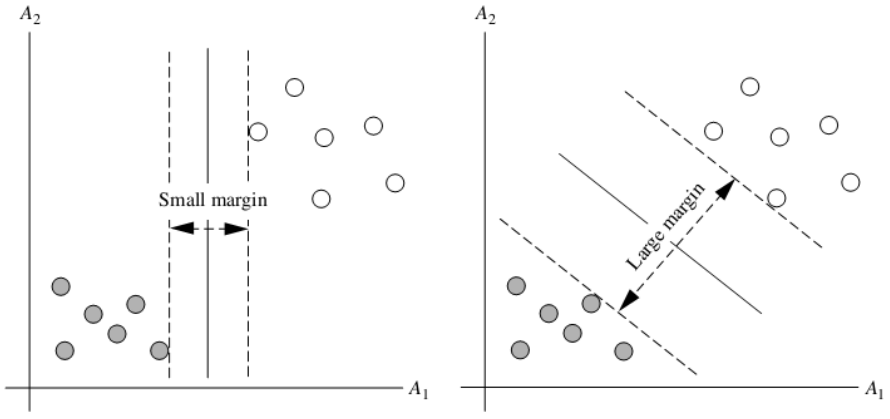


Figura 9: Dois possíveis hiperplanos de separação das classes. A SVM procurará o melhor deles, no caso, o com maior margem de separação (Figura extraída de Han et al. (2011)).

As máquinas de vetores de suporte estão disponíveis para uso através de diversos programas como por exemplo LIBSVM (Chang e Lin,

2001), mySVM (**?**), SVM^{light}(Joachims, 1999), Weka e RAPIDMINER. Nesse trabalho nós escolhemos o programa RAPIDMINER, por permitir executar diversas etapas do processo como normalização dos dados, validação cruzada, cálculo das medidas de desempenho e seleção dos descritores. As SVMs utilizadas no RAPIDMINER foram importadas da biblioteca LIBSVM(Chang e Lin, 2001). Foram usados dois tipos de SVMs, as lineares e as com função de kernel RBF para dados não linearmente separáveis. Está escolha por dois tipos de SVMs ocorreu devido ao número variável de descritores que utilizaremos. Inicialmente, teremos um número grande de descritores, semelhante ao número de resíduos selecionados. Em casos como esse, onde os números de descritores e o de resíduos no conjunto são parecidos, recomenda-se a utilização de SVMs lineares (Hsu et al., 2003). No entanto, a medida que os descritores forem selecionados, seu número tende a diminuir, e em casos onde o número de descritores seja muito menor que o número de resíduos no conjunto de dados é recomendado o uso de SVMs em conjunto com funções de kernel (Hsu et al., 2003).

2.5 ANÁLISE DOS CLASSIFICADORES

A análise dos classificadores tem o objetivo de calcular o desempenho do classificador na tarefa de distinguir corretamente os resíduos da interface entre *hot spots* e não *hot spots*. Além disso, a análise é essencial para observarmos quais são os melhores descritores e para comparar os resultados obtidos neste trabalho com outros da literatura.

2.5.1 *Treinamento e validação*

Antes do treinamento, os dados foram organizados em uma tabela, onde as linhas correspondiam aos resíduos e as colunas correspondiam aos descritores desses resíduos. Cada um dos descritores foram normalizados aplicando a seguinte fórmula:

$$X^{'} = \frac{X - \overline{X}}{s} \tag{10}$$

onde $X^{'}$representa o novo valor do descritor para o resíduo, X é o valor original, $\overline{X}$ é o valor médio do descritor para os resíduos que serão utilizados nesse trabalho e s é o desvio padrão do descritor. Está normalização não é obrigatória, mas recomendada para facilitar o treinamento de alguns métodos de aprendizado de máquina, incluindo máquinas de vetores de suporte (Chang e Lin, 2001).

O método empregado para estimar a performance do classificador durante a seleção dos melhores descritores foi a validação cruzada.

Esse método foi escolhido devido ao número reduzido de resíduos em nosso conjunto de dados. Na validação cruzada, o conjunto dos dados são divididos em subconjuntos, nesse caso, 10 subconjuntos, com um número de resíduos similar e também proporções semelhantes de *hot spots* e não *hot spots.* A divisão em 10 subconjuntos com amostragem estratificada como a utilizada nesse trabalho é o método de validação cruzada mais comumente empregado (Witten e Frank, 2005). Após a divisão, o classificador é treinado 10 vezes, cada uma delas utilizando 9 subconjuntos para o treinamento e 1 subconjunto para validação, o qual é usado para calcular o desempenho do classificador de acordo com os descritores selecionados. Os valores finais das medidas de performance da validação cruzada são obtidos através da média observada para as 10 vezes que o classificador é treinado e validado.

2.5.2 *Medidas de performance*

Diversas medidas de desempenho são usadas para classificados binários como o empregado neste trabalho. A seguir, descreveremos as medidas que utilizamos.

2.5.2.1 *Acurácia*

Acurácia é uma medida da proporção de resultados corretos na predição dos resíduos. Ela é calculada como:

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP+FP+TN+FN}} \tag{11}$$

onde TP e TN são respectivamente denominados verdadeiros positivos e verdadeiros negativos, ou seja, é o número de resíduos preditos corretamente. Os outros dois parâmetros da equação, FP e FN, são os falsos positivos e falsos negativos e representam o número de resíduos preditos incorretamente.

2.5.2.2 *Precisão*

A precisão, diferentemente da acurácia, é uma medida da proporção de resíduos preditos corretamente, mas apenas entre os classificados como *hot spots.*

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

2.5.2.3 *Sensibilidade (Recall)*

A sensibilidade ou *recall* indica a capacidade do classificador de identificar os casos positivos, isto indica que quanto maior a sensibilidade, maior o número de *hot spots* que o método classifica corretamente.

$$\text{Sensibilidade} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

2.5.2.4 *Especificidade*

A especifidade indica a capacidade do classificador de identificar corretamente o casos negativos.

$$\text{Especificidade} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

2.5.2.5 *F-score*

O F-score é uma medida composta por duas outras medidas, a precisão e a sensibilidade. O uso dessas duas medidas o tornam um teste mais completo, indicando não apenas a precisão com que os resíduos são classificados corretamente como *hot spots*, mas também a porcentagem deles que são preditos.

$$\text{F-score} = \frac{2 \times \text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} \times \text{Sensibilidade}} \tag{15}$$

2.5.2.6 *AUC*

A medida AUC (*area under the ROC curve*), área sob a curva ROC (*receiver operating characteristic*) é a medida principal desse trabalho, uma vez que ela é principal medida para o treinamento e também é utilizada para selecionar os melhores descritores. A AUC tem uma vantagem sobre as medidas anteriores por ser uma medida independente do limiar escolhido para a classificação. As SVMs treinadas não classificam os resíduos simplesmente como *hot spots* e não *hot spots*, ou +1 e -1, mas sim atribuem valores entre +1 e -1. Normalmente, valores maiores que 0 são considerados *hot spots* e menores que 0 são considerados não *hot spots*, o que indica que o limiar é 0. Entretanto esse limiar pode ser alterado para mais ou para menos de forma a maximizar por exemplo, a precisão ou a sensibilidade. Desse modo, cada vez que esse limiar é alterado, as medidas acima também se alteram. No entanto, isso não ocorre com a AUC, pois nessa medida,

todos os valores possíveis para os limiares são utilizados no cálculo da curva ROC, o que o torna uma medida simples e mais eficiente para comparar descritores .

A curva ROC é um gráfico da sensibilidade, ou proporção de verdadeiros positivos, pela (1 − especificidade), ou proporção de falsos positivos, para cada valor do limiar. Consequentemente, o classificador ideal possui uma área sob a curva ROC igual a 1 e um método que classificasse os resíduos aleatoriamente exibiria uma AUC próxima a 0,5 (Fawcett, 2006).

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

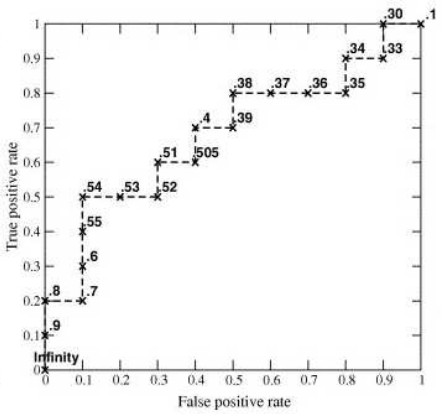


Figura 10: Exemplo de uma curva ROC com 20 dados, cada um com um score atribuído pelo classificador e a classe real a que pertencem (Figura extraída de Fawcett (2006).

2.6 SELEÇÃO DOS MELHORES DESCRITORES

A seleção dos melhores descritores teve dois objetivos: melhorar a performance do classificador e fornecer uma melhor compreensão dos dados através da observação dos descritores selecionados. Neste trabalho foram usados dois métodos distintos de seleção de descritores, um chamado de *Forward Selection* e o outro de *Backward Elimination* que são comumente utilizados em métodos de aprendizado de máquina (Guyon e Elisseeff, 2003).

No método de seleção *Forward Selection* um descritor é acrescentado por vez ao conjunto de descritores e os três melhores conjuntos selecionados de acordo com a medida AUC do classificador são mantidos para o rodada seguinte onde outro descritor será acrescentado. Ao final, quando todos os descritores já tiverem sido acrescentados ao conjunto, o método retorna o melhor conjunto de descritores, onde melhor significa que apresentou o maior valor de AUC na validação.

O método de seleção *Backward Elimination* é semelhante ao método anterior, sendo que a única diferença é que ao invés do método acrescentar descritores ao conjunto, ele começa com o conjunto contendo todos os descritores ele vai removendo um a um e testando o desem-

penho do conjunto resultante. Os três melhores conjuntos de uma rodada são conservados para a rodada seguinte onde outro descritor será eliminado. Ao final, quando todos os descritores forem removidos, o método retorna o melhor conjunto de descritores encontrado, onde novamente chamamos de melhor conjunto de descritores aquele que ao ser utilizado pela SVM apresente o maior valor de AUC na validação.

2.7 TESTE E COMPARAÇÃO COM OUTROS CLASSIFICADORES

Com o objetivo de obtermos uma melhor comparação dos 12 classificadores obtidos neste trabalho, com outros classificadores de *hot spots* publicados na literatura, nós treinamos os nossos classificadores em todo conjunto de dados de treinamento e validação que possui resíduos obtidos do ASEdb, e os aplicamos no conjunto de teste, que possui resíduos obtidos do BID.

3 RESULTADOS E DISCUSSÃO

3.1 ANÁLISE ESTATÍSTICA DOS DESCRITORES

Nesta seção serão expostos os 76 dos 298 descritores que apresentaram um valor médio diferente entre o grupo de resíduos que são *hot spots* e os que não *hot spots*. A análise estatística foi feita apenas para os dados do conjunto de treinamento e validação com resíduos obtidos do ASE. Nesse caso os valores médios foram considerados diferentes apenas se apresentaram um $p\,valor \leq 0,05$ calculado utilizando o teste t de Welch bicaudal. Separamos os descritores entre os provenientes do STING_DB e os propostos nesse trabalho.

3.1.1 *Descritores provenientes do STING_DB*

3.1.1.1 *Acessibilidade ao solvente*

A acessibilidade do resíduo ao solvente é uma característica conhecida dos *hot spots* descrita desde o trabalho de Bogan e Thorn (1998), que observou que os *hot spots* em geral tem pouco acessibilidade ao solvente. Na tabela 13 podemos notar que a diferença da acessibilidade ao solvente quando a proteína esta em complexo (ACCC) é a maior entre os dois grupos (*hot spots* e não *hot spots*). Outro resultado importante que observamos é a diferença entre os grupos observada na acessibilidade ao solvente calculada com as proteínas isoladas (ACCI). Essa diferença indica que os resíduos que serão *hot spots* tem em geral uma maior área acessível ao solvente, consequentemente, eles terão uma maior área de contato com a outra proteína ligante quando o complexo for formado. A acessibilidade relativa (ACCR) confirma a diferença da acessibilidade em isolamento, uma vez que torna independente o tamanho do resíduo.

Tabela 13: Descritores de acessibilidade ao solvente que apresentaram diferenças significativas entre os grupos *hot spots* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
ACCC	0,000	<	16,54	46,27
ACCI	0,030	>	97,30	78,55
ACCR	0,038	>	0,29	0,24

Diferença indica se a média é maior ou menor nos hot spots.

3.1.1.2 *Curvatura*

A curvatura também apresentou uma diferença significativa entre os dois grupos. No entanto, notamos que em alguns casos os valores para a curvatura do resíduo em complexo apresentava um valor de -999. Para analisarmos o quanto isso influenciou na diferença entre os grupos, nós repetimos a análise retirando esses valores discrepantes. Mesmo sem esses valores, foi observada uma diferença significativa (*p valor* = 0,000) da curvatura em complexo entre os grupos, entretanto, os *hot spots* apresentaram um valor maior (superfície côncava) para a curvatura que os não *hot spots* (Tabela 14). Apesar dessa diferença da curvatura média do resíduo em complexo, a curvatura calculada em isolamento não apresentou diferenças estatisticamente significante entre os grupos, sugerindo que não há preferência por superfícies côncavas ou convexas.

Tabela 14: Descritores da curvatura que apresentaram diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots*. Curvatura*é o resultado da análise do descritor após a remoção dos *outliers* com valor -999.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
Curvature	0,001	<	-301,41	-78,07
Curvature*	0,000	>	0,27	0,14

3.1.1.3 *Densidade*

A maior densidade atômica no microambiente ao redor do resíduo é uma característica conhecida em *hot spots.* Diversos trabalhos, tanto de predição de *hot spots* como trabalhos de análise de suas características indicam que a maior densidade atômica observada em regiões de *hot spots* não ocorre apenas quando o complexo está formado, mas preexiste nas proteínas isoladas (Li et al., 2004; Rajamani et al., 2004; Haliloglu et al., 2005; Cho et al., 2009). Neste trabalho, apenas descritores da densidade com esferas centradas na cadeia lateral apresentaram diferença (Tabela 15).

3.1.1.4 *Hidrofobicidade*

Apenas um dos seis descritores de hidrofobicidade do resíduo apresentou diferenças entre os dois grupos, a hidrofobicidade em complexo utilizando o índice de hidrofobicidade dos aminoácidos proposto pora Radzicka (1988). Aparentemente, os resíduos mais hidrofílicos são mais encobertos após a formação do complexo quando os mesmos são *hot spot.* Essa hipótese torna-se mais favorável quando

Tabela 15: Descritores de densidade atômica com diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
DensityLHA(3)	0,000	>	0,95	0,80
DensityLHAsw(3,3)	0,000	>	0,84	0,72
DensityLHAsw(5,3)	0,000	>	0,81	0,71
DensityLHAsw(7,3)	0,000	>	0,79	0,71
DensityLHAsw(7,4)	0,011	>	0,64	0,60
DensityLHAsw(7,5)	0,016	>	0,70	0,65
DensityLHAsw(7,6)	0,016	>	0,69	0,64
DensityLHAsw(9,3)	0,000	>	0,78	0,70
DensityLHAsw(9,4)	0,003	>	0,64	0,60
DensityLHAsw(9,5)	0,002	>	0,71	0,65
DensityLHAsw(9,6)	0,001	>	0,70	0,64

observamos o resultado da hidrofobicidade em complexo com o índice proposto po Kyte e Doolittle (1982) (*hot spots* = 0,24; não *hot spots* = -0,08) e comparamos os resultado com a hidrofobicidade em isolamento utilizando as duas escalas: HydroRI (*hot spots* = -1,01; não *hot spots* = -0,88) e HydroKDI (*hot spots* = -0,42; não *hot spots* = -0,35).

Tabela 16: Descritor de hidrofobicidade que demonstrou diferença significativa entre os grupos *hot spots* e não *hot spots.*

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
HydroRC	0,023	>	0,23	-0,33

3.1.1.5 *Densidade na interface*

A densidade na interface é um dos descritores que apresentou maior diferença entre os grupos *hot spots* e não *hot spots*, com todos os 6 descritores apresentando um *p valor* = 0,000 para o teste *t* (Tabela 17) de Welch. A maior densidade na interface quando os resíduos são *hot spots* está provavelmente relacionada a complementaridade entre as superfícies de ambas as proteínas, mas pode estar sendo influenciada também pela complementaridade dos *hot spots.* Alguns trabalhos sugerem que os *hot spots* de uma proteína, normalmente estão em contatos com os *hot spots* da outra proteína. Como os *hot spots* estão em regiões com maior densidade, isso pode também aumentar a densidade na interface.

Tabela 17: Descritores da densidade atômica na interface que apresentaram diferenças significativas estatisticamente entre os grupos *hot spots* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
IFRDensityCA(3)	0,000	>	0,96	0,59
IFRDensityCA(4)	0,000	>	0,73	0,44
IFRDensityCA(5)	0,000	>	0,65	0,39
IFRDensityCA(6)	0,000	>	0,62	0,36
IFRDensityLHA(3)	0,000	>	0,54	0,29
IFRDensityLHA(4)	0,000	>	0,51	0,26
IFRDensityLHA(5)	0,000	>	0,54	0,28
IFRDensityLHA(6)	0,000	>	0,54	0,27

3.1.1.6 *Esponjosidade na interface*

A esponjosidade na interface, juntamente com a densidade na interface, são os descritores que apresentam maior diferença entre os dois grupos de resíduos exibindo um *p valor* = 0,000 para o teste *t* dos 6 descritores (Tabela 18). A provável explicação para essa diferença é a mesma dos descritores de densidade na interface, ou seja, complementaridade das superfícies em contato e talvez a complementaridade dos *hot spots*.

Tabela 18: Descritores de esponjosidade na interface que demonstraram diferenças significativas entre os grupos *hot spots* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
IFRSpongeCA(3)	0,000	>	0,73	0,44
IFRSpongeCA(4)	0,000	>	0,67	0,40
IFRSpongeCA(5)	0,000	>	0,62	0,37
IFRSpongeCA(6)	0,000	>	0,58	0,34
IFRSpongeLHA(3)	0,000	>	0,58	0,33
IFRSpongeLHA(4)	0,000	>	0,54	0,29
IFRSpongeLHA(5)	0,000	>	0,54	0,28
IFRSpongeLHA(6)	0,000	>	0,53	0,27

3.1.1.7 *Energia de contatos na interface*

A energia de contatos na interface é a soma da energia dos contatos do resíduo analisado com os resíduos da outra proteínas. Estes des-

critores estão diretamente relacionados com o conceito de *hot spots* da interação proteína-proteína, o qual nos diz que os resíduos considerados *hot spots* são os que mais contribuem para a energia de ligação. Consequentemente, era esperado observamos uma maior energia de contatos na interface para esses resíduos e isso foi exatamente o que observamos.

Os dois primeiros descritores da tabela 19 somam todos os contatos, entretanto, o segundo descritor, diferentemente do primeiro, exclui do cálculo os contatos que envolvam molécula de água. Comparando esses dois descritores, é interessante notar que, apesar dos *hot spots* apresentarem menor superfície acessível ao solvente em complexo, a diferença da energia de contatos na interface quando consideramos os contatos com moléculas de água foi estatisticamente mais significativa.

Tabela 19: Descritores de energia de contatos na interface que apresentaram diferenças significativas em os grupos *hot spots* e não *hot spots*. Os três tipos de energia de contatos na tabela são: (true, true) todos os contatos, considerando contatos com moléculas de água; (true, false) todos os contatos excluindo contatos com moléculas de água; (false, true) contatos a curta distância e incluindo contatos com moléculas de água.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
InterfaceContactsEnergy(true,true)	0,004	>	18,49	6,86
InterfaceContactsEnergy(true,false)	0,007	>	17,27	6,51
InterfaceContactsEnergy(false,true)	0,000	>	10,28	3,54

3.1.1.8 *Número de contatos na interface*

O número de contatos na interface é separado em 14 tipos de contato que o resíduo faz na interface. Desses 14 tipos, apenas 5 apresentaram diferenças significativas (Tabela 20). O primeiro descritor da tabela indica o número de contatos hidrofóbicos, e como podemos notar esse número é maior para resíduos que são *hot spots*, provavelmente devido a complementaridade das interfaces que mantem os átomos mais próximos. O segundo descritor indica interações eletrostáticas atrativas e que assim como o anterior apresenta um número maior em *hot spots*. O terceiro descritor indica o número de ligações de hidrogênio entre átomos da cadeia principal de um resíduo e um átomo da cadeia lateral de um resíduo da outra proteína. É interessante notar que o número de ligações de hidrogênio entre átomos de cadeias laterais também é maior, no entanto, não há uma diferença estatisticamente significativa e por isso não está na tabela 20. O quarto descritor da tabela indica o número de ligações de hidrogênio entre átomos da

cadeia lateral dos resíduos, mas com intermédio de duas moléculas de água e como esperado, o número de ligações de hidrogênio desse tipo é menor em *hot spots*, mais especificamente, não foi observado interações desse tipo em resíduos que são *hot spots*. O quinto e último tipo de contato presente na tabela indica o número de interações π. O maior número de contatos dos tipos mencionados anteriormente reflete a preferência dos *hot spots* por determinados aminoácidos (Tabela 1), como triptofano e tirosina, que podem participar de interações π, hidrofóbicas e ligações de hidrogênio e a argininina que além dessas três interações também participa de interações eletrostáticas.

Tabela 20: Descritores do número de contatos na interface que apresentaram diferenças significativas entre os grupos *hot spots* e não *hot spots*. Os descritores estão separados de acordo com o tipo de interação.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
NumberofIFRContacts(1)	0,003	>	1,26	0,52
NumberofIFRContacts(2)	0,034	>	1,09	0,40
NumberofIFRContacts(7)	0,000	>	0,60	0,10
NumberofIFRContacts(12)	0,045	<	0,00	0,02
NumberofIFRContacts(13)	0,013	>	0,47	0,17

3.1.1.9 *Número de resíduos em contato na interface*

O número de resíduos em contato na interface é separado de acordo com o tipo, assim como no descritor anterior. O número de resíduos indica com quantos resíduos diferentes o resíduo analisado interage e assim como no caso anterior, os mesmos tipos apresentaram diferenças estatisticamente significativas (Tabela 21).

Tabela 21: Descritores do número de resíduos da interface que interagem com o resíduo analisado que apresentaram diferenças significativas. Os descritores são separados de acordo o tipo de interação.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
NumberofIFResidues(1)	0,000	>	0,79	0,30
NumberofIFResidues(2)	0,018	>	0,45	0,15
NumberofIFResidues(7)	0,000	>	0,55	0,10
NumberofIFResidues(12)	0,045	<	0,00	0,02
NumberofIFResidues(13)	0,013	>	0,47	0,17

3.1.1.10 *Número de contatos internos*

O número de contatos internos, assim com o número de contatos na interface, é dividido em 14 tipos diferentes de contatos. Desses 14 tipos, 2 apresentam diferenças significativas entre os grupos *hot spot* e não *hot spot* (Tabela 22). O primeiro descritor indica o número de ligações de hidrogênio entre átomos da cadeia principal de dois resíduos da mesma proteína, intermediados por uma molécula de água. O segundo indica o número de ligações de hidrogênio entre átomos da cadeia lateral de dois resíduos da mesma proteína, mas intermediados por duas moléculas de água. As moléculas de água, neste caso, são moléculas de água cocristalizadas com o complexo proteico. Apesar desse tipo de interação ser pequeno em resíduos que não são *hot spots*, ele foi observado algumas vezes, diferentemente dos *hot spots* que não apresentaram nenhuma vez esse tipo de interação. Acreditamos que isso seja resultado da pequena superfície acessível ao solvente apresentada pelos *hot spots*.

Tabela 22: Descritores do número de contatos internos que demonstraram diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots*. Os descritores estão separados pelo tipo de interação.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
NumberofINTContacts(5)	0,011	<	0,00	0,03
NumberofINTContacts(12)	0,045	<	0,00	0,02

3.1.1.11 *Número de resíduos internos em contato*

Esse descritor apresentou exatamente os mesmos resultado do descritor anterior que representa o número de contatos internos e por isso a motivo da diferença é o mesmo comentado anteriormente.

Tabela 23: Descritores do número de resíduos interagindo internamente que demonstraram diferenças significativas entre os grupos *hot spots* e não *hot spots*. Os descritores estão separados de acordo com o tipo de interação.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
NumberofINTResidues(5)	0,011	<	0,00	0,03
NumberofINTResidues(12)	0,045	<	0,00	0,02

3.1.2 *Descritores propostos neste trabalho*

3.1.2.1 *Número de átomos acessíveis ao solvente em região definida*

O número de resíduos acessíveis ao solvente em uma região definida é calculado apenas para a proteína isolada. Este descritor é calculado tanto para uma área ao redor do resíduo, incluindo ele próprio, quanto para uma coroa circular ao redor dele, na qual os átomos do resíduo são sempre excluídos. O número de átomos acessíveis ao solvente em diversas regiões calculadas apresentaram diferenças estatisticamente significantes. Essas diferenças ocorreram em regiões que incluíam os resíduos analisados (6 primeiros descritores da tabela 24) quanto em regiões da coroa circular que acreditamos que melhor representam a região do anel-O (5 últimos descritores da tabela 24). O número maior de átomos observados nessas regiões pode indicar a maior densidade, assim como pode indicar a possibilidade de realizar um maior número de interações entre as proteínas nessas regiões.

Tabela 24: Descritores do número de átomos acessíveis ao solvente em uma região definida que apresentaram diferenças estatisticamente significativas.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
NumAtomDist(3)	0,038	>	14,43	12,45
NumAtomDist(4)	0,041	>	19,49	17,21
NumAtomDist(6)	0,014	>	36,42	33,30
NumAtomDist(7)	0,003	>	45,85	41,94
NumAtomDist(8)	0,003	>	56,49	52,04
NumAtomDist(9)	0,003	>	68,43	62,87
NumAtomCCirc(6, 5)	0,046	>	9,11	8,13
NumAtomCCirc(7, 5)	0,024	>	18,55	16,78
NumAtomCCirc(8, 5)	0,038	>	29,19	26,88
NumAtomCCirc(9, 5)	0,025	>	41,13	37,71
NumAtomCCirc(9, 3)	0,048	>	54,00	50,43

3.1.2.2 *Área acessível ao solvente em região definida*

A área acessível ao solvente também é um descritor calculado em regiões ao redor do resíduo analisado e em regiões da coroa circular, as quais não incluem a área do resíduo. Dos 28 descritores da área acessível ao solvente em região definida, apenas um apresentou uma diferença significativa estatisticamente (Tabela 25). Não conseguimos compreender porque apenas esse descritor de área na coroa circular apresentou diferença. No entanto, o fato dos demais 27 descritores

não terem apresentado diferença é compreensível, pois a finalidade desse descritor era capturar diferenças que ocorressem devido a rugosidade e curvaturas da superfície nessas regiões e essa diferença não foi observada. Como o único descritor com diferença estatisticamente significativa apresenta um valor muito pequeno entre o raio maior da coroa circular (5 Å) e o raio menor (4 Å), acreditamos que essa diferença possa ser desprezada.

Tabela 25: Descritor da área acessível ao solvente em região definida que apresentou diferença significativa em os grupos *hot spot* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
ÁreaAccSolv(5, 4)	0,006	<	71,95	86,65

3.1.2.3 *Entalpia de hidratação em região definida*

Assim como os outros 3 descritores que foram propostos nesse trabalho, a entalpia de hidratação é calculada para regiões ao redor do resíduo analisado, incluindo ele próprio, e regiões da coroa circular que não incluem o resíduo analisado. Este descritor não apresentou diferença significativa em regiões muito próximas ao resíduo, entretanto, a entalpia de hidratação apresentou valores significativamente menores quando ampliamos o raio da esfera para 9 Å (Tabela 26). A soma das áreas acessíveis ao solvente dos átomos distantes até 9 Å dos átomos do resíduo analisado apresenta média de 750,4 $Å^2$e desvio padrão de 171,6 $Å^2$. Como o tamanho padrão das interfaces proteína-proteínas está na faixa de 1200 $Å^2$ a 2000$Å^2$ (Conte et al., 1999), a área analisada pelo descritor pode estar contida inteiramente na região de interface. No entanto, mesmo que isso não ocorra, acreditamos que isso não acarrete diferenças entre hot spots e não hot spots, uma vez que ambos estão sujeitos a inclusão de áreas que não pertencem a interface no cálculo da entalpia.

Também observamos uma entalpia de hidratação menor para algumas regiões da coroa circular que acreditamos conter átomos de resíduos que compõem o anel-O. No entanto, há a possibilidade dessa variação ocorrer devido a variação das área e não apenas a composição dos átomos expostos a superfície e por esse motivo nós propomos o tipo de descritor abaixo que representa a entalpia de hidratação normalizada pela área dos átomos dessa região.

3.1.2.4 *Entalpia de hidratação por área em região definida*

Como mencionado anteriormente, este descritor representa a entalpia de hidratação dividida pela área. Este descritor apresentou diferença significativa para várias regiões da coroa circular ao redor do resíduo

Tabela 26: Descritores da entalpia de hidratação calculado em uma região definida que apresentaram diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots*.

Descritor	*p valor*	Diferença	Média	
			hot spots	não *hot spots*
EntalpiaHidratação(9)	0,048	<	-421,05	-389,35
EntalpiaHidratação(5, 4)	0,046	>	-35,85	-42,69
EntalpiaHidratação(8, 5)	0,049	<	-176,81	-159,65
EntalpiaHidratação(9, 5)	0,042	<	-247,00	-223,20

analisado e apenas para essas regiões que excluem o resíduo analisado (Tabela 27). Este resultado foi o mais inesperado entre todos os descritores, pois indica que a região do anel-O apresenta menor entalpia de hidratação, o que consideramos maior hidrofilicidade, ao redor de resíduos que são *hot spots.* Nenhum dos trabalhos publicados na literatura que pesquisamos indica a ocorrência de tal característica. Na verdade, alguns trabalho, entre eles o de Moreira e colaboradores (2007b), citam exatamente o oposto, sugerindo que o anel-O é uma região hidrofóbica responsável por impedir o acesso do solvente aos *hot spots.* Outros trabalhos como o de Li e colaboradores(2005) propõem uma maior hidrofobicidade no centro da interface, o que equivale em seu estudo à *hot spots,* em relação aos resíduos da periferia, o qual ele mesmo refere-se como a região que forma o anel-O. Ainda em relação ao anel-O, alguns trabalhos como o de DeLano (2002) sugerem que o fato dos resíduos do anel-O não apresentarem grande variações na $\Delta\Delta G_{\text{ligação}}$ quando substituídos por alanina não necessariamente indicam que sua única função seja impedir o acesso do solvente, mas poderia ser explicado devido a substituições das interações que eles participam por moléculas de água. Acreditamos que nenhuma dessas hipóteses explica completamente o resultado inesperado observado analisando este descritor e que por isso, uma análise mais detalhada dessa região deveria ser feita em trabalhos futuros.

3.2 CLASSIFICADORES

Inicialmente, nós testamos dois conjuntos de descritores para construir os classificadores. O primeiro conjunto foi formado por todos os descritores mencionados neste trabalho, enquanto que o segundo, foi composto apenas pelos 76 descritores que demonstraram diferenças significativas entre os grupos *hot spots* e não *hot spots* quando comparados com o teste *t* de Welch. Para cada conjunto de descritores, foram aplicadas as máquinas de vetores de suporte linear e a não linear usando a função de kernel RBF. Apesar de haver recomendações para se utilizar a SVM linear em casos onde o número de descritores

Tabela 27: Descritores da entalpia de hidratação por área calculados em uma região definida que apresentaram diferenças significativas estatisticamente.

Descritor	*p valor*	Diferença	Média	
			*hot spots**	não *hot spots**
EHporÁrea(7, 3)	0,026	<	-0,54	-0,50
EHporÁrea(7, 4)	0,026	<	-0,54	-0,49
EHporÁrea(7, 5)	0,024	<	-0,54	-0,49
EHporÁrea(8, 3)	0,003	<	-0,55	-0,51
EHporÁrea(8, 4)	0,002	<	-0,55	-0,50
EHporÁrea(8, 5)	0,000	<	-0,56	-0,50
EHporÁrea(8, 6)	0,013	<	-0,59	-0,53
EHporÁrea(9, 3)	0,007	<	-0,54	-0,51
EHporÁrea(9, 4)	0,007	<	-0,54	-0,50
EHporÁrea(9, 5)	0,003	<	-0,55	-0,50
EHporÁrea(9, 6)	0,019	<	-0,57	-0,52

*Como a tabela possui apenas descritores da região ao redor dos resíduos analisados, as diferenças observadas correspondem a diferenças da região ao redor de *hot spots* (Anel-O) e da região ao redor de não *hot spots*.

é próximo ao número de dados, ou neste caso o número de resíduos presentes no banco de dados, e de utilizar a SVM não linear com funções de kernel quando o número de descritores é menor, não é possível afirmar com antecedência qual SVM apresentará os melhores resultados, e por isso testamos os dois métodos.

3.2.1 *SVM com todos os descritores*

A SVM linear identificou corretamente 45 do 53 hot spots na validação cruzada e apresentou 43 falsos positivos, resíduos preditos como hot spots, mas que não são (Tabela 28). Este classificador apresentou melhor desempenho que a SVM não linear utilizando a função RBF, o que está de acordo com a recomendação de utilizar-se a SVM linear quando o número de descritores (298) é igual ou próximo ao número de resíduos do conjunto de dados (283). No entanto, no conjunto de teste a SVM com kernel RBF apresentou 2 falso positivos a menos.

O resultado obtido pela SVM linear utilizando todos os descritores apresentou na validação cruzada o mesmo F-score observado no trabalho de Cho, Kim e Lee (2009), autores do Minerva. Entretanto o Minerva demonstrou maior precisão, enquanto que o nosso classificador apresentou maior sensibilidade (Tabelas 4 e 29). Isso indica que o Minerva identifica um menor número de *hot spots* aproximadamente 58% deles, mas quando classifica um resíduo como *hot spot* ele

Tabela 28: Matrizes de confusão dos classificadores (linear e com kernel RBF) que utilizam todos os descritores na predição dos *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	45	43	21	23
Preditos como não *hot spots*	8	187	10	51
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	44	62	21	21
Preditos como não *hot spots*	9	168	10	53
Total	53	230	31	74

apresenta 73% de chance de acertar. Nosso método, apesar de acertar apenas 54% dos resíduos classificados como *hot spots*, 85% dos *hot spots* analisados foram classificados como tal.

No conjunto de teste, tanto a SVM linear quanto a SVM com kernel RBF, apresentaram resultados semelhantes de desempenho. Ambos inferiores ao resultado da validação cruzada observado na SVM linear, mas com F-score semelhante ao obtido SVM com kernel RBF.

Tabela 29: Medidas de desempenho dos classificadores que utilizam todos os descritores na predição de *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,822	0,763	0,721	0,719
Acurácia	0,82	0,75	0,69	0,70
Precisão	0,54	0,45	0,48	0,50
Sensibilidade	0,85	0,84	0,68	0,68
Especificidade	0,81	0,73	0,69	0,72
F-score	0,65	0,57	0,56	0,58

3.2.2 *SVM com descritores estatisticamente diferentes*

Dos 298 descritores utilizados no classificador anterior, apenas 76 apresentaram diferenças significantes entre os grupos *hot spots* e não

hot spots quando submetidos ao teste *t* de Welch. Para avaliarmos se esses descritores seriam suficientes para obtermos um classificador com desempenho igual ou superior ao anterior, nós testamos duas SVMs, uma linear e um não linear, usando apenas esses descritores. O resultado observado na validação cruzada foi levemente superior ao obtido na validação cruzada utilizando todos os descritores e novamente, a SVM linear mostrou-se mais eficaz que a SVM com função de kernel RBF (Tabelas 30 e 31).

Tabela 30: Matrizes de confusão dos classificadores (linear e com kernel RBF) que utilizam somente os 76 descritores selecionados por apresentarem diferenças significativas entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	48	39	23	23
Preditos como não *hot spots*	5	191	8	51
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	47	52	23	24
Preditos como não *hot spots*	6	178	8	50
Total	53	230	31	74

Novamente, quando os classificadores foram aplicados ao conjunto de teste, o desempenho de ambas SVMs foram inferiores ao desempenho observado na validação cruzada. Entretanto, um pouco superior as SVMs testadas que utilizaram todos os classificadores.

3.3 SELEÇÃO DOS DESCRITORES

A seleção dos descritores foi realizada utilizando dois métodos diferentes, um chamado *forward selection* e o outro *backward elimination*, ambos já descritos anteriormente. Os métodos de seleção foram utilizados em conjunto com as mesmas máquinas de vetores de suporte, uma linear e outra com função de kernel RBF. A seleção foi realizada inicialmente a partir do conjunto com todos os descritores e a combinação de descritores que apresentar maior valor da AUC na validação é selecionado. Neste caso, 115 descritores dos 298 totais foram selecionados pelo menos uma vez por um dos métodos. Destes 115 descritores, 43 deles pertencem ao conjunto de 76 descritores que apresentaram diferenças significativas estatisticamente.

Tabela 31: Medidas de desempenho dos classificadores que utilizam somente os 76 descritores selecionados por serem estatisticamente diferentes entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,870	0,822	0,745	0,779
Acurácia	0,84	0,79	0,70	0,70
Precisão	0,57	0,55	0,50	0,49
Sensibilidade	0,90	0,89	0,74	0,74
Especificidade	0,83	0,77	0,69	0,68
F-score	0,68	0,64	0,60	0,59

Realizamos também a seleção de descritores a partir do conjunto de 76 descritores que apresentaram diferenças quando analisados com o teste *t* de Welch. Nesse caso, 12 dos descritores não foram selecionados nenhuma vez. Acreditamos que isso ocorreu porque alguns descritores contenham informação que outros descritores também fornecem.

Os descritores mais selecionados foram a acessibilidade ao solvente em complexo; a curvatura também em complexo; a densidade atômica com a esfera centrada no último átomo da cadeia lateral com exceção do hidrogênio e a média desse descritor calculada utilizando janelas deslizantes; a densidade e a esponjosidade na interface; o número de contatos na interface para alguns tipos de interação; o número de átomos em regiões definidas ao redor do resíduo e em coroas circulares; e a entalpia de hidratação por área para algumas regiões da coroa circular. A tabela completa dos descritores selecionados em cada processo de seleção está disponível no anexo A.1.

A seguir estão os resultados dos classificadores obtidos com a seleção de descritores.

3.3.1 *SVM com forward selection*

O método de seleção *forward selection* em conjunto com a SVM linear selecionou 36 descritores e com a SVM com kernel RBF selecionou 25 descritores. Novamente, a SVM linear apresentou um desempenho superior na validação cruzada a SVM com kernel RBF apesar do pequeno número de descritores selecionados.

A SVM linear identificou corretamente 48 dos 53 *hot spots* na validação cruzada com apenas 20 falsos positivos, enquanto que a SVM com kernel RBF identificou 49 *hot spots*, mas com 31 falsos positivos (Tabela 32).

Tabela 32: Matriz de confusão dos classificadores que utilizaram descritores selecionados pelo método de seleção de *forward selection*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	48	20	21	22
Preditos como não *hot spots*	5	210	10	52
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	49	31	13	22
Preditos como não *hot spots*	4	199	18	52
Total	53	230	31	74

Quando comparamos o desempenho na validação cruzada com o obtido pelo Minerva observamos um desempenho consideravelmente superior do nosso método, o qual demostrou um F-score de aproximadamente 80%, enquanto no Minerva esse valor é de aproximadamente 65% (Tabela 4 e 33). Essa superioridade é atribuída a maior sensibilidade do nosso classificador (0,91 vs. 0,58 no Minerva), mas sem comprometer a precisão (0,74 vs. 0,73 no Minerva). No entanto, quando aplicado ao conjunto de teste, os classificadores não mantiveram o bom desempenho, sendo que a SVM com kernel RBF apresentou o pior desempenho entre todos os outros propostos neste trabalho.

Tabela 33: Medidas de desempenho dos classificadores quando combinados com o método de seleção de descritores *forward selection.* A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,919	0,911	0,772	0,626
Acurácia	0,91	0,88	0,70	0,62
Precisão	0,74	0,66	0,49	0,37
Sensibilidade	0,91	0,92	0,68	0,42
Especificidade	0,91	0,87	0,70	0,70
F-score	0,80	0,75	0,57	0,39

3.3.2 *SVM com backward elimination*

O método de seleção *backward elimination* selecionou 35 descritores quando utilizado com a SVM linear e 64 descritores quando combinado com a SVM com kernel RBF. Apesar desse método resultar em classificadores com F-score superior ao obtido pelo Minerva quando calculado pela validação cruzada, ele não foi tão eficiente quanto o método *forward selection* em relação a mesma medida. No entanto, como mencionado anteriormente, os descritores foram selecionados de acordo com a área sob a curva ROC (AUC), e em relação a esta medida a SVM linear em conjunto com a seleção por *backward elimination* obteve o maior valor na validação cruzada (0,933) entre todos os classificadores nesse trabalho (Tabela 35).

Tabela 34: Matriz de confusão dos classificadores que utilizaram descritores selecionados pelo método de seleção de *backward elimination*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	50	27	25	23
Preditos como não *hot spots*	3	203	6	51
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	47	23	22	24
Preditos como não *hot spots*	6	207	9	50
Total	53	230	31	74

3.3.3 *SVM com forward selection (Seleção estatística)*

O método de seleção de descritores *forward selection* foi também aplicado aos 76 descritores selecionados por apresentarem diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots* quando analisados através do teste *t* de Welch. Foram selecionados 14 descritores quando o método foi aplicado em conjunto com a SVM linear e 34 quando combinado com a SVM com kernel RBF. Novamente a SVM linear apresentou um melhor desempenho na validação cruzada quando comparada a SVM com kernel RBF. Este desempenho superior foi, em grande parte, resultado da maior precisão, ou seja, menor número de falsos positivos, que na SVM linear foi de 29 resíduos e na SVM com kernel RBF foi de 37 resíduos (Tabela 36). En-

Tabela 35: Medidas de desempenho dos classificadores quando combinados com o método de seleção de descritores *backward elimination*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,933	0,908	0,769	0,728
Acurácia	0,89	0,90	0,72	0,69
Precisão	0,69	0,70	0,52	0,48
Sensibilidade	0,95	0,89	0,81	0,71
Especificidade	0,88	0,90	0,69	0,68
F-score	0,78	0,77	0,63	0,57

tretanto, quando os classificadores foram aplicados no conjunto de dados teste o desempenho da SVM com kernel RBF foi superior ao obtido com a SVM linear.

Tabela 36: Matriz de confusão dos classificadores (SVM linear e SVM com kernel RBF) combinados com o método de seleção de descritores *forward selection* aplicado ao conjunto de 76 descritores selecionados por apresentarem diferenças significativas entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	49	29	22	29
Preditos como não *hot spots*	4	201	9	45
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	48	37	25	27
Preditos como não *hot spots*	5	193	6	47
Total	53	230	31	74

3.3.4 *SVM com backward elimination (Seleção estatística)*

O método de seleção *backward selection* também foi aplicado aos 76 descritores selecionados de acordo com o resultado do teste t de Welch. Neste caso a SVM com kernel RBF utilizando 43 descritores

Tabela 37: Medidas de desempenho dos classificadores combinados com o método de seleção de descritores *forward selection* aplicado ao conjunto de 76 descritores selecionados por apresentarem diferenças significativas entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,913	0,895	0,745	0,756
Acurácia	0,88	0,85	0,64	0,69
Precisão	0,66	0,60	0,43	0,48
Sensibilidade	0,93	0,91	0,71	0,81
Especificidade	0,87	0,84	0,61	0,64
F-score	0,76	0,72	0,54	0,60

obteve um desempenho na validação cruzada um pouco superior ao da SVM linear (Tabela 38) com 23 descritores selecionados.

A SVM com kernel RBF em conjunto com os 43 descritores selecionados foi o classificador deste trabalho que demonstrou o melhor desempenho no conjunto de teste, segundo as medidas de AUC (0,799) e F-score (0,64). Este desempenho é explicado pela alta sensibilidade do classificador (0,90), o que indica que identificou 90% dos *hot spots* presentes nos 13 complexos representados no conjunto de dados de teste (BID). No entanto, apesar da alta sensibilidade, a precisão foi de apenas 0,49, o que indica que dos resíduos que o método classifica com *hot spots*, apenas aproximadamente metade são realmente *hot spots.* Devido ao melhor desempenho no conjunto de testes, esse foi o classificador escolhido para comparação com outros métodos da literatura.

Os 43 descritores selecionados e utilizados pelo classificador deste trabalho que apresentou o melhor desempenho são provenientes de 9 tipos de descritores. São eles:

1. Acessibilidade do resíduo ao solvente - dois classificadores de acessibilidade ao solvente, um que corresponde ao resíduo quando as proteínas estão em complexo (ACCC) e o outro a acessibilidade do resíduo relativa (ACCR);

2. Densidade - sete descritores de densidade todos calculados com janela deslizante e esfera centrada no último átomo da cadeia lateral com excessão do hidrogênio e com valores de raios de 3 a 5 Å;

3. Densidade na interface - seis descritores de densidade na interface, sendo três com esferas centradas no C_α e três com esferas

Tabela 38: Matriz de confusão dos classificadores (SVM linear e SVM com kernel RBF) combinados com o método de seleção de descritores *backward selection* aplicado ao conjunto de 76 descritores selecionados por apresentarem diferenças significativas entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

SVM (linear)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	52	46	23	25
Preditos como não *hot spots*	1	184	8	49
Total	53	230	31	74

SVM (RBF)	Validação cruzada		Teste	
	hot spots	não *hot spots*	*hot spots*	não *hot spots*
Preditos como *hot spots*	52	44	28	29
Preditos como não *hot spots*	1	186	3	45
Total	53	230	31	74

centradas no último átomo da cadeia lateral com excessão do hidrogênio;

4. Esponjosidade na interface - sete descritores de esponjosidade na interface, sendo quatro calculados com esferas centradas no C_αe três com esferas centradas no último átomo da cadeia lateral com excessão do hidrogênio;

5. Número de contatos na interface - dois descritores que correspondem a dois tipos de contatos, um correspondente ao número de ligações de hidrogênio entre um átomo da cadeia lateral, duas moléculas de água, e outro átomo de uma cadeia lateral, e o outro correspondente ao número de interações π;

6. Número de resíduos em contato na interface - dois descritores, um que indica o número de resíduos realizando interações eletrostáticas atrativas com o resíduo analisado, e o outro que indica o número de resíduos participando de ligações de hidrogênio envolvendo a cadeia principal de um resíduo com a cadeia lateral de outro resíduo;

7. Número de átomos acessíveis ao solvente em região definida - nove descritores que indicam o número de átomos acessíveis ao solvente foram selecionados. Seis deles são de regiões que incluem os átomos do resíduo analisado e os três restantes são de regiões ao redor do resíduo analisado excluindo os seus átomos;

Tabela 39: Medidas de desempenho dos classificadores combinados com o método de seleção de descritores *backward elimination* aplicado ao conjunto de 76 descritores selecionados por apresentarem diferenças significativas entre os grupos *hot spots* e não *hot spots*. A validação cruzada foi feita com resíduos obtidos do ASEdb e o teste, com o conjunto de teste composto por resíduos do BID.

Medida	Validação Cruzada		Teste	
	SVM (linear)	SVM (RBF)	SVM (linear)	SVM (RBF)
AUC	0,917	0,902	0,769	0,799
Acurácia	0,83	0,84	0,69	0,70
Precisão	0,56	0,59	0,48	0,49
Sensibilidade	0,98	0,98	0,74	0,90
Especificidade	0,80	0,81	0,66	0,61
F-score	0,71	0,72	0,58	0,64

8. Entalpia de hidratação em região definida - dois descritores da entalpia de hidratação foram selecionados. Um deles corresponde a entalpia de hidratação dos átomos expostos ao solvente com distância de até 9 Å dos átomos expostos do resíduo analisado. O outro descritor corresponde entalpia de hidratação dos átomos expostos ao solvente com distância inferior a 9 Å e superior a 3 Å dos átomos expostos ao solvente do resíduo analisado;

9. Entalpia de hidratação por área em região definida - foram selecionados seis descritores desse tipo. Todos eles calculados em regiões de coroa circular com raio maior entre 7 e 9 Å, e raio menor entre 3 e 5 Å.

3.4 RESULTADOS DOS TESTES E COMPARAÇÃO COM OUTROS CLASSIFICADORES

Todos os classificadores desenvolvidos neste trabalho demonstraram um desempenho consideravelmente inferior no conjunto de teste quando comparados aos seus resultados na validação cruzada. Para avaliar se isso era uma consequência da ineficácia dos classificadores deste trabalho, nós comparamos o desempenho do nosso melhor classificador, isto é, o que obteve o melhor desempenho no conjunto de teste, com outros classificadores da literatura que utilizam dados da estrutura do complexo e que foram discutidos anteriomente neste trabalho. Esses outros classificadores foram o Minerva, o KFC , o Rosetta Alanine Scanning e o FOLDEF . Quando comparamos o desempenho obtido na validação cruzada por nosso melhor classificador com esses outros classificadores, notamos que o nosso classificador apresentou o

maior F-score (0,72). Comparando o com o Minerva que apresentou o segundo maior F-score (0,65) notamos que as principais diferenças entre o desempenho dos dois está na precisão e na sensibilidade (Tabela 40). Enquanto o Minerva apresentou maior precisão, ele identificou um pouco mais da metade dos resíduos que são *hot spots*, como podemos notar pela sensibilidade de 0,58. Já nosso método apresentou precisão menor que o Minerva, mas próxima aos demais classificadores. Entretanto, a sensibilidade apresentada pelo nosso classificador indica que 98% dos resíduos que são *hot spots* foram classificados como tal. Além disso, outros classificadores propostos neste trabalho apresentaram um desempenho ainda maior na validação cruzada, como por exemplo a SVM linear em conjunto com o método de seleção de descritores *forward selection* que apresentou F-score médio igual a 0,80, com precisão de 0,74 e sensibilidade de 0,91. No entanto, esse classificador não obteve um bom desempenho no conjunto de teste, indicando que as regras aprendidas não foram generalizadas com eficácia, e por isso, este classificador não foi utilizado nesta comparação.

Tabela 40: Comparação do desempenho observado na validação cruzada do melhor classificador de *hot spots* obtido neste trabalho com outros classificadores da literatura.

Método	Precisão	Sensibilidade	Especificidade	F-score
Nosso Classificador*	0,59	0,98	0,81	0,72
Minerva	0,73	0,58	0,89	0,65
KFC	0,58	0,55	0,85	0,56
Rosetta	0,62	0,49	0,90	0,55
FOLDEF	0,59	0,32	0,93	0,41

* Nosso classificador corresponde a SVM com kernel RBF que utilizou 43 descritores estatisticamente diferentes seleciodos pelo método *backward elimination*.

Quando comparamos o desempenho do nosso classificador entre o obtido na validação cruzada e no conjunto de teste, notamos que ele apresenta um desempenho inferior. Isso ocorreu também com os outros classificados publicados que comparamos (Tabelas 40 e 41). Comparando o desempenho obtido pelo nosso classificador no conjunto teste com o desempenho dos outros classificadores quando aplicados ao mesmo conjunto, notamos novamente que o nosso classificador obteve um melhor desempenho de acordo com o F-score (0,64). Dentre os outros classificadores, o Minerva foi, como na validação cruzada, o que obteve o melhor desempenho. Comparando o nosso classificador ao Minerva, notamos que novamente, o Minerva apresenta maior precisão enquanto o nosso classificador apresenta maior sensibilidade (Tabela 41).

Como os resíduos do conjunto teste foram obtidos do BID, e neste banco de dados eles são classificados em ”forte”, ”intermediário”, ”fraco” e ”insignificante”, de acordo com sua contribuição na energia

Tabela 41: Comparação do desempenho observado do melhor classificador de *hot spots* obtido neste trabalho com outros classificadores da literatura quando aplicados ao conjunto de teste.

Método	Precisão	Sensibilidade	Especificidade	F-score
Nosso Classificador*	0,49	0,90	0,61	0,64
Minerva	0,70	0,52	0,91	0,59
KFC	0,45	0,29	0,85	0,35
Rosetta	0,52	0,35	0,86	0,42
FOLDEF	0,56	0,32	0,89	0,41

* Nosso classificador corresponde a SVM com kernel RBF que utilizou 43 descritores estatisticamente diferentes seleciodos pelo método *backward elimination*.

de ligação, nós decidimos avaliar qual a porcentagem de resíduos dessas classes são classificados como *hot spots*. Neste caso, nossa hipótese seria que quanto maior a contribuição energética dos resíduos da classes, maior seria a porcentagem deles classificados como *hot spots* (Tabela 42). Através dessa análise, observamos que quanto maior a contribuição do resíduo na energia de ligação, maior é a probabilidade dele ser classificado como *hot spot*. Quando comparamos essas porcentagens com outros classificadores, notamos que nosso classificador infelizmente apresenta uma probabilidade significativamente maior de classificar resíduos "fracos" e "insignificantes" como *hot spots.* O mesmo ocorreu com resíduos "intermediários", onde mais da metade deles foram classificados como *hot spots*. Entretando, quando observamos os resíduos "fortes", notamos que nosso classificador destaca-se dos demais, classificando corretamente como *hot spots* mais de 90% deles. Uma tabela completa com todos os resíduos do conjunto de teste e a predição de *hot spots* feita utilizando nosso classificador e os outros aos quais comparamos, encontra-se em anexo (Anexo A.2).

Tabela 42: Porcentagem de resíduos do conjunto de teste preditos como *hot spots* de acordo com a classificação fornecida pelo BID.

Classe	Resíduos classificados como *hot spots* (%)				
	Rosetta	FOLDEF	KFC	Minerva	Nosso Classificador
forte	35,5%	32,3%	29,0%	51,6%	90,3%
intermediário	27,8%	22,2%	33,3%	22,2%	61,1%
fraco	14,3%	4,8%	9,5%	4,8%	42,9%
insignificante	5,7%	8,6%	8,6%	5,7%	25,7%

Analisamos também, qual é eficácia do classificador por aminoácido, a qual pode ser observada na tabela 43. Através desta tabela, notamos que alguns aminoácidos como metioninas, serinas, treoninas, triptofanos e tirosinas, apresentaram um desempenho excelente

em nosso classificador sendo preditos com máxima acurácia. Lisinas, glutamatos e glicinas também apresentaram um bom desempenho. Os dois que destacam-se por apresentar um desempenho bem inferior aos demais foram a prolina e cisteína. Acreditamos que esse baixa acurácia obtida por esses aminoácidos seja consequência do pequeno número deles presentes no conjunto de treinamento (Tabela 8), entretanto, também há poucas metioninas e fenilalinas no conjunto de treinamento, mas mesmo assim, esses resíduos apresentaram boa acurácia.

Tabela 43: Desempenho do nosso classificador de acordo com o aminoácido.

Resíduo	Número de resíduos		Desempenho		
	hot spots	não *hot spots*	Precisão	Sensibilidade	Acurácia
ALA	0	1	-	-	1,00
ARG	3	7	0,33	0,67	0,50
ASN	0	2	0,00	-	0,50
ASP	2	4	0,40	1,00	0,50
CYS	0	1	0,00	-	0,00
GLN	0	3	0,00	-	0,67
GLU	2	6	0,50	0,50	0,75
GLY	1	3	0,50	1,00	0,75
HIS	1	1	0,50	1,00	0,50
ILE	2	4	0,50	1,00	0,67
LEU	4	6	0,44	1,00	0,50
LYS	1	10	0,33	1,00	0,82
MET	1	1	1,00	1,00	1,00
PHE	6	6	0,63	0,83	0,67
PRO	0	3	0,00	-	0,33
SER	0	4	-	-	1,00
THR	1	5	1,00	1,00	1,00
TRP	4	0	1,00	1,00	1,00
TYR	3	2	1,00	1,00	1,00
VAL	0	5	0,00	-	0,60

As medidas de desempenho que não apresentam um valor númerico, resultam em divisão por zero.

Outra fator que avaliamos foi a eficácia do classificador de acordo com o complexo (Tabela 44). Dentre os complexos que possuem resíduos no conjunto de teste, três deles destacam-se por terem apresentado acurácia máxima (1G3I, 1MQ8 e 2HHB) e um apresentou acurácia visivelmente inferior aos demais (1NUN). O complexo com código do PDB 1NUN, corresponde ao fator de crescimento de fibro-

Tabela 44: Desempenho do nosso classificador por complexos proteicos presentes no conjunto de teste.

Complexo	Número de resíduos		Desempenho		
PDB ID	*hot spots*	não *hot spots*	Precisão	Sensibilidade	Acurácia
1CDL	6	6	0,60	1,00	0,67
1DVA	5	18	0,42	1,00	0,70
1DX5	3	14	0,40	0,67	0,76
1EBP	4	5	0,60	0,75	0,67
1ES7	1	5	0,25	1,00	0,50
1FAK	2	17	0,25	1,00	0,68
1G3I	6	0	1,00	1,00	1,00
1IHB	0	4	0,00	-	0,75
1JAT	2	0	1,00	0,50	0,50
1MQ8	1	0	1,00	1,00	1,00
1NFI	1	1	0,50	1,00	0,50
1NUN	0	3	0	-	0,33
2HHB	0	1	-	-	1,00

blasto 10 ligado ao receptor do fator de crescimento de fibroblasto 2 (Yeh et al., 2003). No trabalho publicado por Yeh e colaboradores (2003) os autores resolvem a estrutura do complexo e estudam três mutantes: D76A, R78A e R155A, que correspondem aos três resíduos disponíveis no BID e portanto no conjunto de teste. Estes mutantes foram analisados por eles em ensaios de atividade. Dentre os mutantes, o R78A, predito pelo nosso classificador, pelo Rosetta Alanine Scanning e pelo FOLDEF como um *hot spot*, reduziu a atividade em 44% em relação a proteína tipo selvagem, enquanto os demais não apresentaram uma redução tão significativa (Yeh et al., 2003). No entanto, esses resíduos são descritos no BID, todos como "intermediários".

O método de classificação utilizado pelo BID (2003) não encontra-se no artigo publicado pelos autores, mas sabemos que os resíduos são classificados manualmente utilizando como fonte artigos da literatura. Diferentemente do ASEdb, onde todos os resíduos possuem valores da $\Delta\Delta G_{\text{ligação}}$, os resíduos do BID não possuem e muitas vezes são classificados de acordo com experimentos que não fornecem tal informação como no caso dos três resíduos mencionados anteriomente. Isso nos leva a questionar a exatidão da classificação desses resíduos, entretanto, mesmo com a possibilidade de erros de classificação dos resíduos no BID, acreditamos que esse conjunto de teste seja útil para efetuar uma comparação entre os classificadores.

Os classificadores de *hot spots* propostos neste trabalho foram também testados para classificar resíduos com $\Delta\Delta G_{\text{ligação}} >= 1,00$ Kcal/-

mol. Entretanto, todos os classificadores apresentaram AUC próxima a 0,5, o que indica que ele não é melhor que um classificador aleatório. Por esse motivo, o desempenho dos classificados para predizer esses resíduos não foi melhor analisado neste trabalho.

3.5 RESUMO GRÁFICO DOS RESULTADOS

A seguir, estão apresentados na forma gráfica para melhor visualização, os resultados discutidos anteriormente que foram obtidos pelos classificadores testados neste trabalho (Figuras de 11 a 18).

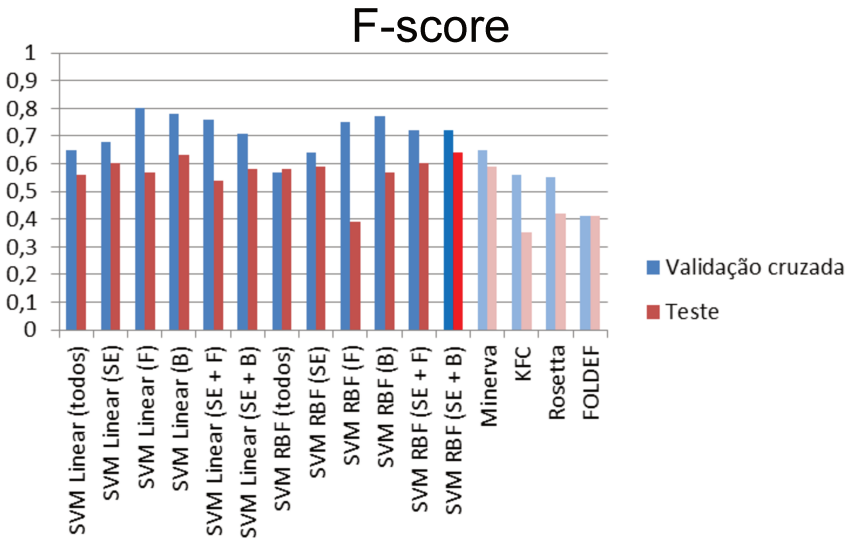


Figura 11: Comparação dos classificadores de *hot spots* de acordo com o F-score. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. As tonalidades mais claras são de classificadores da literatura e o com tonalidade mais viva o classificador que apresentou o melhor desempenho neste trabalho.

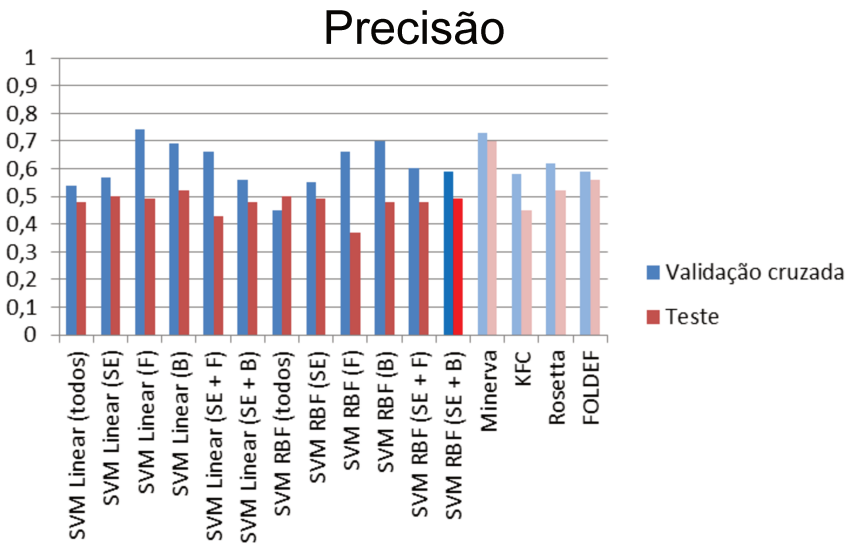


Figura 12: Comparação dos classificadores de *hot spots* de acordo com a precisão. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. As tonalidades mais claras são de classificadores da literatura e o com tonalidade mais viva o classificador que apresentou o melhor desempenho neste trabalho.

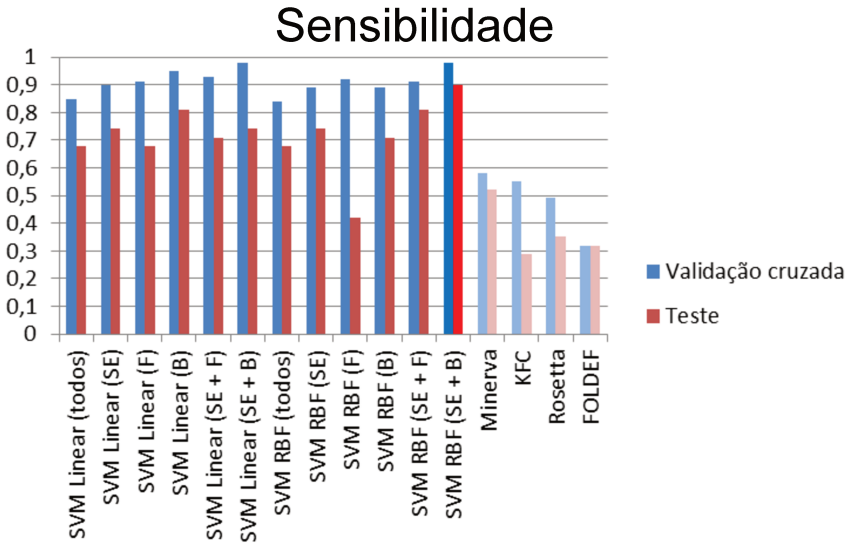


Figura 13: Comparação dos classificadores de *hot spots* de acordo com a sensibilidade. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. As tonalidades mais claras são de classificadores da literatura e o com tonalidade mais viva o classificador que apresentou o melhor desempenho neste trabalho.

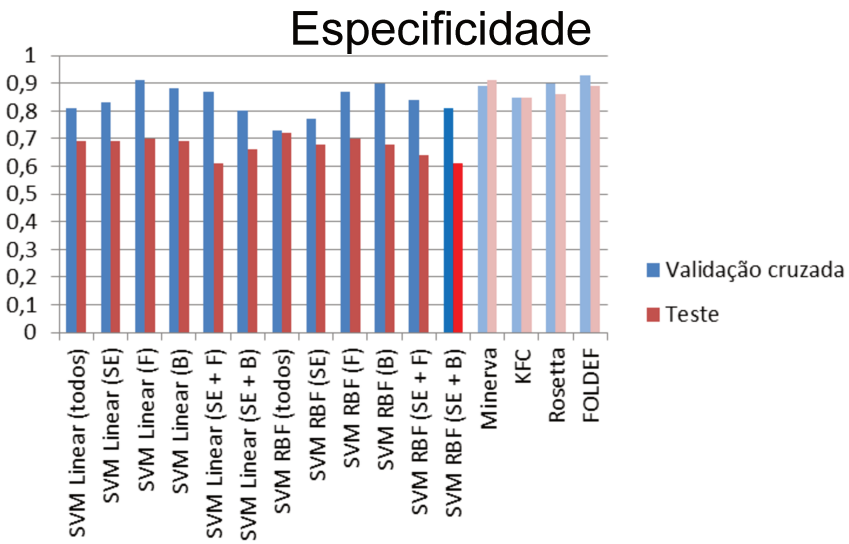


Figura 14: Comparação dos classificadores de *hot spots* de acordo com a especificidade. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. As tonalidades mais claras são de classificadores da literatura e o com tonalidade mais viva o classificador que apresentou o melhor desempenho neste trabalho.

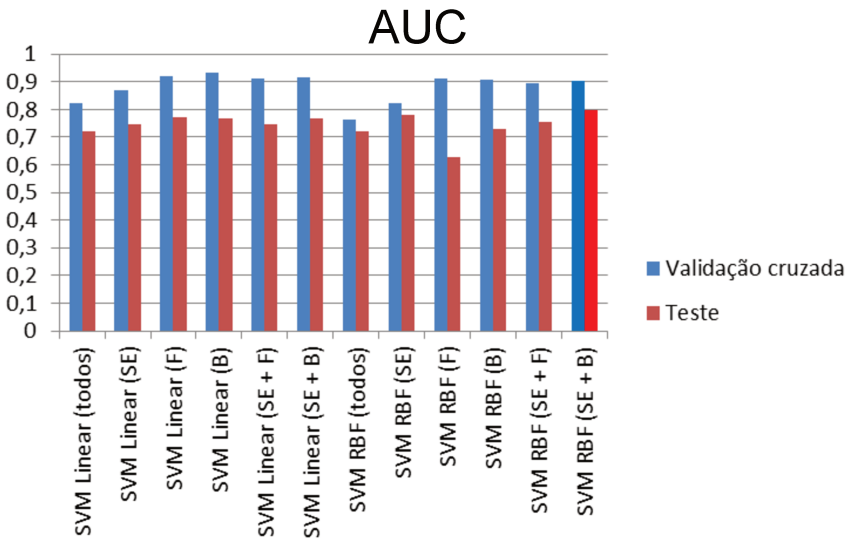


Figura 15: Comparação dos classificadores de *hot spots* de acordo com a AUC. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. Os classificadores publicados que foram utilizados para comparação não possuem dados da AUC.

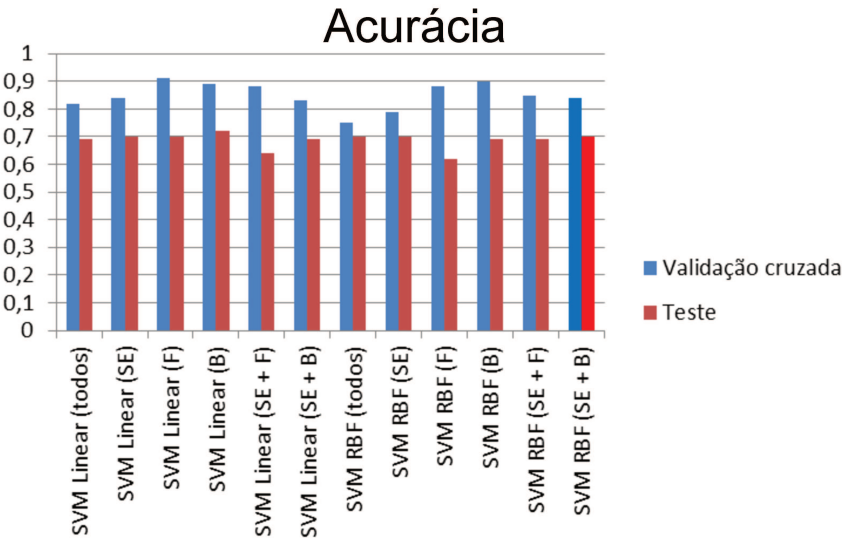
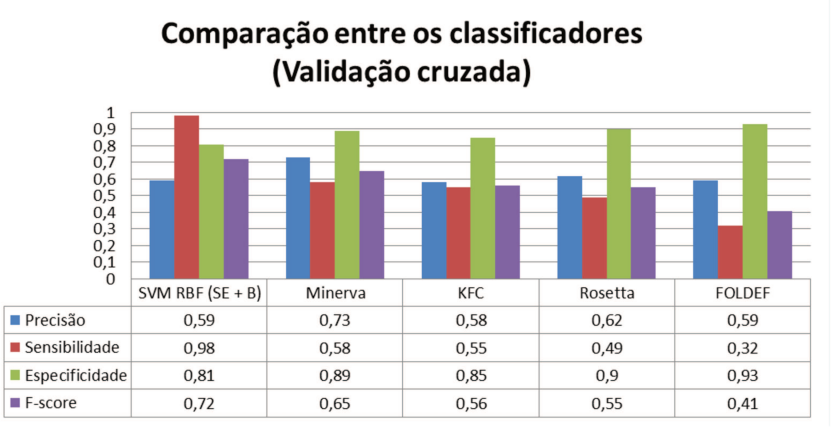
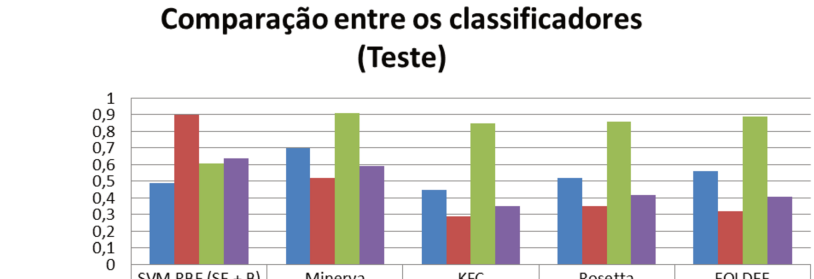


Figura 16: Comparação dos classificadores de *hot spots* de acordo com a acurácia. Em azul estão os resultados obtidos através da validação cruzada e em vermelho o resultado obtido no conjunto de teste. Os classificadores publicados que foram utilizados para comparação não possuem dados da acurácia.



	SVM RBF (SE + B)	Minerva	KFC	Rosetta	FOLDEF
Precisão	0,59	0,73	0,58	0,62	0,59
Sensibilidade	0,98	0,58	0,55	0,49	0,32
Especificidade	0,81	0,89	0,85	0,9	0,93
F-score	0,72	0,65	0,56	0,55	0,41

Figura 17: Comparação do desempenho obtido no conjunto de validação cruzada entre o melhor classificador deste trabalho e outros classificadores da literatura.



	SVM RBF (SE + B)	Minerva	KFC	Rosetta	FOLDEF
Precisão	0,49	0,7	0,45	0,52	0,56
Sensibilidade	0,9	0,52	0,29	0,35	0,32
Especificidade	0,61	0,91	0,85	0,86	0,89
F-score	0,64	0,59	0,35	0,42	0,41

Figura 18: Comparação do desempenho obtido no conjunto de teste entre o melhor classificador deste trabalho e outros classificadores da literatura.

4

CONCLUSÃO

O estudo de interações proteína-proteína tem diversas aplicações práticas como por exemplo o estudo do efeito de algumas mutações ou o desenvolvimento de fármacos que atuem impedindo que essas interações ocorram e consequentemente, influenciando o funcionamento de redes de interações (Arkin e Wells, 2004). Dentro desse contexto, o Grupo de Pesquisa em Biologia Computacional, CNPTIA, EMBRAPA, tem desenvolvido diversos trabalhos computacionais com o objetivo, tanto de caracterizar e melhor compreender as interações proteicas, como criar ferramentas computacionais disponíveis publicamente para auxiliar outras pesquisas. Entre estes trabalhos, temos a caracterização das interfaces proteicas e a predição dos resíduos que constituem estas interfaces através de métodos de aprendizado de máquina, sendo ambos os trabalhos desenvolvidos utilizando a plataforma STING e seus mais de 700 descritores estruturais disponíveis no STING_DB. Desta forma, este trabalho é complementar aos demais trabalhos desenvolvidos em nosso grupo, fornecendo uma visão ainda mais detalhada das interações proteína-proteína.

Este trabalho, assim como outros trabalhos já publicados na literatura sobre predição de *hot spots* em interfaces proteína-proteína, demonstrou que a utilização de métodos de aprendizado de máquina combinados com descritores estruturais da proteína é uma abordagem eficaz para identificar os resíduos com maior contribuição energética quando a estrutura do complexo estiver disponível. Esta eficácia ocorre devido dois fatores principais:(1) o grande número de descritores estruturais disponíveis, como por exemplo os disponíveis no STING_DB e (2) à rapidez e bom desempenho dos métodos de aprendizado de máquina, como demonstrado pelas SVMs em nosso trabalho, que uma vez treinadas em um conjunto de dados, podem ser aplicadas a quaisquer outros complexos proteicos com estrutura resolvida para predizer os *hot spots.* Isto permite construir programas computacionais específicos para essa tarefa que serão fáceis de serem utilizados pelos usuários, como por exemplo o KFC Server , ou incluí-lo como mais um dos descritores do Java Protein Dossier. Outros métodos, como a dinâmica molecular, também são aplicados a predição de *hot spots*, mas como exigem um poder computacional muito superior aos métodos de aprendizado de máquina sua aplicação é limitada a um número reduzido de resíduos e complexos proteicos, além de exigirem do usuário um bom conhecimento de como realizar a simulação e analisar os dados.

Em relação ao desempenho do melhor classificador obtido em nosso trabalho, observamos que ele superou o desempenho apresentado pelo Minerva, o melhor classificador ao qual comparamos. Entretanto, devido a maior precisão apresentada pelo Minerva e a maior sensibilidade apresentada pelo nosso classificador, acreditamos que eles possam atuar de forma complementar, cabendo ao usuário optar aquele que mais atenderá os seus propósitos.

A análise estatística dos descritores, juntamente com a seleção dos descritores combinada com as SVMs, demonstrou que a superfície do resíduo acessível ao solvente quando as proteínas estão em complexo; a densidade; as interações hidrofóbicas, eletrostáticas e π e as ligações de hidrogênio apresentam, como já descrito na literatura, características distintas entre os resíduos que são *hot spots* e os que não são. Outros descritores, especificamente os propostos nesse trabalho, buscaram caracterizar as regiões ao redor dos *hot spots*, chamado de anel-O. Até onde sabemos, nenhum outro trabalho utilizou descritores dessas regiões com o objetivo de predizer *hot spots*. Esses descritores demonstraram que há diferenças significativas na região do anel-O que podem ser utilizados na criação de classificadores com maior acurácia. Dentre os descritores do anel-O, a entalpia de hidratação por área foi o que apresentou maior diferença entre os *hot spots* e não *hot spots*. No entanto, esse resultado foi o oposto do esperado, demonstrando que o anel-O aparentemente é mais hidrofílico ao redor de *hot spots*. Não conseguimos explicar plenamente esse resultado e como não há nenhum trabalho publicado que comenta ou sustenta tal característica, acreditamos que análises complementares sobre essa região, com métodos mais rigosos como experimentos de dinâmica molecular para observar a flutução da densidade de moléculas de água próximas a essa região como feito por Patel et al. (2010) e Jamadagni et al. (2011), necessitam serem feitas antes de concluirmos qual a finalidade do anel-O ser mais hidrofílico.

Apesar dos bons resultados mencionados anteriormente, uma característica negativa deste trabalho, assim como de outros trabalhos que analisam as características de *hot spots* ou que tentam identificá-los nas interfaces proteicas, é o número reduzido de dados experimentais que podem ser utilizados nestes trabalhos. A aplicação dos métodos de aprendizado de máquina na classificação de *hot spots* é uma área de grande interesse, com diversos artigos sendo publicados nos últimos anos sobre o tema. No entanto, os dados utilizados nestes trabalhos são basicamente os mesmos analisados no trabalho de Bogan e Thorn (1998) e depois disponibilizados por eles mesmos no banco de dados ASEdb (2001). Este número reduzido de dados experimentais disponíveis torna as análises estatísticas menos significativas e podem diminuir a capacidade de predição de *hot spots* uma vez que há apenas um conjunto pequeno de exemplos para o aprendizado.

Entretanto, tal limitação só poderá ser diminuída com a produção e disponibilização de novos dados experimentais.

O classificador proposto neste trabalho que apresentou o melhor desempenho está sendo implantado no STING, e assim que esta etapa for finalizada conseguiremos fazer a predição de *hot spots* para todos os complexos proteicos depositados no PDB. Esses dados, assim como o dos descritores de entalpia de hidratação propostos neste trabalho, serão integrados ao STING_DB e poderão ser obtidos através do Java Protein Dossier.

APÊNDICE A

A.1 DESCRITORES SELECIONADOS

Os métodos de seleção da tabela abaixo são:

- SE - os 76 descritores selecionados por apresentarem diferenças estatisticamente significativas entre os grupos *hot spots* e não *hot spots* de acordo com o teste *t* de Welch;
- SVM^1+ FS - SVM linear em conjunto com o método de seleção *forward selection*;
- SVM^2+ FS - SVM com função de kernel RBF em conjunto com o método de seleção *forward selection*;
- SVM^1+ BE - SVM linear em conjunto com o método de seleção *backward elimination*;
- SVM^2+ BE - SVM com função de kernel RBF em conjunto com o método de seleção *backward elimination*;
- SVM^1+ FS(SE) - SVM linear em conjunto com o método de seleção *forward selection*, restrito apenas a descritores com diferenças significativas entre os grupos *hot spots* e não *hot spots* (SE);
- SVM^2+ FS(SE) - SVM com função de kernel RBF em conjunto com o método de seleção *forward selection*, restrito a descritores presentes na SE;
- SVM^1+ BE(SE) - SVM linear em conjunto com o método de seleção *backward elimination*, restrito a descritores presentes na SE;
- SVM^2+ BE(SE) - SVM com função de kernel RBF em conjunto com o método de seleção *backward elimination*, mas restrito a descritores presentes na SE.

Os descritores selecionados aparecem marcados com ✓. Descritores que foram selecionados em mais da metade dos métodos de seleção de descritores estão destacados em azul. Ao final da tabela encotra-se o número de descritores selecionados por cada método.

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
ACCC	✓	✓	✓	✓	✓	✓	✓	✓	✓
ACCI	✓			✓		✓		✓	

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
ACCR	✓					✓	✓		✓
CloCA									
CloLHA									
CpoCA		✓							
CpoLHA		✓							
Curvature	✓	✓		✓	✓	✓	✓	✓	
CurvatureIsolation				✓					
DensityCA(3)				✓					
DensityCA(4)					✓				
DensityCA(5)				✓					
DensityCA(6)									
DensityCAsw(3,3)			✓						
DensityCAsw(3,4)					✓				
DensityCAsw(3,5)									
DensityCAsw(3,6)									
DensityCAsw(5,3)					✓				
DensityCAsw(5,4)									
DensityCAsw(5,5)					✓				
DensityCAsw(5,6)									
DensityCAsw(7,3)									
DensityCAsw(7,4)					✓				
DensityCAsw(7,5)									
DensityCAsw(7,6)									
DensityCAsw(9,3)				✓					
DensityCAsw(9,4)									
DensityCAsw(9,5)									
DensityCAsw(9,6)									
DensityLHA(3)	✓						✓		
DensityLHA(4)									
DensityLHA(5)		✓		✓	✓				
DensityLHA(6)									
DensityLHAsw(3,3)	✓								✓
DensityLHAsw(3,4)				✓					
DensityLHAsw(3,5)		✓		✓					
DensityLHAsw(3,6)									

Descritor	SE	SVM¹+ FS	SVM²+ FS	SVM¹+ BE	SVM²+ BE	SVM¹+ FS(SE)	SVM²+ FS(SE)	SVM¹+ BE(SE)	SVM²+ BE(SE)
DensityLHAsw(5,3)	✓				✓			✓	
DensityLHAsw(5,4)			✓		✓				
DensityLHAsw(5,5)		✓							
DensityLHAsw(5,6)									
DensityLHAsw(7,3)	✓				✓			✓	
DensityLHAsw(7,4)	✓				✓		✓		✓
DensityLHAsw(7,5)	✓				✓	✓			✓
DensityLHAsw(7,6)	✓								
DensityLHAsw(9,3)	✓	✓		✓	✓		✓	✓	✓
DensityLHAsw(9,4)	✓								✓
DensityLHAsw(9,5)	✓			✓	✓				✓
DensityLHAsw(9,6)	✓						✓	✓	
EnergyDensityCAsw(3,3)			✓						
EnergyDensityCAsw(3,4)									
EnergyDensityCAsw(3,5)			✓		✓				
EnergyDensityCAsw(3,6)									
EnergyDensityCAsw(5,3)									
EnergyDensityCAsw(5,4)									
EnergyDensityCAsw(5,5)									
EnergyDensityCAsw(5,6)									
EnergyDensityCAsw(7,3)				✓					
EnergyDensityCAsw(7,4)									
EnergyDensityCAsw(7,5)									
EnergyDensityCAsw(7,6)									
EnergyDensityCAsw(9,3)									
EnergyDensityCAsw(9,4)				✓					
EnergyDensityCAsw(9,5)									
EnergyDensityCAsw(9,6)									
EnergyDensityLHAsw(3,3)									
EnergyDensityLHAsw(3,4)									
EnergyDensityLHAsw(3,5)									
EnergyDensityLHAsw(3,6)									
EnergyDensityLHAsw(5,3)									
EnergyDensityLHAsw(5,4)					✓				
EnergyDensityLHAsw(5,5)					✓				

Descritor	SE	SVM¹+ FS	SVM²+ FS	SVM¹+ BE	SVM²+ BE	SVM¹+ FS(SE)	SVM²+ FS(SE)	SVM¹+ BE(SE)	SVM²+ BE(SE)
EnergyDensityLHAsw(5,6)					✓				
EnergyDensityLHAsw(7,3)									
EnergyDensityLHAsw(7,4)									
EnergyDensityLHAsw(7,5)									
EnergyDensityLHAsw(7,6)									
EnergyDensityLHAsw(9,3)									
EnergyDensityLHAsw(9,4)									
EnergyDensityLHAsw(9,5)					✓				
EnergyDensityLHAsw(9,6)									
EPabsolute()									
EPaverage()			✓		✓				
EPca()					✓				
EPlha()									
EPsurface()		✓		✓					
HydroKD()			✓		✓				
HydroKDC()			✓		✓				
HydroKDI()			✓		✓				
HydroR()									
HydroRC()	✓				✓				
HydroRI()									
IFRDensityCA(3)	✓						✓		
IFRDensityCA(4)	✓		✓		✓		✓	✓	✓
IFRDensityCA(5)	✓								✓
IFRDensityCA(6)	✓	✓	✓		✓				✓
IFRDensityLHA(3)	✓		✓		✓	✓	✓	✓	✓
IFRDensityLHA(4)	✓		✓						
IFRDensityLHA(5)	✓		✓						✓
IFRDensityLHA(6)	✓	✓					✓		✓
IFRSpongeCA(3)	✓	✓	✓		✓	✓	✓		✓
IFRSpongeCA(4)	✓								✓
IFRSpongeCA(5)	✓				✓				✓
IFRSpongeCA(6)	✓		✓			✓	✓		✓
IFRSpongeLHA(3)	✓								
IFRSpongeLHA(4)	✓	✓			✓				✓
IFRSpongeLHA(5)	✓	✓			✓		✓		✓

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
IFRSpongeLHA(6)	✓	✓					✓		✓
InterfaceContactsEnergy(true,true)	✓	✓							
InterfaceContactsEnergy(true,false)	✓								
InterfaceContactsEnergy(false,true)	✓				✓	✓			
InternalContEnergy(true,true)		✓							
InternalContEnergy(true,false)									
InternalContEnergy(false,true)					✓				
ContactsEnergyAllsw(true,3)									
ContactsEnergyAllsw(true,5)									
ContactsEnergyAllsw(true,7)					✓				
ContactsEnergyAllsw(true,9)									
ContactsEnergyAllsw(false,3)									
ContactsEnergyAllsw(false,5)									
ContactsEnergyAllsw(false,7)									
ContactsEnergyAllsw(false,9)				✓					
ContactsEnergyShortsw(true,3)									
ContactsEnergyShortsw(true,5)									
ContactsEnergyShortsw(true,7)									
ContactsEnergyShortsw(true,9)									
ContactsEnergyShortsw(false,3)									
ContactsEnergyShortsw(false,5)									
ContactsEnergyShortsw(false,7)									
ContactsEnergyShortsw(false,9)									
NumberofIFRContacts(1)	✓								
NumberofIFRContacts(2)	✓						✓		
NumberofIFRContacts(3)									
NumberofIFRContacts(4)									
NumberofIFRContacts(5)									
NumberofIFRContacts(6)									
NumberofIFRContacts(7)	✓	✓		✓	✓	✓	✓	✓	
NumberofIFRContacts(8)		✓							
NumberofIFRContacts(9)									
NumberofIFRContacts(10)									
NumberofIFRContacts(11)									
NumberofIFRContacts(12)	✓	✓	✓	✓		✓	✓	✓	✓

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
NumberofIFRContacts(13)	✓								✓
NumberofIFRContacts(14)									
NumberofIFResidues(1)	✓	✓						✓	
NumberofIFResidues(2)	✓						✓	✓	✓
NumberofIFResidues(3)					✓				
NumberofIFResidues(4)									
NumberofIFResidues(5)									
NumberofIFResidues(6)									
NumberofIFResidues(7)	✓	✓							✓
NumberofIFResidues(8)									
NumberofIFResidues(9)									
NumberofIFResidues(10)									
NumberofIFResidues(11)									
NumberofIFResidues(12)									
NumberofIFResidues(13)	✓						✓		
NumberofIFResidues(14)	✓					✓			
NumberofINTContacts(1)									
NumberofINTContacts(2)									
NumberofINTContacts(3)					✓				
NumberofINTContacts(4)					✓				
NumberofINTContacts(5)	✓	✓					✓		
NumberofINTContacts(6)									
NumberofINTContacts(7)									
NumberofINTContacts(8)									
NumberofINTContacts(9)				✓	✓				
NumberofINTContacts(10)									
NumberofINTContacts(11)									
NumberofINTContacts(12)	✓						✓		
NumberofINTContacts(13)									
NumberofINTContacts(14)									
NumberofINTResidues(1)									
NumberofINTResidues(2)				✓					
NumberofINTResidues(3)									
NumberofINTResidues(4)		✓		✓					
NumberofINTResidues(5)	✓			✓	✓		✓	✓	

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
NumberofINTResidues(6)									
NumberofINTResidues(7)									
NumberofINTResidues(8)					✓				
NumberofINTResidues(9)									
NumberofINTResidues(10)					✓				
NumberofINTResidues(11)									
NumberofINTResidues(12)	✓			✓					
NumberofINTResidues(13)									
NumberofINTResidues(14)									
NumAtomDist(3)	✓								✓
NumAtomDist(4)	✓	✓							✓
NumAtomDist(5)									
NumAtomDist(6)	✓								✓
NumAtomDist(7)	✓				✓				✓
NumAtomDist(8)	✓		✓			✓	✓		✓
NumAtomDist(9)	✓								✓
NumAtomCCirc(4,3)									
NumAtomCCirc(5,3)									
NumAtomCCirc(5,4)									
NumAtomCCirc(6,3)				✓					
NumAtomCCirc(6,4)				✓					
NumAtomCCirc(6,5)	✓	✓			✓				
NumAtomCCirc(7,3)					✓				
NumAtomCCirc(7,4)					✓				
NumAtomCCirc(7,5)	✓								
NumAtomCCirc(7,6)									
NumAtomCCirc(8,3)				✓					
NumAtomCCirc(8,4)									
NumAtomCCirc(8,5)	✓		✓		✓		✓	✓	✓
NumAtomCCirc(8,6)			✓	✓					
NumAtomCCirc(8,7)									
NumAtomCCirc(9,3)	✓			✓	✓		✓	✓	✓
NumAtomCCirc(9,4)		✓							
NumAtomCCirc(9,5)	✓								✓
NumAtomCCirc(9,6)									

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
NumAtomCCirc(9,7)									
NumAtomCCirc(9,8)									
ÁreaAccSolv(3)									
ÁreaAccSolv(4)									
ÁreaAccSolv(5)									
ÁreaAccSolv(6)									
ÁreaAccSolv(7)									
ÁreaAccSolv(8)									
ÁreaAccSolv(9)									
ÁreaAccSolv(4,3)									
ÁreaAccSolv(5,3)									
ÁreaAccSolv(5,4)	✓						✓		
ÁreaAccSolv(6,3)		✓							
ÁreaAccSolv(6,4)			✓						
ÁreaAccSolv(6,5)									
ÁreaAccSolv(7,3)			✓						
ÁreaAccSolv(7,4)									
ÁreaAccSolv(7,5)									
ÁreaAccSolv(7,6)									
ÁreaAccSolv(8,3)			✓						
ÁreaAccSolv(8,4)		✓							
ÁreaAccSolv(8,5)				✓	✓				
ÁreaAccSolv(8,6)									
ÁreaAccSolv(8,7)									
ÁreaAccSolv(9,3)									
ÁreaAccSolv(9,4)					✓				
ÁreaAccSolv(9,5)					✓				
ÁreaAccSolv(9,6)									
ÁreaAccSolv(9,7)									
ÁreaAccSolv(9,8)									
EntalpiaHidratação(3)									
EntalpiaHidratação(4)									
EntalpiaHidratação(5)		✓							
EntalpiaHidratação(6)									
EntalpiaHidratação(7)									

Descritor	SE	SVM^1+ FS	SVM^2+ FS	SVM^1+ BE	SVM^2+ BE	SVM^1+ FS(SE)	SVM^2+ FS(SE)	SVM^1+ BE(SE)	SVM^2+ BE(SE)
EntalpiaHidratação(8)									
EntalpiaHidratação(9)	✓					✓	✓		✓
EntalpiaHidratação(4,3)									
EntalpiaHidratação(5,3)									
EntalpiaHidratação(5,4)	✓		✓✓					✓	
EntalpiaHidratação(6,3)		✓							
EntalpiaHidratação(6,4)		✓		✓					
EntalpiaHidratação(6,5)									
EntalpiaHidratação(7,3)									
EntalpiaHidratação(7,4)									
EntalpiaHidratação(7,5)									
EntalpiaHidratação(7,6)									
EntalpiaHidratação(8,3)									
EntalpiaHidratação(8,4)									
EntalpiaHidratação(8,5)	✓						✓		
EntalpiaHidratação(8,6)									
EntalpiaHidratação(8,7)									
EntalpiaHidratação(9,3)	✓						✓	✓	✓
EntalpiaHidratação(9,4)					✓				
EntalpiaHidratação(9,5)									
EntalpiaHidratação(9,6)									
EntalpiaHidratação(9,7)					✓				
EntalpiaHidratação(9,8)									
EHporÁrea(3)					✓				
EHporÁrea(4)									
EHporÁrea(5)									
EHporÁrea(6)			✓		✓				
EHporÁrea(7)									
EHporÁrea(8)									
EHporÁrea(9)			✓		✓				
EHporÁrea(4,3)				✓					
EHporÁrea(5,3)									
EHporÁrea(5,4)				✓					
EHporÁrea(6,3)									
EHporÁrea(6,4)		✓		✓					

Descritor	SE	SVM¹+ FS	SVM²+ FS	SVM¹+ BE	SVM²+ BE	SVM¹+ FS(SE)	SVM²+ FS(SE)	SVM¹+ BE(SE)	SVM²+ BE(SE)
EHporÁrea(6,5)									
EHporÁrea(7,3)	✓								✓
EHporÁrea(7,4)	✓				✓				
EHporÁrea(7,5)	✓							✓	✓
EHporÁrea(7,6)									
EHporÁrea(8,3)	✓						✓		✓
EHporÁrea(8,4)	✓					✓	✓		✓
EHporÁrea(8,5)	✓	✓	✓		✓				✓
EHporÁrea(8,6)	✓								
EHporÁrea(8,7)									
EHporÁrea(9,3)	✓				✓		✓	✓	
EHporÁrea(9,4)	✓				✓			✓	
EHporÁrea(9,5)	✓			✓	✓		✓	✓	✓
EHporÁrea(9,6)	✓							✓	
EHporÁrea(9,7)					✓				
EHporÁrea(9,8)					✓				
Número de descritores	76	36	25	35	64	14	34	23	43

A.2 COMPARAÇÃO DOS RESULTADOS DE TESTE

Na tabela estão os resíduos que compõem o conunto de dados de teste, indicando o código PDB do complexo ao qual pertencem, a cadeia, o tipo do resíduo, sua numeração na estrutura e a "força" de acordo com o BID, onde resíduos "fortes" possuem $\Delta\Delta G_{\text{ligação}}$ >= 2,0 Kcal/mol e por isso somente eles foram considerados como *hot spots*.

Abaixo são comparados cinco métodos de predição de *hot spots* e os resultados obtidos no conjunto de teste. Os métodos foram Rosetta Alanine Scanning , FOLDEF , KFC e Minerva, além do melhor classificador obtido neste trabalho. Como o Rosetta e o FOLDEF fazem a predição do valor númerico da $\Delta\Delta G_{\text{ligação}}$, resíduos que resultaram em valores maiores que 2,00 Kcal/mol foram considerados preditos como *hot spots*. Na tabela, resíduos preditos incorretamente como *hot spots* estão marcados com ✗ e resíduos preditos corretamente como *hot spots* estão marcados com ✓.

PDB ID	Cadeia	Resíduo		Força	Rosetta	FOLDEF	KFC	Minerva	Nosso Classificador
1CDL	A	PHE	12	insignificante	-	-	-	✗	-
1CDL	A	PHE	19	fraco	-	-	✗	-	✗
1CDL	A	PHE	92	forte	-	-	✓	✓	✓
1CDL	E	LYS	799	insignificante	-	✗	-	-	✗
1CDL	E	TRP	800	forte	✓	✓	✓	✓	✓
1CDL	E	LYS	802	intermediário	-	✗	-	-	✗
1CDL	E	GLY	804	forte	-	-	-	-	✓
1CDL	E	ARG	808	intermediário	✗		✗		✗
1CDL	E	ILE	810	forte	✓	-	✓	✓	✓
1CDL	E	GLY	811	intermediário	-	-	-	-	-
1CDL	E	ARG	812	forte	-	-	-	✓	✓
1CDL	E	LEU	813	forte	✓	✓	✓	-	✓
1DVA	H	GLY	38	intermediário	-	-	-	-	✗
1DVA	H	ILE	65	insignificante	-	-	-	-	-
1DVA	H	VAL	67	insignificante	-	-	-	-	✗
1DVA	H	GLU	70	fraco	-	-	-	-	✗
1DVA	H	LEU	73	insignificante	-	-	✗	✗	-
1DVA	H	SER	74	insignificante	-	-	-	-	-
1DVA	H	GLU	75	insignificante	-	-	-	-	-
1DVA	H	HIS	76	forte	-	-	-	-	✓
1DVA	H	GLU	80	insignificante	-	-	-	-	-
1DVA	H	SER	82	insignificante	-	-	-	-	-
1DVA	H	LEU	144	insignificante	-	-	-	-	✗
1DVA	H	LEU	153	fraco	-	-	-	-	✗
1DVA	X	ALA	1	insignificante	-	-	-	-	-
1DVA	X	LEU	2	forte	✓	✓	-	✓	✓
1DVA	X	ASP	5	fraco	-	-	-	-	-
1DVA	X	ARG	7	fraco	✗	-	-	-	✗
1DVA	X	VAL	8	intermediário	-	-	-	-	-
1DVA	X	ASP	9	intermediário	-	-	-	✗	✗
1DVA	X	TRP	11	forte	✓	✓	-	-	✓
1DVA	X	TYR	12	forte	✓	-	✓	-	✓
1DVA	X	GLN	14	insignificante	-	-	-	-	-

PDB ID	Cadeia	Resíduo		Força	Rosetta	FOLDEF	KFC	Minerva	Nosso Classificador
1DVA	X	PHE	15	forte	-	-	✓	✓	✓
1DVA	X	VAL	16	insignificante	-	-	-	-	-
1DX5	N	ILE	24	insignificante	-	-	-	-	-
1DX5	N	PHE	34	intermediário	-	-	✗	-	-
1DX5	N	LYS	36	fraco	-	-	-	-	-
1DX5	N	PRO	37	fraco	-	-	-	-	-
1DX5	N	GLN	38	fraco	✗	-	-	✗	✗
1DX5	N	GLU	39	insignificante	-	-	-	-	-
1DX5	N	LEU	65	fraco	-	-	-	-	✗
1DX5	N	ARG	67	forte	-	-	-	-	-
1DX5	N	THR	74	fraco	-	-	-	-	-
1DX5	N	ARG	75	fraco	-	-	-	-	-
1DX5	N	TYR	76	forte	-	-	✓	-	✓
1DX5	N	GLU	80	forte	✓	-	-	-	✓
1DX5	N	LYS	81	fraco	-	-	-	-	-
1DX5	N	ILE	82	intermediário	-	-	-	✗	✗
1DX5	N	MET	84	insignificante	-	-	-	-	-
1DX5	N	LYS	110	insignificante	-	-	-	-	-
1DX5	N	LYS	235	insignificante	-	-	-	-	-
1EBP	A	PHE	93	forte	-	-	-	-	✓
1EBP	A	MET	150	forte	-	✓	-	✓	✓
1EBP	A	THR	151	fraco	-	-	-	-	-
1EBP	A	PHE	205	forte	-	-	-	-	-
1EBP	C	GLY	9	intermediário	-	-	-	-	-
1EBP	C	PRO	10	intermediário	-	-	-	-	✗
1EBP	C	LEU	11	insignificante	-	-	-	-	✗
1EBP	C	THR	12	fraco	-	-	-	-	-
1EBP	C	TRP	13	forte	-	-	-	✓	✓
1ES7	A	VAL	26	intermediário	-	-	-	-	✗
1ES7	A	TRP	31	forte	✓	✓	-	✓	✓
1ES7	A	PHE	49	intermediário	✗	✗	✗	✗	✗
1ES7	A	PRO	50	intermediário	✗	✗	✗	✗	✗
1FAK	T	LYS	15	insignificante	-	-	-	-	-

PDB ID	Cadeia	Resíduo		Força	Rosetta	FOLDEF	KFC	Minerva	Nosso Classificador
1FAK	T	THR	17	insignificante	-	-	-	-	-
1FAK	T	ASN	18	insignificante	-	-	-	-	✗
1FAK	T	LYS	20	forte	-	-	-	✓	✓
1FAK	T	ILE	22	fraco	-	-	-	-	✗
1FAK	T	GLU	24	fraco	-	-	-	-	-
1FAK	T	GLN	37	fraco	-	-	✗	-	-
1FAK	T	LYS	41	insignificante	-	-	-	-	-
1FAK	T	SER	42	insignificante	-	-	-	-	-
1FAK	T	ASP	44	fraco	-	✗	-	-	✗
1FAK	T	SER	47	insignificante	-	-	-	-	-
1FAK	T	LYS	48	insignificante	-	-	✗	-	-
1FAK	T	PHE	50	insignificante	✗	✗	-	-	✗
1FAK	T	ASP	58	forte	-	-	-	✓	✓
1FAK	T	TYR	94	fraco	-	-	-	-	-
1FAK	T	GLU	128	insignificante	-	-	-	-	-
1FAK	T	LEU	133	insignificante	✗	✗	-	-	✗
1FAK	T	ARG	135	insignificante	-	-	-	-	✗
1FAK	T	PHE	140	intermediário	-	-	-	-	-
1FAK	T	THR	203	insignificante	-	-	✗	-	-
1FAK	T	VAL	207	insignificante	-	-	-	-	-
1G3I	A	ASP	438	forte	-	-	-	-	✓
1G3I	A	LEU	439	forte	-	-	-	-	✓
1G3I	A	ARG	441	forte	-	-	-	-	✓
1G3I	A	PHE	442	forte	-	-	-	✓	✓
1G3I	A	ILE	443	forte	-	✓	-	✓	✓
1G3I	A	LEU	444	forte	-	-	-	✓	✓
1IHB	B	ASN	101	insignificante	-	-	-	-	-
1IHB	B	ARG	133	fraco	✗	-	-	-	-
1IHB	B	HIS	135	fraco	-	-	-	-	✗
1IHB	B	LYS	136	intermediário	-	-	-	-	-
1JAT	A	GLU	55	forte	✓	-	-	-	-
1JAT	B	PHE	8	forte	✓	✓	✓	✓	✓
1MQ8	B	THR	206	forte	✓	-	✓	-	✓

PDB ID	Cadeia	Resíduo		Força	Rosetta	FOLDEF	KFC	Minerva	Nosso Classificador
1NFI	F	TYR	181	forte	✓	✓	-	✓	✓
1NFI	F	CYS	215	insignificante	-	-	-	-	✗
1NUN	A	ASP	76	intermediário	✗	-	✗	-	✗
1NUN	A	ARG	78	intermediário	✗	✗	-	-	✗
1NUN	A	ARG	155	intermediário	-	-	✗	-	-
2HHB	B	TYR	35	insignificante	-	-	-	-	-

Arkin, M. R., Wells, J. A., Abr. 2004. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discov 3 (4), 301–317.
URL http://dx.doi.org/10.1038/nrd1343 (Citado nas páginas 2 e 79.)

Bahadur, R., Zacharias, M., Abr. 2008. The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. Cellular and Molecular Life Sciences 65 (7), 1059–1072.
URL http://dx.doi.org/10.1007/s00018-007-7451-x (Citado na página 4.)

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., Sarai, A., Jan. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Research 32 (Database issue), D120–121, PMID: 14681373.
URL http://www.ncbi.nlm.nih.gov/pubmed/14681373 (Citado na página 11.)

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., Jan. 2000. The protein data bank. Nucleic Acids Research 28 (1), 235–242, PMID: 10592235.
URL http://www.ncbi.nlm.nih.gov/pubmed/10592235 (Citado na página 20.)

Bogan, A. A., Thorn, K. S., Jul. 1998. Anatomy of hot spots in protein interfaces. Journal of Molecular Biology 280 (1), 1–9.
URL http://www.sciencedirect.com/science/article/B6WK7-45S49GB-9C/2/b3d9c6f299c1eec3933d2774dffaf67d (Citado nas páginas 2, 3, 4, 5, 49 e 80.)

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. Em: Proceedings of the fifth annual workshop on Computational learning theory. COLT '92. ACM, New York, NY, USA, p. 144–152.
URL http://doi.acm.org/10.1145/130385.130401 (Citado na página 42.)

Butterfoss, G. L., Kuhlman, B., 2006. Computer-based design of novel protein structures. Annual Review of Biophysics and Biomolecular Structure 35, 49–65, PMID: 16689627.
URL http://www.ncbi.nlm.nih.gov/pubmed/16689627 (Citado na página 2.)

Chang, C.-c., Lin, C., 2001. LIBSVM: a library for support vector machines.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9020 (Citado nas páginas 42 e 43.)

Cho, K.-i., Kim, D., Lee, D., Maio 2009. A feature-based approach to modeling protein-protein interaction hot spots. Nucl. Acids Res. 37 (8), 2672–2687.
URL http://nar.oxfordjournals.org/cgi/content/abstract/37/8/2672 (Citado nas páginas 7, 15, 17, 20, 25, 50 e 59.)

Chong, L. T., Swope, W. C., Pitera, J. W., Pande, V. S., Mar. 2006. Kinetic computational alanine scanning: application to p53 oligomerization. Journal of Molecular Biology 357 (3), 1039–1049, PMID: 16457841.
URL http://www.ncbi.nlm.nih.gov/pubmed/16457841 (Citado na página 7.)

Clackson, T., Wells, J. A., Jan. 1995. A hot spot of binding energy in a hormone-receptor interface. Science (New York, N.Y.) 267 (5196), 383–386, PMID: 7529940.
URL http://www.ncbi.nlm.nih.gov/pubmed/7529940 (Citado na página 2.)

Conte, L. L., Chothia, C., Janin, J., Fev. 1999. The atomic structure of protein-protein recognition sites. Journal of Molecular Biology 285 (5), 2177–2198.
URL http://www.sciencedirect.com/science/article/B6WK7-45R884M-RY/2/6f4a9866e2495a34273695de046893dc (Citado nas páginas 1 e 57.)

Cusick, M. E., Klitgord, N., Vidal, M., Hill, D. E., Out. 2005. Interactome: gateway into systems biology. Human Molecular Genetics 14 Spec No. 2, R171–181, PMID: 16162640.
URL http://www.ncbi.nlm.nih.gov/pubmed/16162640 (Citado na página 1.)

Darnell, S. J., Page, D., Mitchell, J. C., 2007. An automated decision-tree approach to predicting protein interaction hot spots. Proteins: Structure, Function, and Bioinformatics 68 (4), 813–823.
URL http://dx.doi.org/10.1002/prot.21474 (Citado nas páginas 7, 14, 15 e 20.)

Das, R., Baker, D., 2008. Macromolecular modeling with rosetta. Annual Review of Biochemistry 77, 363–382, PMID: 18410248.
URL http://www.ncbi.nlm.nih.gov/pubmed/18410248 (Citado na página 10.)

DeLano, W. L., Fev. 2002. Unraveling hot spots in binding interfaces: progress and challenges. Current Opinion in Structural Biology

12 (1), 14–20, PMID: 11839484.
URL http://www.ncbi.nlm.nih.gov/pubmed/11839484 (Citado nas páginas 3 e 58.)

Deremble, C., Lavery, R., Abr. 2005. Macromolecular recognition. Current Opinion in Structural Biology 15 (2), 171–175.
URL http://www.sciencedirect.com/science/article/B6VS6-4FFX9JD-1/2/c033f8edd0769b3cd76165da0e9149cd (Citado na página 2.)

Diller, D. J., Humblet, C., Zhang, X., Westerhoff, L. M., 2010. Computational alanine scanning with linear scaling semi-empirical quantum mechanical methods. Proteins: Structure, Function, and Bioinformatics 9999 (999A), NA.
URL http://dx.doi.org/10.1002/prot.22745 (Citado na página 7.)

Fauchere, J. L., Pliska, V., 1983. Hydrophobic parameters ? of amino acid side chains from the partitioning of n-acetyl-amio-acid amides. J. Eur. Med. Chem., 369–375. (Citado na página 16.)

Fawcett, T., Jun. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.
URL http://www.sciencedirect.com/science/article/pii/S016786550500303X (Citado na página 46.)

Fischer, T. B., Arunachalam, K. V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., Huang, D., Paczkowski, M., Lalchandani, V., Ramachandra, C., Ellison, B., Galer, S., Shapley, J., Fuentes, E., Tsai, J., Jul. 2003. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics (Oxford, England) 19 (11), 1453–1454, PMID: 12874065.
URL http://www.ncbi.nlm.nih.gov/pubmed/12874065 (Citado nas páginas 22 e 72.)

Fromer, M., Shifman, J. M., Dez. 2009. Tradeoff between stability and multispecificity in the design of promiscuous proteins. PLoS Computational Biology 5 (12), e1000627, PMID: 20041208.
URL http://www.ncbi.nlm.nih.gov/pubmed/20041208 (Citado na página 2.)

Gao, Y., Wang, R., Lai, L., Fev. 2004. Structure-based method for analyzing protein–protein interfaces. Journal of Molecular Modeling 10 (1), 44–54.
URL http://dx.doi.org/10.1007/s00894-003-0168-3 (Citado na página 7.)

Guerois, R., Nielsen, J. E., Serrano, L., Jul. 2002. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology 320 (2),

369–387.
URL http://www.sciencedirect.com/science/article/B6WK7-463X8HC-P/2/c5991690487be77731a8bc0f31f4129d (Citado nas páginas 7 e 12.)

Guharoy, M., Chakrabarti, P., 2009. Empirical estimation of the energetic contribution of individual interface residues in structures of protein–protein complexes. Journal of Computer-Aided Molecular Design 23 (9), 645–654.
URL http://dx.doi.org/10.1007/s10822-009-9282-3 (Citado na página 25.)

Guyon, I., Elisseeff, A., Mar. 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
URL http://dl.acm.org/citation.cfm?id=944919.944968 (Citado na página 46.)

Haliloglu, T., Keskin, O., Ma, B., Nussinov, R., Mar. 2005. How similar are protein folding and protein binding nuclei? examination of vibrational motions of energy hot spots and conserved residues. Biophysical Journal 88 (3), 1552–1559.
URL http://www.cell.com/biophysj/abstract/S0006-3495(05)73222-7 (Citado na página 50.)

Han, J., Kamber, M., Pei, J., Jul. 2011. Data Mining: Concepts and Techniques, 3 EdiÃ§Ã£o. Morgan Kaufmann. (Citado na página 42.)

Honig, B., Nicholls, A., Maio 1995. Classical electrostatics in biology and chemistry. Science (New York, N.Y.) 268 (5214), 1144–1149, PMID: 7761829.
URL http://www.ncbi.nlm.nih.gov/pubmed/7761829 (Citado na página 30.)

Hsu, C., Chang, C., Lin, C., 2003. A practical guide to support vector classification. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University.
URL http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (Citado nas páginas 42 e 43.)

Hu, Z., Ma, B., Wolfson, H., Nussinov, R., 2000. Conservation of polar residues as hot spots at protein interfaces. Proteins: Structure, Function, and Genetics 39 (4), 331–342.
URL http://dx.doi.org/10.1002/(SICI)1097-0134(20000601)39:4<331::AID-PROT60>3.0.CO;2-A (Citado na página 25.)

Humphris, E. L., Kortemme, T., Ago. 2007. Design of multi-specificity in protein interfaces. PLoS Computational Biology 3 (8), e164, PMID: 17722975.
URL http://www.ncbi.nlm.nih.gov/pubmed/17722975 (Citado na página 2.)

Jamadagni, S. N., Godawat, R., Garde, S., 2011. Hydrophobicity of proteins and interfaces: insights from density fluctuations. Annual Review of Chemical and Biomolecular Engineering 2, 147–171, PMID: 22432614.
URL http://www.ncbi.nlm.nih.gov/pubmed/22432614 (Citado na página 80.)

Janin, J., Jan. 1999. Wet and dry interfaces: the role of solvent in protein–protein and protein–DNA recognition. Structure 7 (12), R277–R279.
URL http://www.cell.com/structure/abstract/S0969-2126(00)88333-1 (Citado na página 3.)

Joachims, T., 1999. Making large-scale support vector machine learning practical. Em: Advances in kernel methods: support vector learning. MIT Press, pp. 169–184.
URL http://portal.acm.org/citation.cfm?id=299104 (Citado na página 43.)

Karanicolas, J., Kuhlman, B., Ago. 2009. Computational design of affinity and specificity at protein-protein interfaces. Current Opinion in Structural Biology 19 (4), 458–463.
URL http://www.sciencedirect.com/science/article/B6VS6-4WWDCR0-2/2/66d1ea49662797d713913fee2768b849 (Citado na página 2.)

Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., Meiler, J., Abr. 2010. Practically useful: What the rosetta protein modeling suite can do for you. Biochemistry 49 (14), 2987–2998.
URL http://dx.doi.org/10.1021/bi902153g (Citado na página 10.)

Knuth, D. E., Dez. 1974. Computer programming as an art. Commun. ACM 17 (12), 667–673.
URL http://doi.acm.org/10.1145/361604.361612 (Citado na página vi.)

Kortemme, T., Baker, D., Out. 2002. A simple physical model for binding energy hot spots in protein–protein complexes. Proceedings of the National Academy of Sciences of the United States of America 99 (22), 14116–14121.
URL http://www.pnas.org/content/99/22/14116.abstract (Citado nas páginas 7, 10, 11, 12 e 20.)

Kortemme, T., Baker, D., Fev. 2004. Computational design of protein-protein interactions. Current Opinion in Chemical Biology 8 (1), 91–97.
URL http://www.sciencedirect.com/science/article/B6VRX-4BDY9F2-1/2/cf5af1f55ef5d888d6de19106907e16b (Citado nas páginas 7 e 20.)

Kortemme, T., Kim, D. E., Baker, D., Fev. 2004. Computational alanine scanning of Protein-Protein interfaces. Sci. STKE 2004 (219), pl2.
URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;2004/219/pl2 (Citado nas páginas 2 e 10.)

Kyte, J., Doolittle, R. F., Maio 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157 (1), 105–132.
URL http://www.sciencedirect.com/science/article/B6WK7-4DN8WFH-4J/2/40ae87fac04734a1b6db3acfd6b80652 (Citado nas páginas 31 e 51.)

Lazaridis, T., Karplus, M., Maio 1999. Effective energy function for proteins in solution. Proteins 35 (2), 133–152, PMID: 10223287.
URL http://www.ncbi.nlm.nih.gov/pubmed/10223287 (Citado na página 11.)

Li, X., Keskin, O., Ma, B., Nussinov, R., Liang, J., Nov. 2004. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. Journal of Molecular Biology 344 (3), 781–795, PMID: 15533445.
URL http://www.ncbi.nlm.nih.gov/pubmed/15533445 (Citado na página 50.)

Li, Y., Huang, Y., Swaminathan, C. P., Smith-Gill, S. J., Mariuzza, R. A., Fev. 2005. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. Structure (London, England: 1993) 13 (2), 297–307, PMID: 15698573.
URL http://www.ncbi.nlm.nih.gov/pubmed/15698573 (Citado na página 58.)

Li, Z., Li, J., Dez. 2010. Geometrically centered region: A "wet" model of protein binding hot spots not excluding water molecules. Proteins: Structure, Function, and Bioinformatics 78 (16), 3304–3316.
URL http://onlinelibrary.wiley.com/doi/10.1002/prot.22838/abstract (Citado nas páginas 38 e 39.)

Makhatadze, G. I., Privalov, P. L., Jul. 1993. Contribution of hydration to protein folding thermodynamics. i. the enthalpy of hydration. Journal of Molecular Biology 232 (2), 639–659, PMID: 8393940.
URL http://www.ncbi.nlm.nih.gov/pubmed/8393940 (Citado nas páginas 39 e 40.)

Massova, I., Kollman, P. A., 1999. Computational alanine scanning to probe Protein-Protein interactions: A novel approach to evaluate binding free energies. Journal of the American Chemical Society 121 (36), 8133–8143.
URL http://dx.doi.org/10.1021/ja990935j (Citado na página 7.)

Moreira, I. S., Fernandes, P. A., Ramos, M. J., 2007a. Computational alanine scanning mutagenesis - an improved methodological approach. Journal of Computational Chemistry 28 (3), 644–654.
URL http://dx.doi.org/10.1002/jcc.20566 (Citado na página 7.)

Moreira, I. S., Fernandes, P. A., Ramos, M. J., 2007b. Hot spots - a review of the protein-protein interface determinant amino-acid residues. Proteins: Structure, Function, and Bioinformatics 68 (4), 803–812.
URL http://dx.doi.org/10.1002/prot.21396 (Citado nas páginas 2, 4 e 58.)

Morrison, K. L., Weiss, G. A., Jun. 2001. Combinatorial alanine-scanning. Current Opinion in Chemical Biology 5 (3), 302–307, PMID: 11479122.
URL http://www.ncbi.nlm.nih.gov/pubmed/11479122 (Citado nas páginas 4 e 5.)

Neshich, G., Mazoni, I., Oliveira, S. R. M., Yamagishi, M. E. B., Kuser-Falcão, P. R., Borro, L. C., Morita, D. U., Souza, K. R. R., Almeida, G. V., Rodrigues, D. N., Jardine, J. G., Togawa, R. C., Mancini, A. L., Higa, R. H., Cruz, S. A. B., Vieira, F. D., Santos, E. H., Melo, R. C., Santoro, M. M., 2006. The star STING server: a multiplatform environment for protein structure analysis. Genetics and Molecular Research: GMR 5 (4), 717–722, PMID: 17183482.
URL http://www.ncbi.nlm.nih.gov/pubmed/17183482 (Citado na página 23.)

Neshich, G., Rocchia, W., Mancini, A. L., Yamagishi, M. E. B., Kuser, P. R., Fileto, R., Baudet, C., Pinto, I. P., Montagner, A. J., Palandrani, J. F., Krauchenco, J. N., Torres, R. C., Souza, S., Togawa, R. C., Higa, R. H., Jul. 2004. JavaProtein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. Nucleic Acids Research 32 (Web Server issue), W595–601, PMID: 15215458.
URL http://www.ncbi.nlm.nih.gov/pubmed/15215458 (Citado na página 30.)

Ofran, Y., Rost, B., Jul. 2007. Protein–Protein interaction hotspots carved into sequences. PLoS Comput Biol 3 (7), e119.
URL http://dx.plos.org/10.1371/journal.pcbi.0030119 (Citado nas páginas 5, 7, 8, 9 e 10.)

Oliveira, S. R. M., Almeida, G. V., Souza, K. R. R., Rodrigues, D. N., Kuser-Falcão, P. R., Yamagishi, M. E. B., Santos, E. H., Vieira, F. D., Jardine, J. G., Neshich, G., 2007. Sting_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. Genetics and Molecular Research: GMR 6 (4), 911–922, PMID: 18058712.

URL http://www.ncbi.nlm.nih.gov/pubmed/18058712 (Citado na página 23.)

Patel, A. J., Varilly, P., Chandler, D., Fev. 2010. Fluctuations of water near extended hydrophobic and hydrophilic surfaces. The journal of physical chemistry. B 114 (4), 1632–1637, PMID: 20058869 PMCID: PMC3173972. (Citado na página 80.)

Potapov, V., Reichmann, D., Abramovich, R., Filchtinski, D., Zohar, N., Ben Halevy, D., Edelman, M., Sobolev, V., Schreiber, G., Dez. 2008. Computational redesign of a protein-protein interface for high affinity and binding specificity using modular architecture and naturally occurring template fragments. Journal of Molecular Biology 384 (1), 109–119, PMID: 18804117.
URL http://www.ncbi.nlm.nih.gov/pubmed/18804117 (Citado na página 2.)

Radzicka, A., Wolfenden, R., Mar. 1988. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 27 (5), 1664–1670.
URL http://dx.doi.org/10.1021/bi00405a042 (Citado nas páginas 31 e 50.)

Rajamani, D., Thiel, S., Vajda, S., Camacho, C. J., 2004. Anchor residues in protein–protein interactions. Proceedings of the National Academy of Sciences of the United States of America 101 (31), 11287 –11292.
URL http://www.pnas.org/content/101/31/11287.abstract (Citado na página 50.)

Reichmann, D., Rahat, O., Cohen, M., Neuvirth, H., Schreiber, G., Fev. 2007. The molecular architecture of protein-protein binding sites. Current Opinion in Structural Biology 17 (1), 67–76.
URL http://www.sciencedirect.com/science/article/B6VS6-4MVDVF3-3/2/60d484544e9900423f219f319b0936b5 (Citado nas páginas 1 e 2.)

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., Honig, B., Jan. 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. Journal of Computational Chemistry 23 (1), 128–137, PMID: 11913378.
URL http://www.ncbi.nlm.nih.gov/pubmed/11913378 (Citado na página 30.)

Ruxton, G. D., 2006. The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney u test. Behavioral Ecology 17 (4), 688 –690.

URL http://beheco.oxfordjournals.org/content/17/4/688.short (Citado na página 41.)

Ryan, D. P., Matthews, J. M., Ago. 2005. Protein-protein interactions in human disease. Current Opinion in Structural Biology 15 (4), 441–446.
URL http://www.sciencedirect.com/science/article/B6VS6-4GHRBY5-1/2/05b6bb5d77b2a424a5c252bdcf684d4c (Citado na página 1.)

Sridharan, S., Nicholls, A., Honig, B., 1992. A new vertex algorithm to calculate solvent accessible surface areas. Biophysical Journal, A174. (Citado nas páginas 26 e 38.)

Suárez, M., Jaramillo, A., Ago. 2009. Challenges in the computational design of proteins. Journal of the Royal Society, Interface / the Royal Society 6 Suppl 4, S477–491, PMID: 19324680.
URL http://www.ncbi.nlm.nih.gov/pubmed/19324680 (Citado na página 2.)

Thorn, K. S., Bogan, A. A., Mar. 2001. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics (Oxford, England) 17 (3), 284–285, PMID: 11294795.
URL http://www.ncbi.nlm.nih.gov/pubmed/11294795 (Citado nas páginas 10, 20 e 80.)

Tsai, C., Ma, B., Nussinov, R., Dez. 2009. Protein-protein interaction networks: how can a hub protein bind so many different partners? Trends in Biochemical Sciences 34 (12), 594–600.
URL http://www.sciencedirect.com/science/article/B6TCV-4XG5NVG-1/2/2fc43c24021c4701a519b412f4fae68e (Citado na página 1.)

Tsodikov, O. V., Record Jr., M. T., Sergeev, Y. V., Abr. 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. Journal of Computational Chemistry 23 (6), 600–609.
URL http://onlinelibrary.wiley.com/doi/10.1002/jcc.10061/abstract (Citado na página 29.)

Tuncbag, N., Gursoy, A., Keskin, O., Jun. 2009. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics 25 (12), 1513–1520.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/12/1513 (Citado na página 7.)

Vriend, G., Mar. 1990. WHAT IF: a molecular modeling and drug design program. Journal of Molecular Graphics 8 (1), 52–56, 29,

PMID: 2268628.
URL http://www.ncbi.nlm.nih.gov/pubmed/2268628 (Citado na página 15.)

Witten, I., Frank, E., Jun. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0120884070 (Citado na página 44.)

Yeh, B. K., Igarashi, M., Eliseenkova, A. V., Plotnikov, A. N., Sher, I., Ron, D., Aaronson, S. A., Mohammadi, M., Mar. 2003. Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. Proceedings of the National Academy of Sciences 100 (5), 2266 –2271.
URL http://www.pnas.org/content/100/5/2266.abstract (Citado na página 72.)