Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Roberto Hiroshi Higa

Predição de regiões de interface proteína-proteína baseada em informações estruturais

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: Engenharia de Computação.

Orientador: Clésio Luis Tozzi

Campinas, SP
2009

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

H533p

Higa, Roberto Hiroshi
Predição de regiões de interface proteína-proteína baseada em informações estruturais / Roberto Hiroshi Higa. –Campinas, SP: [s.n.], 2009.

Orientador: Clésio Luis Tozzi.
Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Proteína. 2. Previsão. 3. Reconhecimento de padrões. I. Tozzi, Clésio Luis. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês:	Prediction of protein-protein interface region based on structural information
Palavras-chave em Inglês:	Protein, Prediction, Pattern recognition
Área de concentração:	Engenharia de Computação
Titulação:	Doutor em Engenharia Elétrica
Banca Examinadora:	Fernando José Von Zuben, José Eduardo Cogo Castanho, Márcio Luiz de Andrade Netto e Paula Regina Kuser Falcão
Data da defesa:	30/07/2009
Programa de Pós Graduação:	Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidato: Roberto Hiroshi Higa

Data da Defesa: 30 de julho de 2009

Título da Tese: "Predição de Regiões de Interface Proteína-Proteína Baseada em Informações Estruturais"

Prof. Dr. Clésio Luis Tozzi (Presidente): _____

Prof. Dr. José Eduardo Cogo Castanho: _____

Dra. Paula Regina Kuser Falcão: _____

Prof. Dr. Márcio Luiz de Andrade Netto: _____

Prof. Dr. Fernando José Von Zuben: _____

A verdadeira origem da descoberta consiste não em procurar novas paisagens, mas em ter novos olhos.
(Marcel Proust)

Resumo

Este trabalho aborda o problema de predição de regiões de interface proteína-proteína, baseado em medidas de propriedades físico-químicas e estruturais. A abordagem considera os aminoácidos da superfície da proteína como as unidades básicas para classificação, o que elimina a restrição de uso de *patches*, considerada por preditores do mesmo tipo. Este preditor é complementado por um segundo preditor, que identifica, dentre os aminoácidos da interface, os mais relevantes do ponto de vista de energia de ligação proteína-proteína, conhecidos como *hot spots*. A abordagem apresentada permite a utilização dos classificadores de forma independente na predição dos aminoácidos de interface e dos *hot spots*. Diferente de outras abordagens encontradas na literatura para predição de *hot spots*, a abordagem empregada não depende do conhecimento da estrutura do complexo protéico e permite que a predição de *hot spots* seja realizada em complemento à predição dos aminoácidos da interface. O desempenho alcançado pelo preditor na identificação de *hot spots* é superior ao obtido por outros preditores descritos na literatura e que utilizam o mesmo conjunto de dados e critério de avaliação de desempenho.

Palavras-chave: Proteína, Previsão, Reconhecimento de padrões.

Abstract

This work approaches the problem of protein-proteins interface region prediction based on measures of physical-chemical and structural properties. The approach considers the amino acids of the protein surface as the basic units for classification, such that the restriction of use of patches, considered by similar predictors, is eliminated. This predictor is complemented by a second one, which identifies, among the interface amino acids, those which are most relevant from the standpoint of protein-protein binding energy. The classifiers can be used independently, for predicting interface amino acids and hot spots. Unlike other approaches for prediction of hot spots, described in the literature, the proposed approach does not depend on the knowledge of the protein structure in complex. This allows predicting hot spots in complement to the prediction of the interface amino acids. Concerning the identification of hot spots, the proposed predictor outperformed those, described in the literature, which use the same data set and criterion for performance evaluation.

Keywords: Protein, Prediction, Pattern recognition.

Agradecimentos

Ao meu orientador, Prof. Dr. Clésio Luis Tozzi, sou grato pela orientação, principalmente, pela paciência em me ouvir e à disposição para se engajar em nossas longas discussões. Vou sentir falta disso!

A meus familiares, meus pais e irmãos, pelo apoio demonstrado ao longo desta jornada.

Em especial, à minha sogra, Jandira Aparecida Moura, pelo suporte com a casa e as crianças. Este trabalho seria inviável sem o seu apoio!

À Embrapa Informática Agropecuária, na pessoa do Chefe Geral Dr. Eduardo Delgado Assad, pelo apoio financeiro e incentivo.

Aos colegas de trabalho e de pós-graduação, pelas críticas e sugestões.

Ao Prof. Eleri Cardozo, por disponibilizar este modelo de tese.

À minha amada esposa, Maria Fernanda Moura, pelo amor sincero, apoio irrestrito e incentivo constante ao longo do desenvolvimento desta tese.

Às minhas filhas, Yasmin e Sayuri, simplesmente, por existirem. Vocês são a minha vida.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

ASP - *Atomic solvation parameter*

CN - *Coordination number*

DNA - *Deoxyribonucleic acid*

HSE - *Half sphere exposure*

HSEd - *Half sphere exposure down*

HSEu - *Half sphere exposure upper*

mmCIF - *MacroMolecular Crystllograhic InFormation*

MS - *Molecular surface*

PDB - *Protein Data Bank*

PDBML - *PDB Markup Language*

RD - *Residue depth*

RMN - Ressonância Nuclear Magnética

SAS - *Solvent accessible surface*

VWS - *van der Waals surface*

XML - *eXtensible Markup Language*

Lista de Símbolos

m, n, i e j	-	Constantes $\in \Re$.
x, y e z	-	Variáveis $\in \Re$.
$\mathbf{x}$	-	Vetor $\mathbf{x}$.
$\|\mathbf{x}\|$	-	Norma do vetor $\mathbf{x}$.
$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{mn} \end{pmatrix}$	-	Matriz $\mathbf{X}$ de dimensão $n \times m$.
$\mathbf{X}^t$	-	Matriz transposta de $\mathbf{X}$.
$\mathbf{X}^{-1}$	-	Matriz inversa de $\mathbf{X}$.
$p(\mathbf{x})$	-	Função densidade de probabilidade de $\mathbf{x}$.
$P(\mathbf{x})$	-	Função distribuição de probabilidade de $\mathbf{x}$.
$p(\mathbf{x}, \mathbf{y})$	-	Função densidade de probabilidade conjunta de $\mathbf{x}$ e $\mathbf{y}$.
$p(\mathbf{x}\|\mathbf{y})$	-	Função densidade de probabilidade de $\mathbf{x}$ condicional a $\mathbf{y}$.
$N_{\mathrm{d}}(\mu, \mathbf{\Sigma})$	-	Função densidade de probabilidades normal d-dimensional com vetor de médias μ e matriz de covariâncias $\mathbf{\Sigma}$.
$var(x)$	-	Variância da variável aleatória x.
$cov(x, y)$	-	Covariância entre as variáveis aleatórias x e y.
$\mathcal{C}$	-	Conjunto $\mathcal{C}$.
$\|\mathcal{C}\|$	-	Cardinalidade do conjunto $\mathcal{C}$.
Π_i	-	i-ésima classe de objetos.
$J(.)$	-	Função J.
$\arg\max_{\mathbf{w}} J(\mathbf{w})$	-	Valor de $\mathbf{w}$ que maximiza a função $J(\mathbf{w})$.
$\dot{\gamma}(t)$	-	Derivada temporal de $\gamma(t)$ em t.
$\frac{\partial}{\partial x}$	-	Operador de deferenciação parcial com relação a x.
$\Delta = \nabla^2 = \nabla.\nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$	-	Operador de Laplace, definido como o divergente do gradiente.
$\Delta\Delta$G	-	Diferença de energia livre.
Å	-	Unidade de medida de distância. Um Å (Ångstron) é igual a 10^{-10} m.
d	-	Unidade de medida de massa. Um Dalton (1 d) é aproximadamente igual à massa de um átomo de hidrogênio ($1.7\text{x}10^{-24}$ g) e 1 kd (kilodalton) é igual a 10^3 daltons.

Notação

Termos em idiomas diferentes do português para os quais não haja uma tradução consagrada são escritos em seu idioma original. Eles aparecem no texto em estilo *italico* (ex: *a priori*, *hot spots*, *curvednesss*, etc.).

Publicações

1. R. H. Higa, C. L. Tozzi. “Prediction of binding Hot Spot residues by using structural and evolutionary parameters”. *Genetics and Molecular Biology* Vol. 32 No. 3, pg. 626-633, 2009.

2. R. H. Higa, C. L. Tozzi. “A simple and efficient method for predicting protein-protein interaction sites”. *Genetics and Molecular Research* Vol. 7 No. 3, pg. 898-909, 2008.

3. R. H. Higa, C. L. Tozzi. “Prediction of Protein-Protein Binding Hot Spots: A Combination of Classifiers Approach”. *Brazilian Symposium on Bioinformatics 2008* (Lecture Notes in Computer Science 5167), pg. 165-168.

Capítulo 1

Introdução

Esta tese aborda o problema de predição de regiões de interface proteína-proteína, utilizando informações estruturais de proteínas. Neste capítulo, é feita uma breve introdução ao tema interação proteína-proteína, sua relevância biológica e a utilização de preditores de regiões de interface no estudo de interações proteína-proteína. Em seguida, são apresentadas as motivações, os objetivos e o trabalho desenvolvido e, por fim, a organização do restante do trabalho.

1.1 Proteínas, interação proteína-proteína e preditores de regiões de interface

As proteínas são macromoléculas presentes nas células e que desempenham funções de fundamental importância para a sustentação do processo que denominamos vida. Elas são formadas pelo encadeamento de aminoácidos, que podem ser de vinte tipos diferentes. Cada proteína é caracterizada por uma seqüência específica de aminoácidos, conhecida como estrutura primária. O nível seguinte de organização estrutural de proteínas refere-se a conformações geométricas locais assumida pela seqüência de aminoácidos e é conhecida como estrutura secundária. O empacotamento dessas estruturas secundárias, então, resulta na conformação tridimensional característica da molécula de proteína, denominada estrutura terciária. Essa conformação lhe confere propriedades físico-químicas que determinam a especificidade e a afinidade com que ela interage com outras moléculas.

Para realizar suas funções, a proteína depende de interações com outras moléculas e, em particular, com outras moléculas de proteína, com as quais forma complexos, que podem ter uma duração transiente ou permanente. Exemplos incluem vários complexos formados em processos de reconhecimento de sinais celulares e catálise e/ou inibição de enzimas. A região de interface proteína-proteína é a região da superfície da proteína através da qual ela interage fisicamente com outras proteínas. A

identificação das interações em que uma proteína está envolvida e de sua correspondente região de interface representa um importante passo para a sua caracterização funcional, pois elas proporcionam uma melhor compreensão de seu mecanismo de interação com outras proteínas e de sua inserção em redes de interação proteína-proteína.

Importantes aplicações biotecnológicas na medicina e na agricultura baseiam-se na manipulação de processos biológicos em nível molecular. Por exemplo, um processo biológico específico pode ser regulado ou até mesmo bloqueado através de uma droga (pequena molécula) que tenha como alvo a região de interface de uma das proteínas envolvidas no processo. Na medicina, isso poderia significar o bloqueio do processo de formação de um complexo essencial para a replicação de um vírus, tornando-o não infeccioso. Na agricultura, se a proteína alvo é específica do processo de digestão de um organismo considerado como praga, ter-se-ia uma forma de combate a pragas ambientalmente segura, uma vez que a droga afetaria especificamente o organismo considerado como praga, matando-o por inanição.

A caracterização dos diferentes aspectos funcionais de uma proteína, em nível celular ou de organismo, envolve a utilização de métodos experimentais específicos. Em particular, informações sobre a estrutura de proteínas em nível atômico são obtidas através dos métodos de cristalografia por raios X e de ressonância magnética nuclear. Contudo, estes métodos consomem grandes quantidades de recursos[1] e estão sujeitos a restrições técnicas que, por vezes, limitam sua aplicação. Os métodos de cristalografia por raios X dependem da obtenção de cristais de proteínas com boa qualidade de difração, o que muitas vezes se mostra uma tarefa extremamente desafiadora, enquanto o método de ressonância magnética nuclear é limitado pelo tamanho da proteína em estudo. Estas dificuldades são ainda maiores para estruturas formadas por mais de uma cadeia de aminoácidos, os complexos protéicos. Isto se reflete no número muito menor de estruturas de complexos de proteínas, resolvidas experimentalmente e depositadas no *Protein Data Bank*[2] (PDB) (Berman et al., 2000), comparado ao número de estruturas de proteínas formadas por apenas uma cadeia de aminoácido.

Neste contexto, métodos computacionais de predição capazes de auxiliar o processo de determinação de regiões de interface para uma proteína, quando dados estruturais do complexo formado com sua proteína parceira não estão disponíveis, são de grande valia. Um preditor deste tipo pode ser útil em diversas situações, por exemplo:

- Muitas das proteínas estudadas no escopo dos projetos de genoma estrutural tiveram suas estruturas determinadas experimentalmente, mas sem uma caracterização funcional. Nessa situação,

[1] O custo médio para a solução da estrutura de uma proteína é da ordem centenas de milhares de dólares (Chandonia e Brenner, 2006)

[2] O PDB é o banco de dados mundial, de acesso público, onde são depositadas as estruturas de proteínas determinadas experimentalmente.

preditores de regiões funcionais, tais como as de interface com outras proteínas, podem ser úteis na elaboração de hipóteses sobre a função da proteína;

- Em estudos experimentais de mutagênese sítio-dirigido, a seqüência do gene que produz a proteína de interesse é modificada de forma que uma posição específica da proteína tenha seu aminoácido alterado. A importância funcional do aminoácido modificado pode ser aferida avaliando-se o nível de atividade da proteína mutante em relação ao da proteína original. Supondo que a estrutura tridimensional para a proteína de interesse seja conhecida, determinada experimentalmente ou modelada computacionalmente, um preditor de regiões de interface de proteínas pode sugerir um conjunto de aminoácidos a serem priorizados em análises experimentais;

- *Molecular docking* proteína-proteína refere-se à construção de modelos computacionais de complexos de proteínas. Basicamente, os algoritmos de *molecular docking* executam um processo de busca para encontrar a orientação espacial preferencial de uma molécula com relação a uma segunda, de acordo com uma função de mérito pré-estabelecida. Um preditor de regiões de interface proteína-proteína pode fornecer informações sobre regiões preferenciais de interação na superfície das proteínas, que podem ser usadas por algoritmos de *molecular docking* para guiar o processo de busca das orientações ótimas.

Preditores de regiões de interface proteína-proteína baseados em métodos computacionais podem ser considerados como classificadores. Eles têm por objetivo predizer os aminoácidos que pertencem à região de interface, quando os dados sobre a proteína com a qual ela interage não estão disponíveis, e podem ser baseados no conhecimento da seqüência da proteína ou de sua estrutura tridimensional.

Estes preditores baseiam-se no fato de que regiões de interface possuem características específicas que as diferenciam do restante da superfície da proteína, o que é corroborado por diferentes resultados experimentais. Por exemplo, DeLano et al. (2000) construíram um conjunto de peptídeos com seqüências aleatórias e mostraram que aqueles que se ligavam à proteína imunoglobina G de humanos, o faziam de forma consistente na mesma região da superfície da proteína. Já Lim et al. (2001) mostraram que dois inibidores diferentes de proteínas da família β-lactamase se ligavam exatamente na mesma região da superfície de uma das proteínas dessa família. Em adição, resultados experimentais também mostraram que as regiões de interface possuem propriedades mensuráveis e que apresentam diferenças quando comparadas com as correspondentes medidas sobre os demais aminoácidos da superfície da proteína. Algumas das diferenças encontradas entre os aminoácidos pertencentes à região de interface e os demais aminoácidos da superfície da proteína incluem os tipos preferenciais de aminoácidos, o grau de conservação evolucionário, a geometria da superfície e propriedades físico-químicas como a hidrofobicidade. Dados estruturais de complexos de proteínas

são estudados há pelo menos 30 anos, visando descrever as propriedades estruturais características da região de interface proteína-proteína (Chothia e Janin, 1975; Jones e Thornton, 1997a; Lo Conte et al., 1999; Chakrabarti e Janin, 2002; Bahandur et al., 2003; Ofran e Rost, 2003a; De et al., 2005). Embora muito se tenha avançado, esses estudos não foram capazes de identificar uma propriedade única capaz de caracterizar a região de interface de maneira inequívoca, tal que Jones e Thornton (1997a) sugeriram a utilização de uma combinação de propriedades físico-químicas e estruturais para se obter uma melhor caracterização.

Diversos preditores de regiões de interface proteína-proteína foram propostos na literatura, baseados no conhecimento da seqüência de aminoácidos ou da estrutura tridimensional da proteína. Ofran e Rost (2003b) e Yan et al. (2004) utilizam uma codificação para os 20 tipos de aminoácidos que compõem as proteínas e constroem o vetor de características para cada aminoácido da estrutura primária da proteína concatenando o seu código com os dos n aminoácidos precedentes e dos n subseqüentes, com n igual a 4. Reš et al. (2005) e Ofran e Rost (2006) utilizam um conjunto de seqüências de proteínas homólogas [3] à proteína de interesse, a partir do qual um alinhamento múltiplo é construído e a freqüência relativa de ocorrência de cada um dos 20 tipos de aminoácidos padrão em cada posição do alinhamento é extraída (perfil evolucionário). O vetor de características para cada aminoácido da seqüência é formado pela concatenação do seu perfil evolucionário com o dos n aminoácidos precedentes e n subseqüentes. Já Zhou e Shan (2001), Fariselli et al. (2002), Koike e Takagi (2004), Chen e Zhou (2005), Bordner e Abagyan (2005), Wang et al. (2006), Chung et al. (2006) e Li et al. (2007) consideram os aminoácidos da superfície da estrutura tridimensional da proteína e constroem o vetor de características para cada aminoácido na superfície da proteína concatenando o perfil evolucionário para o aminoácido e os m aminoácidos de superfície mais próximos, onde m geralmente é igual a 10. Finalmente, Jones e Thornton (1997b), Neuvirth et al. (2004), Bradford e Westhead (2005) e Bradford et al. (2006) consideram diferentes propriedades físico-químicas e estruturais para construir o vetor de características. Essas propriedades são medidas sobre uma região contínua de forma circular, formada por um conjunto de aminoácidos amostrados sobre a superfície da proteína, denominada *patch*. Nesses trabalhos, os *patches* são considerados como os objetos a serem classificados e os aminoácidos pertencentes àqueles identificados como interface são igualmente considerados como pertencentes à região de interface, apesar de terem sido amostrados de forma arbitrária.

1.2 Motivação, objetivo e trabalho desenvolvido

Conforme mencionado na seção 1.1, preditores que se baseiam no conhecimento da estrutura tridimensional da proteína (Jones e Thornton, 1997b; Neuvirth et al., 2004; Bradford e Westhead, 2005;

[3] Duas proteínas são consideradas homólogas se elas são evolutivamente relacionadas.

Bradford et al., 2006) realizam medidas de propriedades físico-químicas e estruturais sobre *patches* (um conjunto de aminoácidos), amostrados sobre a superfície da proteína. Assim, um aminoácido do conjunto que constitui um *patch* é igualmente considerado como pertencente ou não à região de interface, independente de sua importância para o processo de interação. Da mesma forma que em todos os preditores mencionados na seção 1.1, a interface é tratada como uma região com propriedades uniformes.

Contudo, diferentes estudos têm caracterizado a região de interface como sendo composta por regiões com propriedades distintas. Por exemplo, Lo Conte et al. (1999), Chakrabarti e Janin (2002), Bahandur et al. (2003) e De et al. (2005) analisaram os aminoácidos da região de interface do ponto de vista estrutural, considerando a perda de área acessível a solvente [4] na formação do complexo. Eles propuseram um modelo onde a região de interface é formada por: (a) um núcleo contendo aminoácidos que são acessíveis a solvente quando a proteína está isolada, mas que quando em complexo se tornam completamente obstruídos; e (b) um anel periférico contendo aminoácidos que permanecem acessíveis a solvente, mesmo após a formação do complexo. Eles também relataram que o núcleo apresenta uma preferência por aminoácidos com características mais hidrofóbicas, enquanto que o anel periférico, de forma similar aos aminoácidos encontrados no restante da superfície da proteína, apresenta uma preferência por aminoácidos com características mais hidrofílicas.

Com base nos estudos que indicam que a interface é uma região com propriedades não homogêneas e a indicação de Bradford e Westhead (2005) quanto à necessidade de relaxamento da restrição relativa à forma do *patch*, nosso estudo teve como objetivo o desenvolvimento de um classificador em que os aminoácidos da superfície da molécula de proteína são considerados como as unidades básicas para classificação. Para construir o vetor de características foram utilizadas medidas de propriedades físico-químicas e estruturais, cuja finalidade é caracterizar o aminoácido e a vizinhança em que ele está inserido. Um classificador em dois estágios foi desenvolvido. No primeiro, é utilizado um classificador linear com opção de rejeição. Os aminoácidos cuja classificação pode ser realizada com alta confiança são imediatamente classificados, enquanto que os demais têm sua classificação postergada para o estágio seguinte (sua classificação é rejeitada neste estágio). No segundo estágio, um procedimento empírico que considera a informação de vizinhança (dependência de contexto) é utilizado para decidir a classe dos aminoácidos cuja classificação tenha sido postergada no estágio anterior.

Os resultados obtidos ao testar o classificador desenvolvido mostram que ele é capaz de identificar os aminoácidos pertencentes à região de interface sem a utilização do conceito de *patches*; e que, portanto, atende ao objetivo do projeto inicialmente proposto. Encorajados por estes resultados e pela disponibilidade tanto das ferramentas desenvolvidas quanto de um conjunto de dados previa-

[4] A área acessível a solvente é uma forma de se medir a área da molécula de proteína que está exposta em sua superfície.

mente compilado (Darnell et al., 2007), o objetivo inicial do projeto foi estendido de modo a também considerar um refinamento na caracterização dos aminoácidos pertencentes à região de interface. Neste sentido, foi considerado o fato, evidenciado por resultados de experimentos de mutação sítio-dirigido, de que a energia de ligação não é igualmente distribuída entre os aminoácidos que compõem a região de interface, com uma grande fração atribuída a um pequeno conjunto de aminoácidos, denominados *hot spots* (Clackson e Wells, 1995; Bogan e Thorn, 1998). Assim, o preditor previamente desenvolvido foi estendido com um segundo classificador, de forma que os aminoácidos por ele identificados como pertencentes à região de interface também pudessem ter sua contribuição energética caracterizada como *hot spots* ou não *hot spots* (Figura 1.1). Para implementar o preditor de *hot spots*, é utilizado um classificador SVM (*Support Vector Machine*) (Cristianini e Shawe-Taylor, 2000) com a saída calibrada para representar a probabilidade a *posteriori*, de acordo com o método proposto por Platt (2000).

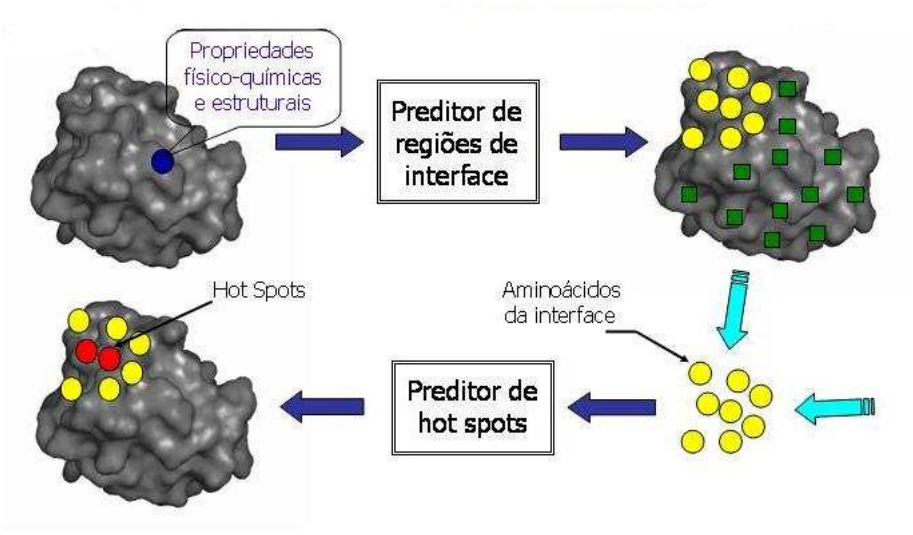

Figura 1.1: Utilização conjunta dos preditores de regiões de interface e de *hot spots*.

Finalmente, cabe ressaltar que na extensa revisão bibliográfica realizada pelo autor não foi identificado nenhum estudo que considere a não uniformidade da região de interface no desenvolvimento de preditores de regiões de interface proteína-proteína e que identifique os aminoácidos mais importantes da região de interface (*hot spots*). De fato, esses problemas, predição de regiões de interface e predição de *hot spots*, são abordados na literatura como dois problemas distintos.

1.3 Organização do trabalho

No capítulo 2, é feita uma introdução aos principais conceitos sobre estrutura de proteínas, seguida pela descrição do conjunto de propriedades físico-químicas e estruturais, a partir dos quais são extraídos os descritores utilizados para caracterizar os aminoácidos da superfície da molécula da proteína.

No capítulo 3, é apresentado um conjunto de técnicas utilizadas no desenvolvimento de classificadores de padrões, incluindo métodos de classificação, seleção e extração de características e critérios para avaliação de desempenho de classificadores. Sua função é a de prover a fundamentação metodológica na construção dos preditores descritos nos capítulos 4 e 5. A apresentação é feita de forma genérica, sem referência ao tipo de objeto sendo tratado.

No capítulo 4, é apresentado o classificador desenvolvido para a predição de regiões de interface proteína-proteína, baseado em medidas de propriedades físico-químicas e estruturais, com os aminoácidos da superfície da molécula de proteína considerados como as unidades básicas para classificação. Também são apresentados a metodologia utilizada em seu desenvolvimento, baseada nos capítulos 2 e 3, os resultados dos experimentos realizados para avaliação de desempenho e dois casos de uso.

No capítulo 5, é apresentado o classificador desenvolvido para predição de *hot spots* dentre os aminoácidos da região de interface, com base em medidas de propriedades físico-químicas e estruturais. Também são apresentados a metodologia utilizada em seu desenvolvimento, baseada nos capítulos 2 e 3, os resultados dos experimentos realizados para avaliação de desempenho e dois casos de uso.

Finalmente, no capítulo 6 são apresentadas as conclusões e propostas para trabalhos futuros.

Capítulo 2

Caracterização de aminoácidos expostos na superfície de moléculas de proteína

O objetivo deste capítulo é apresentar o conjunto de propriedades físico-químicas e estruturais, medidas sobre a estrutura tridimensional da molécula de proteína e utilizados para caracterizar os aminoácidos de sua superfície. A idéia é que os preditores apresentados nos capítulos 4 e 5 identifiquem os aminoácidos de interface e os *hot spots* utilizando essas propriedades para a formação de seus vetores de características.

Na parte inicial do capítulo são apresentados conceitos básicos sobre estrutura de proteínas. O texto apresentado é baseado em livros textos das áreas de biologia celular (Alberts et al., 1994; Alberts et al., 1998), bioquímica (Nelson e Cox, 2000) e estrutura de proteínas (Branden e Tooze, 1999). Também é feita uma breve introdução ao banco de dados *Protein Data Bank* (PDB) (Berman et al., 2000), o repositório mundial para estruturas tridimensionais de proteínas determinadas experimentalmente, a partir do qual são extraídos os conjuntos de dados utilizados no desenvolvimento dos preditores de regiões de interface e de *hot spots*. Na seqüência, são apresentadas as propriedades físico-químicas e estruturais consideradas neste trabalho. Os leitores familiarizados com os conceitos utilizados no estudo de estruturas de proteínas podem iniciar a leitura deste capítulo pela seção 2.3, sem prejuízos à sua compreensão.

2.1 Conceitos básicos sobre estrutura de proteínas

A estrutura de uma molécula de proteína pode ser descrita utilizando uma hierarquia de quatro níveis: primária, secundária, terciária e quaternária. Um tipo especial de ligação química covalente, denominada ligação peptídica, é utilizada na construção da estrutura primária, enquanto as demais estruturas da hierarquia são estabilizadas por um conjunto de ligações químicas, denominadas ligações

fracas.

2.1.1 Ligações químicas covalentes e não covalentes

Uma molécula, orgânica ou inorgânica, é formada por átomos que são unidos por ligações, denominadas covalentes ou fortes, formada por dois átomos que compartilham elétrons de sua órbita mais externa (Figura 2.1(a)). Este tipo de ligação química é bastante estável e se subdivide em ligações simples, duplas ou triplas, dependendo do número de elétrons compartilhados. Além disso, elas também podem apresentar um caráter polar quando os elétrons compartilhados são atraídos com diferentes intensidades pelos átomos que formam a ligação. Neste caso, a ligação é denominada ligação covalente polar. Um exemplo de molécula que exibe este tipo de ligação é a molécula de água, formada por ligações covalentes que unem um átomo de oxigênio (O) e dois átomos de hidrogênio (H) (Figura 2.1(b)).

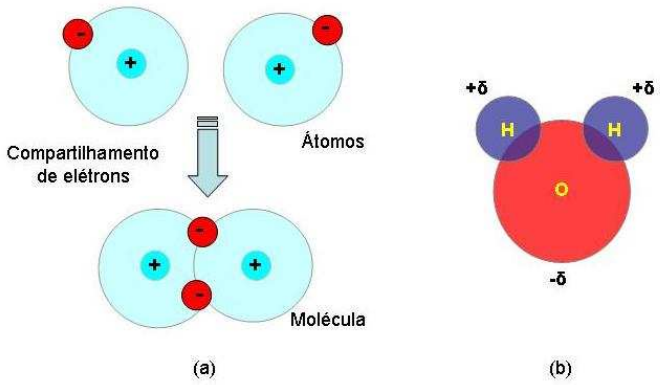

Figura 2.1: (a) Ligação covalente. (b) Exemplo de ligação covalente polar: molécula de água.

Existe ainda um segundo grupo de interações químicas que tem grande importância na análise de interações entre grupos atômicos de proteínas. Elas são denominadas ligações não covalentes ou ligações fracas e se subdividem em três[1] tipos (Alberts et al., 1994; Alberts et al., 1998; Nelson e Cox, 2000):

- **Ligações iônicas**: tipicamente, este tipo de ligação ocorre entre átomos que possuem um déficit ou um superávit de um ou dois elétrons em sua órbita mais externa, tal que elas podem ser preenchidas mais facilmente doando ou recebendo elétrons. Ao doar ou receber elétrons, dois átomos formam íons e passam a se atrair eletrostaticamente. Por exemplo, a Figura 2.2(a) ilustra a ligação que forma a molécula de NaCl. Um átomo de Na doa um elétron de sua órbita

[1] Por vezes, a interação hidrofóbica é considerada como um quarto tipo de interação fraca.

mais externa para o átomo de Cl, formando os íons Na+ e Cl-. Ambos os átomos passam a ter uma carga elétrica e se atraem eletrostaticamente.

- **Pontes de hidrogênio**: este tipo de ligação ocorre em função dos dipolos formados por ligações covalentes polares. Tipicamente, uma ponte de hidrogênio se forma quando um átomo de hidrogênio (H) de uma molécula, carregado positivamente, se encontra próximo de um átomo negativamente carregado, em geral oxigênio (O) ou nitrogênio (N). Novamente, utilizando a molécula de água como exemplo, a Figura 2.2(b) ilustra como elas se ligam através de pontes de hidrogênio.

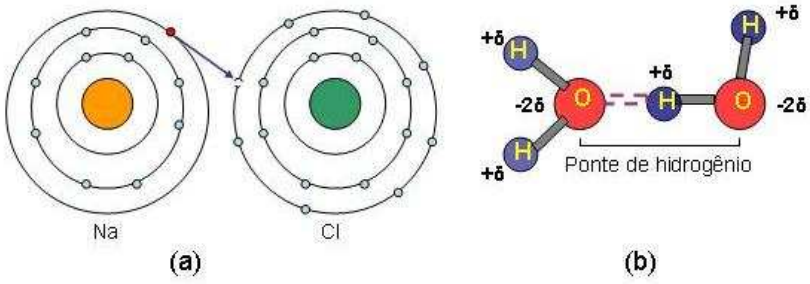

Figura 2.2: (a) Ligação iônica. (b) Ponte de hidrogênio.

- **Interação de van der Walls**: este tipo de interação acontece quando dois átomos não carregados estão muito próximos. Variações aleatórias nas posições dos elétrons em torno de um núcleo podem criar um dipolo transitório que, por sua vez, induz um dipolo de carga elétrica oposta em um átomo próximo. Os dois átomos são eletrostaticamente atraídos até uma distância mínima (distância de van der Walls), quando as suas nuvens eletrônicas se sobrepõem e a interação passa a ser de repulsão.

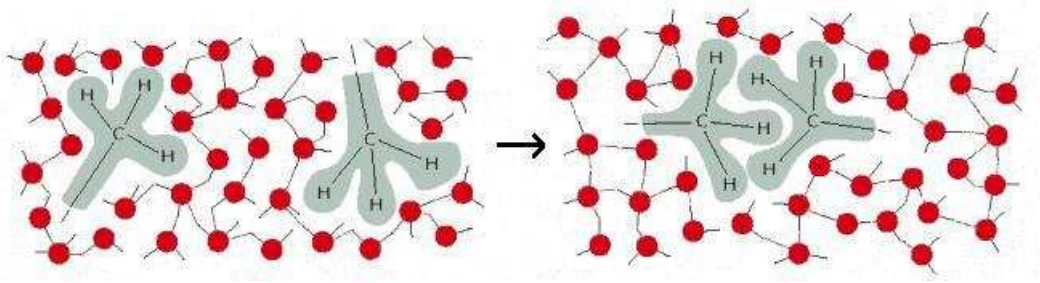

Figura 2.3: Interação hidrofóbica. Adaptado de Alberts et al. (1998).

- **Interação hidrofóbica**: além disso, em um meio aquoso como o interior da célula, grupos atômicos hidrofóbicos ou não polares apresentam uma tendência de se aglutinarem, de forma a

minimizar sua superfície de contato com a água. A força que mantém esse agrupamento não polar, por vezes, é considerada uma quarta força de ligação fraca, denominada interação hidrofóbica (Alberts et al., 1994; Alberts et al., 1998; Nelson e Cox, 2000). A Figura 2.3 ilustra esse fenômeno, considerando uma molécula rica em átomos de carbono (não polares) em um meio aquoso.

2.1.2 Aminoácidos, ligação peptídica e estrutura primária

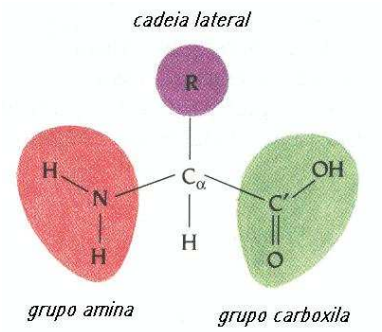

Figura 2.4: Composição química de aminoácidos. Adaptado de Branden e Tooze (1999).

Os aminoácidos são moléculas orgânicas e, conforme ilustrado pela Figura 2.4, são compostas por um grupo amina (NH_2), um grupo carboxila (COOH) e um grupo R, todos ligados a um carbono central, denominado carbono α (Cα). Vinte diferentes tipos de aminoácidos são utilizados na síntese de proteínas, sendo cada tipo caracterizado por um grupo R distinto, que lhe confere propriedades químicas específicas. Por serem observados nas proteínas de todos os organismos vivos, esses 20 aminoácidos são denominados aminoácidos padrões. Apesar de algumas proteínas também apresentarem formas modificadas de aminoácidos, estes resultam de modificações que ocorrem após a sua síntese.

As proteínas, por sua vez, são formadas por uma cadeia de aminoácidos unidos por ligações covalentes, denominadas ligações peptídicas. Essas cadeias de aminoácidos são denominadas cadeias polipeptídicas e seu processo de formação é denominado polimerização de aminoácidos. Durante o processo de síntese de uma proteína, o grupo amina de um aminoácido é ligado de forma covalente ao grupo carboxila do aminoácido seguinte, com a eliminação de uma molécula de água, numa reação denominada reação de condensação (Figura 2.5)

Esse processo se repete à medida que a cadeia polipeptídica é alongada, tal que as terminações NH_2 do primeiro aminoácido e COOH do último permanecem inalteradas. Diz-se, então, que a cadeia polipeptídica possui um sentido de orientação da terminação amina (NH_2) para a terminação carboxila

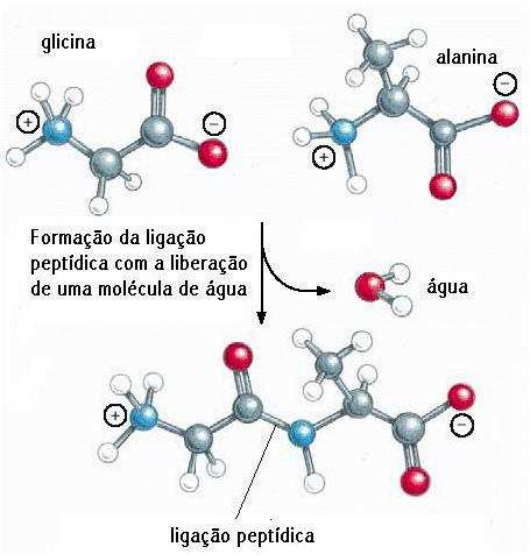

Figura 2.5: Reação de condensação. Adaptado de Alberts et al. (1994).

(COOH). Por convenção, uma cadeia polipeptídica é descrita pela seqüência de amínoácidos que a forma, com a terminação amina à esquerda e a terminação caboxila à direita (Figura 2.6).

A seqüência de átomos que se repete ao longo da cadeia de aminoácidos de uma proteína (... - NH-Cα-CO - ...) é denominada cadeia principal enquanto que os grupos R, ligados aos carbonos α, constituem a cadeia lateral.

Devido ao fato da reação de condensação eliminar uma hidroxila (OH) e um hidrogênio (H) dos aminoácidos envolvidos, as proteínas são, de fato, formadas por resíduos de aminoácidos, razão pela qual esses aminoácidos também são referenciados como resíduos (Nelson e Cox, 2000).

Figura 2.6: Exemplo de cadeia polipeptídica representada pelo código de uma letra (vide anexo B). Cadeia polipeptídica A do arquivo PDB 1gzx (hemoglobina).

2.1.3 Ângulos diedrais e estrutura secundária

Do ponto de vista bioquímico, o resíduo de aminoácido é a unidade básica que é repetidamente utilizada para construção de uma proteína (vide Figura 2.6). Contudo, uma outra forma de dividir a cadeia polipeptídica em unidades básicas, mais apropriadas para descrever propriedades estruturais,

é considerar os átomos compreendidos entre um carbono α e o próximo carbono α (Figura 2.7). Essa unidade é denominada unidade básica estrutural (Branden e Tooze, 1999).

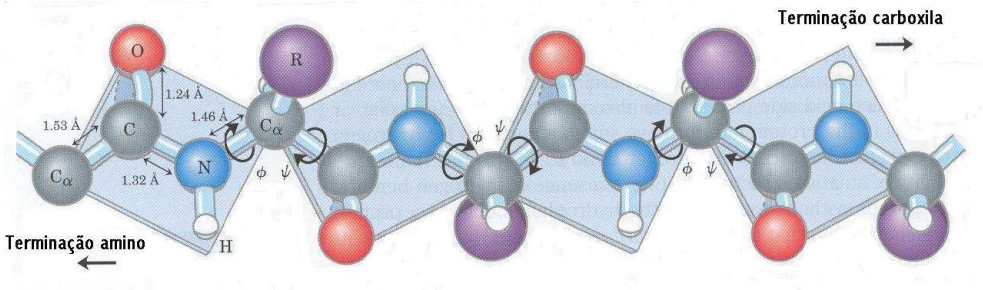

Figura 2.7: Unidade básica estrutural, constituída pelos átomos dos aminoácidos que estão entre dois carbonos α. Adaptado de Nelson e Cox (2000).

Experimentos realizados na década de 50 demonstraram que as ligações peptídicas têm um caráter de ligação covalente dupla, tal que todos os átomos associados com a ligação peptídica (ou em uma mesma unidade básica estrutural) estão em um mesmo plano com praticamente as mesmas distâncias e ângulos relativos em todas as unidades da cadeia polipeptídica (Nelson e Cox, 2000). Assim as unidades estruturais básicas formam grupos de átomos dispostos de forma rígida em um plano e são interligados por ligações covalentes envolvendo os carbonos α. Os dois únicos graus de liberdade que essas unidades possuem são as rotações das ligações N-Cα e Cα-C em torno do carbono α. O ângulo diedral correspondente à rotação em torno da ligação N-Cα é denominado phi (ϕ), enquanto o ângulo correspondente à rotação em torno da ligação Cα-C é denominado psi (ψ) (vide Figura 2.7). Assim, a cada aminoácido da cadeia polipeptídica está associado um par de ângulos ϕ e ψ, sendo as únicas exceções o primeiro aminoácido, que possui apenas o ângulo ψ, e o último aminoácido, que possui apenas o ângulo ϕ. Por convenção, os ângulos ϕ e ψ possuem valores iguais a 180° quando a conformação da proteína é completamente distendida e todas as unidades estruturais básicas estão em um mesmo plano. Como os movimentos de rotação em torno de Cα são os únicos graus de liberdade da cadeia principal, sua conformação espacial é totalmente determinada conhecendo-se os pares ϕ e ψ para todos os aminoácidos da cadeia polipeptídica. Embora, em princípio, ϕ e ψ possam assumir quaisquer valores entre -180° e +180°, a maior parte dessas configurações é proibitiva devido às restrições envolvendo a compatibilidade entre as localizações no espaço dos átomos da cadeia polipeptídica. O espaço de conformações de um polipeptídeo pode ser mapeado em um gráfico de ângulos diedrais ϕ x ψ, denominado *Ramachandran Plot* (Figura 2.8), que permite visualizar as conformações permitidas e não permitidas.

As estruturas secundárias são padrões de conformação local da cadeia polipeptídica que independem do restante da cadeia. Elas envolvem apenas átomos da cadeia principal e são recorrentes na maioria das estruturas tridimensionais de proteínas, sendo que as estruturas mais freqüentes são

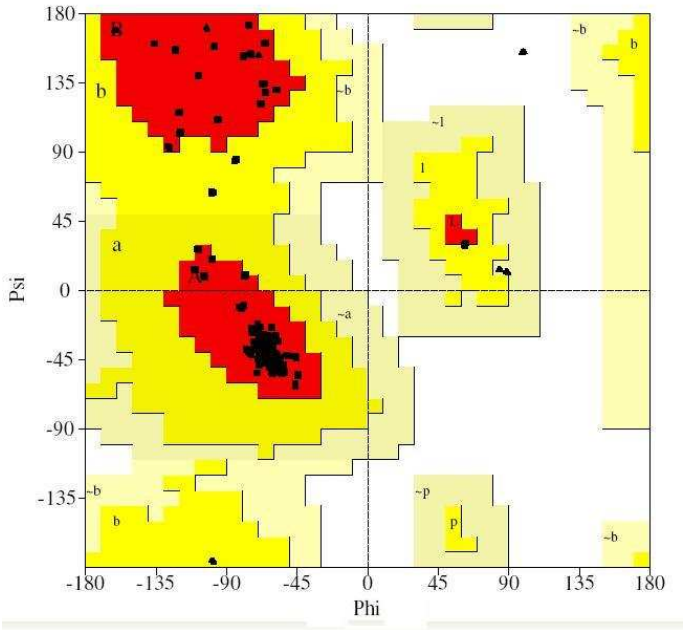

Figura 2.8: Exemplo de *Ramachandran Plot* apresentando os ângulos diedrais, ψ e ϕ, típicos para uma estrutura de proteína.

denominadas hélices α e folhas β. Nas hélices α a cadeia polipeptídica assume uma conformação espiralada em torno de um eixo imaginário no centro da espiral, com os grupos R posicionados na parte externa da espiral (Figura 2.9(a)). Para a maioria das hélices α, olhando-a a partir da terminação carboxila e a espiral girando da terminação amino para a terminação carboxila, o sentido de rotação da espiral é anti-horário (*right-hand*), sendo rara a ocorrência de hélices com rotação no sentido horário (*left-hand*). Cada volta da hélice (*right-hand*) compreende em média 3.6 aminoácidos e se estende por 5.4 Å enquanto os valores de ϕ variam de -45° a -50° e os de ψ apresentam valores em torno de -60°. A estrutura de hélice α é estabilizada por ligações do tipo ponte de hidrogênio entre o átomo nitrogênio da cadeia principal de um aminoácido e o átomo de oxigênio da cadeia principal do aminoácido distante de quatro posições no sentido da terminação amina (Figura 2.9(a)). Outros tipos de hélices também são observados: a hélice 3_{10} apresenta um período de 3 aminoácidos por volta, sendo estabilizada por pontes de hidrogênio entre cada aminoácido e o terceiro aminoácido na direção da terminação amina; a hélice π apresenta um período de 4.4 aminoácidos por volta, sendo estabilizada por pontes de hidrogênio entre cada aminoácido e o quinto aminoácido na direção da terminação amina. Ambos os tipos são raros, tendo sido observados principalmente no final de hélices α.

Nas folhas β, segmentos da cadeia polipeptídica, denominados fitas β, assumem uma conformação em zig-zag e são dispostos lado a lado, formando uma estrutura pregueada. Esta estrutura

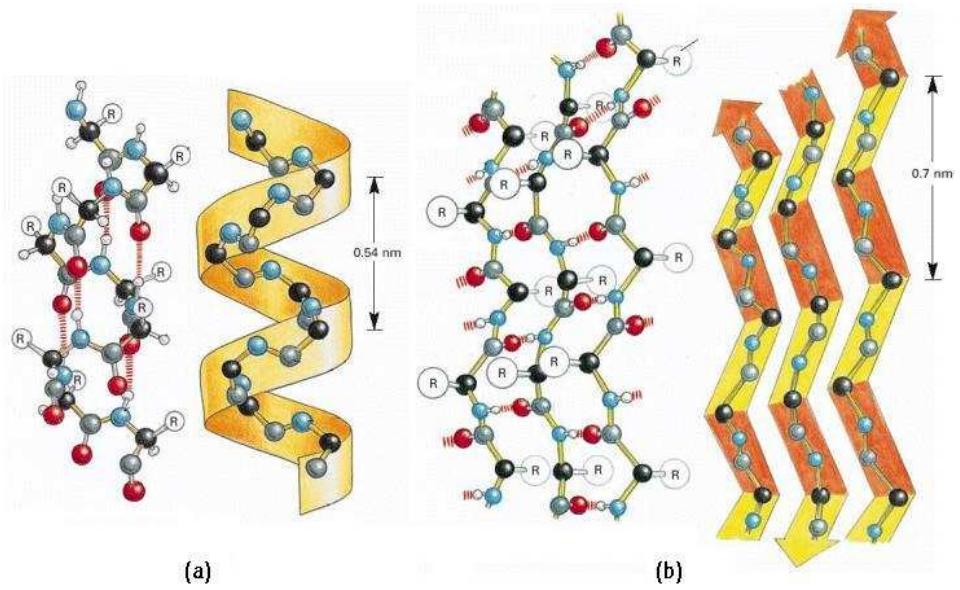

Figura 2.9: Estruturas secundárias mais comuns. (a) hélice α: à esquerda, a disposição dos átomos da cadeia polipeptídica e pontes de hidrogênio que estabilizam a estrutura; à direita, os átomos da cadeia principal e a representação gráfica da hélice α; (b) folha β com configuração anti-paralela: à esquerda, a disposição dos átomos da cadeia polipeptídica e pontes de hidrogênio que estabilizam a estrutura; à direita, os átomos da cadeia principal e a representação gráfica da folha β. Adaptado de Alberts et al. (1994).

é estabilizada por pontes de hidrogênio que se formam entre aminoácidos de segmentos adjacentes (Figura 2.9(b)). Usualmente, os segmentos que formam a folha β estão próximos na cadeia polipeptídica, mas também podem estar muito distantes ou até mesmo pertencer a diferentes cadeias polipeptídicas. Os grupos R de aminoácidos adjacentes emergem da estrutura em zig-zag em direções opostas criando um padrão alternado (Figura 2.9(b)). Se os segmentos das cadeias polipeptídicas adjacentes de uma estrutura em folha ß possuem a mesma orientação da terminação amina para a terminação carboxila, diz-se que a estrutura é paralela; e se eles possuem orientações opostas, diz-se que ela é antiparalela. Embora o período de repetição (6.5 Å para a estrutura paralela e 7 Å para a estrutura antiparalela) e o padrão de pontes de hidrogênio nessas variantes sejam diferentes, as estruturas que elas formam são bastante similares. Variantes de folhas β também foram observadas como, por exemplo, a protuberância β (β *bulge*). Mais freqüentemente observada em estruturas antiparalelas, essa variante é caracterizada pela existência de pontes de hidrogênio entre dois aminoácidos de uma mesma fita β com um único aminoácido da fita adjacente (Scheeff e Fink, 2003).

As hélices α e folhas β são as estrutura secundárias predominantes em proteínas, sendo interligadas por regiões fracamente estruturadas, denominadas *loops* ou *coils*. Particularmente comuns são as regiões que conectam regiões adjacentes em folhas β antiparalelas, denominadas contornos

(*turns*) (Nelson e Cox, 2000). Os contornos β são caracterizados por quatro aminoácidos, cuja conformação é estabilizada por uma ponte de hidrogênio entre o primeiro e o quarto aminoácido. Eles se subdividem em tipo I e tipo II, sendo este último caracterizado pela presença de uma glicina como terceiro aminoácido. Outro tipo de contorno, menos freqüente, é o contorno γ (γ *turn*), caracterizado por três aminoácidos e estabilizado por uma ponte de hidrogênio entre o primeiro e o terceiro aminoácidos. Além disso, em um nível de organização intermediário de caracterização da estrutura da proteína, entre as estruturas secundária e terciária, observa-se a ocorrência de combinações simples de alguns poucos elementos de estruturas secundárias formando arranjos geométricos e conexões topológicas características, denominadas estruturas supersecundárias ou motivos (Branden e Tooze, 1999).

2.1.4 Estruturas terciária e quaternária

O posicionamento relativo dos aminoácidos que compõem a cadeia polipeptídica determina sua conformação tridimensional, denominada estrutura terciária (Figura 2.10). De forma diferente da estrutura secundária, que é estabilizada por interações entre aminoácidos próximos em termos de seqüência, a estabilização da estrutura terciária envolve interações entre átomos de aminoácidos distantes em termos de seqüência. Outra diferença refere-se ao fato de que o tipo de interação envolvido na estabilização das estruturas secundárias é a ponte de hidrogênio, envolvendo os grupos carboxila e amina da cadeia principal, enquanto o principal tipo de interação envolvido na estabilização da estrutura terciária é a interação hidrofóbica, com massiva participação dos átomos dos grupos R (Nelson e Cox, 2000).

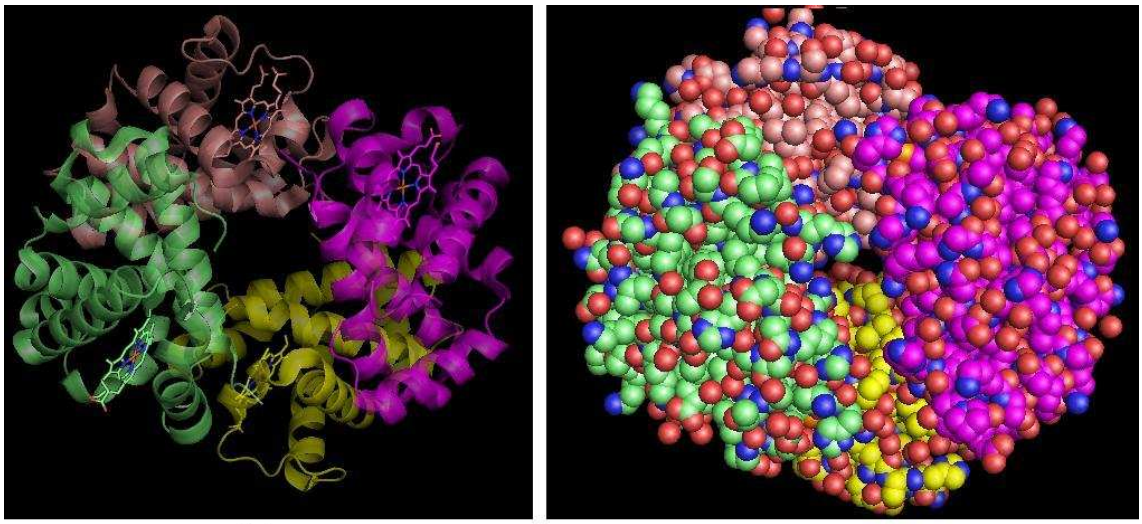

Figura 2.10: Estrutura tridimensional da molécula de hemoglobina, incluindo os grupos prostéticos heme (arquivo pdb 1gzx). (a) Estruturas secundárias. (b) átomos modelados como esferas[3].

Tão logo as primeiras estruturas de proteínas foram determinadas, observou-se que seu interior era constituído por um grande número de aminoácidos de caráter hidrofóbico, que formavam um núcleo hidrofóbico, enquanto aminoácidos polares ficavam expostos na superfície da proteína, formando uma superfície hidrofílica (Branden e Tooze, 1999). Aminoácidos polares também foram observados no núcleo hidrofóbico, mas sempre envolvidos em interações que neutralizavam suas cargas, em geral, através de pontes de hidrogênio com átomos da cadeia lateral de outros aminoácidos polares ou da cadeia principal não envolvidos na formação de estruturas secundárias. Também existe a ocorrência de aminoácidos polares que neutralizam suas cargas interagindo com moléculas de água, aprisionadas em cavidades formadas no interior da molécula de proteína. Além disso, dois grupos thiol (-SH) da cadeia lateral de aminoácidos do tipo cisteína (Cys) podem formar ligações covalentes, denominadas pontes de sulfeto, que contribuem enormemente para a estabilização da estrutura terciária da proteína (Branden e Tooze, 1999).

Proteínas também podem ser formadas por duas ou mais cadeias polipeptídicas, o que implica em um nível adicional de organização estrutural, denominado estrutura quaternária. Proteínas que possuem estrutura quaternária são denominadas multímeros. Quando a estrutura quaternária de uma proteína é composta por poucas cadeias, ela é denominada oligômero. Particularmente, quando o número de cadeias é igual a dois, utiliza-se a denominação dímero. Além disso, os prefixos homo e hetero podem ser utilizados para enfatizar o fato das cadeias polipeptídicas serem idênticas ou diferentes. Em sua maioria, os multímeros são formados por cadeias polipeptídicas idênticas ou grupos de cadeias polipeptídicas não idênticas que se repetem de forma simétrica. A unidade estrutural que se repete, seja ela uma cadeia polipeptídica ou um grupo de cadeias polipeptídicas não idênticas, é denominada protômero (Nelson e Cox, 2000).

Em adição, algumas proteínas também precisam incorporar à sua estrutura modificações de aminoácidos e/ou interações com grupos químicos diferentes de aminoácidos para que sejam capazes de realizar sua função ou mesmo para estabilizar sua conformação tridimensional. Por exemplo, observou-se que a incorporação de carboidratos à estrutura de algumas proteínas serve como um sinalizador da sua localização intracelular e que a incorporação de lipídios pode ajudar a ancorar a proteína a uma membrana celular (Scheeff e Fink, 2003). Proteínas com essas características são denominadas proteínas conjugadas, em contraposição às proteínas constituídas apenas de aminoácidos, denominadas proteínas simples (Nelson e Cox, 2000). As proteínas conjugadas são agrupadas de acordo com o grupo químico que é incorporado à sua estrutura, denominado grupo prostético. Assim, por exemplo, a proteína é denominada metaloproteína se o grupo prostético possui íons metálicos, glicoproteína se contém oligossacarídeo e lipoproteína se contém lipídeos (Nelson e Cox, 2000).

Geralmente, as proteínas são classificadas em três grandes grupos (Nelson e Cox, 2000; Scheeff e Fink, 2003): proteínas fibrosas, formadas por cadeias polipeptídicas arranjadas em uma confor-

mação formando extensas fitas ou folhas; proteínas globulares, formadas por cadeias polipeptídicas com conformação globular; e proteínas de membrana, similares às proteínas globulares, mas que se encontram inseridas em uma membrana celular. A estrutura de proteínas fibrosas contém apenas um tipo de estrutura secundária e, basicamente, têm função estrutural; enquanto que proteínas globulares apresentam diferentes tipos de estruturas secundárias, são encontradas em um ambiente aquoso intracelular e, em geral, têm função associada a processos enzimáticos ou de regulação. As proteínas de membrana, por serem encontradas em ambiente hidrofóbico (membrana celular), diferem das proteínas globulares pelo fato de apresentarem uma superfície altamente hidrofóbica (Scheeff e Fink, 2003).

Ainda neste nível de organização estrutural, observa-se a existência de regiões da cadeia polipeptídica capazes de manter sua estrutura tridimensional estável e, muitas vezes sua funcionalidade, mesmo quando isoladas do restante da molécula de proteína. Essas regiões são denominadas domínios (Scheeff e Fink, 2003) e são as unidades básicas para a classificação estrutural de proteínas.

2.2 Dados experimentais sobre estruturas de proteínas

Um modelo para a estrutura tridimensional de uma molécula de proteína consiste, basicamente, do conjunto de coordenadas xyz de seus átomos. Esses modelos são importantes para a análise da função das proteínas, pois muitos detalhes do seu mecanismo de funcionamento só podem ser desvendados através da análise de sua estrutura em nível atômico. A análise do mecanismo pelo qual uma enzima catalisa uma reação química específica ou a forma como uma proteína de transporte captura e libera a molécula por ela transportada (Rhodes, 1993) são exemplos de estudos que se beneficiam deste tipo de dado. Atualmente, são dois os métodos experimentais mais utilizados para determinação de estruturas de proteínas: a cristalografia por difração de raios X e a ressonância magnética nuclear - RMN (Branden e Tooze, 1999; Wider, 2000). O método de cristalização por raios X pode ser considerado, num certo sentido, um método direto, na medida em que ele consiste em capturar a imagem da molécula (Rhodes, 1993), ou sua densidade eletrônica. A aplicação do método é restrita a proteínas que sejam capazes de produzir cristais e que, por sua vez, sejam capazes de produzir difração de raios X de boa qualidade. Já o método de ressonância magnética nuclear pode ser entendido como um método indireto, uma vez que as estruturas são obtidas pela satisfação de restrições determinadas experimentalmente. O método apresenta a vantagem de determinar a estrutura da proteína em solução, mas sofre restrições quanto ao tamanho da proteína. Atualmente, a utilização do método está restrita a proteínas com peso molecular inferior a 30 kd (Wider, 2000).

2.2.1 Banco de dados de estruturas de proteínas

Em geral, uma vez elaborado o modelo final da estrutura tridimensional da molécula de proteína (um conjunto de modelos no caso de experimentos de RMN), seus autores o depositam no *Protein Data Bank* (PDB) (Berman et al., 2000). O PDB é um banco de dados de âmbito mundial, onde as estruturas de macromoléculas resolvidas experimentalmente ficam publicamente disponíveis para outros cientistas, pesquisadores e o público em geral. Atualmente, são mais de 55.000 estruturas depositadas, incluindo moléculas de proteínas isoladas, em complexo com outras proteínas, DNA e outros ligantes.

As estruturas depositadas no PDB são distribuídas através de arquivos, identificados por um código composto por quatro caracteres, onde o primeiro é sempre um dígito e os demais são caracteres alfanuméricos (ex: 1cho, 1ppf, 1dhk, etc.). Diferentes formatos são utilizados para distribuição desses arquivos (Berman et al., 2000):

- **Formato PDB**, que é o mais antigo e consiste de um arquivo texto formatado, sendo mais apropriado para leitura por humanos;
- **Formato mmCIF** (*MacroMolecular Crystllograhic InFormation*) (Westbrook e Fitzgerald, 2003) que é o formato oficial para distribuição de arquivos PDB e utiliza um conjunto pré-definido de termos (dicionário), sendo mais apropriado para leitura por máquina; e
- **Formato PDBML/XML** (*PDB Markup Language*) (Westbrook e Fitzgerald, 2003) que segue o padrão XML (Harold e Means, 2001) para estruturação de informação e documentos complexos, sendo também mais apropriado para leitura por máquina.

O conteúdo de um arquivo PDB inclui:

- informações de caráter geral como o nome da proteína, os autores do depósito e as publicações relacionadas;
- características bioquímicas relacionadas à proteína;
- detalhes experimentais relacionados à determinação da estrutura;
- algumas características estruturais, incluindo atribuições de estruturas secundárias e localização de pontes de hidrogênio; e
- informações biológicas, como a localização de sítios ativos.

```
ATOM      1  N   VAL A   1      18.432  -2.931   3.579  1.00 37.68           N
ATOM      2  CA  VAL A   1      19.662  -2.549   2.806  1.00 35.41           C
ATOM      3  C   VAL A   1      19.282  -1.939   1.441  1.00 34.04           C
ATOM      4  O   VAL A   1      18.421  -2.497   0.695  1.00 33.95           O
ATOM      5  CB  VAL A   1      20.659  -3.754   2.825  1.00 35.59           C
ATOM      6  CG1 VAL A   1      20.109  -4.992   2.222  1.00 37.84           C
ATOM      7  CG2 VAL A   1      21.982  -3.272   2.245  1.00 36.73           C
ATOM      8  N   LEU A   2      19.905  -0.786   1.169  1.00 29.21           N
ATOM      9  CA  LEU A   2      19.749  -0.064  -0.067  1.00 27.27           C
ATOM     10  C   LEU A   2      20.513  -0.749  -1.213  1.00 27.19           C
ATOM     11  O   LEU A   2      21.748  -0.901  -1.212  1.00 27.58           O
ATOM     12  CB  LEU A   2      20.204   1.339   0.210  1.00 25.79           C
ATOM     13  CG  LEU A   2      19.275   2.508   0.284  1.00 30.66           C
ATOM     14  CD1 LEU A   2      17.858   2.278   0.784  1.00 26.00           C
```

Figura 2.11: Exemplo de descrição dos átomos de uma proteína no arquivo PDB: 1gzx, cadeia A, aminoácidos de números 1 e 2, identificados pela sexta coluna da esquerda para a direita.

Entretanto, a informação fundamental nesses arquivos é o próprio modelo da estrutura tridimensional da proteína, representada pelas coordenadas tridimensionais dos átomos que a compõem. Considerando o formato PDB, essa informação é relatada nas linhas contendo o rótulo ATOM (vide Figura 2.11), sendo os átomos identificados por uma versão em língua inglesa da convenção de letras gregas utilizadas pela química orgânica. Assim, para cada aminoácido, são listados os átomos da cadeia principal (o nitrogênio do grupo amina (N), o carbono α (CA) e o carbono (C) e o oxigênio (O) do grupo carboxila), seguidos pelos átomos da cadeia lateral (carbono ß (CB), carbono γ (CG), e assim por diante). No caso de cadeias laterais com ramificações ou anéis, a identificação de cada átomo é acrescida da numeração 1 ou 2 depois da letra grega. Assim, por exemplo, no caso de um aminoácido do tipo leucina, os átomos são listados na seguinte ordem: N, CA, C, O, CB, CG, CD1 e CD2 (vide aminoácido 2, indicado pela sexta coluna, na Figura 2.11).

2.3 Propriedades físico-químicas e estruturais de aminoácidos

Aminoácidos expostos na superfície de moléculas de proteína podem ser caracterizados por diferentes propriedades físico-químicas e estruturais. Nesta seção, são apresentadas as principais propriedades relatadas na literatura como relevantes para o estudo estrutural do fenômeno bioquímico de interação proteína-proteína.

2.3.1 Área da superfície acessível a solvente e molecular

Uma das propriedades mais importantes para a análise estrutural de macromoléculas é a área de sua superfície que está exposta para interação com o solvente (água) e/ou outras moléculas. Existem três definições bastante difundidas para a representação da superfície de moléculas de proteína: a superfície de van der Waals (VWS), a superfície acessível a solvente (SAS) e a superfície molecular

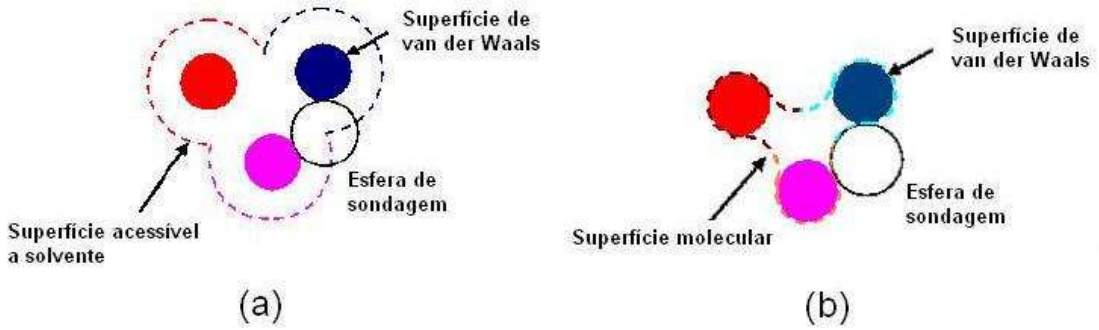

Figura 2.12: Definição de superfície da molécula de proteína (analogia em 2 dimensões) (a) A superfície acessível a solvente é definida pelo lugar geométrico descrito pelo centro da esfera de sondagem à medida que esta tangencia a molécula. (b) A superfície molecular é definida pelo envelope definido pela esfera de sondagem neste mesmo processo.

(MS). A superfície de van der Waals (Lee e Richards, 1971) é definida pela superfície do modelo molecular de preenchimento de espaço (vide Figura 2.10(b)). Neste modelo, os átomos da molécula de proteína são considerados como esferas rígidas centradas nas coordenadas xyz e raios determinados pelos correspondentes raios de van der Waals.

As duas outras definições de superfície molecular, superfície acessível a solvente (SAS) e superfície molecular (MS) (Richards, 1977), consideram uma esfera de sondagem tangenciando a molécula de proteína definida pelo modelo de preenchimento de espaço. Em geral, assume-se que a esfera de sondagem possui raio entre 1.2 Å e 1.5 Å para simular as dimensões de uma molécula de água, a menor molécula que interage com a proteína e também a mais abundante na célula. A SAS é definida como o lugar geométrico descrito pelo centro da esfera de sondagem, à medida que ela tangencia a molécula de proteína (vide Figura 2.12(a)), enquanto a MS é definida como o envelope descrito pela esfera de sondagem no mesmo processo de tangenciamento da molécula de proteína (vide Figura 2.12(b)). Por vezes, a MS também é referenciada como superfície com solvente excluído.

De acordo com estas definições, tanto a SAS quanto a MS são formadas por uma composição de superfícies menores associadas aos átomos das moléculas de proteína (vide Figura 2.12), de tal forma que a obtenção da contribuição individual de cada átomo é direta e a de cada aminoácido pode ser obtida através da somatória das contribuições individuais de seus átomos. Também é possível calcular a área acessível a solvente relativa para cada aminoácido. Neste caso, uma área acessível a solvente de referência, para cada tipo de aminoácido (Ahmed et al., 2004), é utilizada para determinar a contribuição de cada um dos aminoácidos da molécula de proteína, em termos de fração da SAS.

Os métodos para cálculo das áreas da SAS e MS são classificados em dois grupos: métodos exatos e métodos baseados em aproximações numéricas. O cálculo da área da SAS utiliza o fato de que ela também pode ser definida por esferas com raios estendidos, correspondentes aos átomos da

molécula de proteína, com os raios dados pela soma dos raios de van der Waals do átomo e da esfera de sondagem (vide Figura 2.12(a)). Neste caso, a contribuição de cada átomo para a SAS é dada pela parte da superfície da correspondente esfera não obstruída pelas esferas correspondentes a outros átomos na sua vizinhança. Os métodos exatos determinam o conjunto de polígonos curvos que definem a SAS e empregam o teorema de Gauss-Bonnet (Pressley, 2001) para calcular as contribuições em área (Richmond, 1984) de cada esfera. Já os métodos aproximados, distribuem pontos sobre a superfície de cada esfera e eliminam aqueles que estão contidos em alguma das esferas em sua vizinhança. A contribuição de cada átomo para a SAS pode ser aproximada pela proporção da área total da esfera (Shrake e Rupley, 1973; Eisenhaber et al., 1995), dada pelo número de pontos não obstruídos.

Os métodos de cálculo da área da MS a dividem em: (i) regiões convexas, correspondentes às superfícies de van der Waals dos átomos que são tangenciados pela esfera de sondagem; (ii) regiões côncavas descritas pela região interna da superfície de um toróide, correspondentes às situações em que a esfera de sondagem tangencia apenas dois átomos; e (iii) regiões côncavas, descritas por polígonos curvos definidos sobre a esfera de sondagem, correspondentes às situações em que a esfera de sondagem tangencia três ou mais átomos ao mesmo tempo. De forma análoga aos métodos para cálculo das áreas de SAS, os métodos exatos para cálculo da área da MS utilizam o teorema de Gauss-Bonnet (Connolly, 1983; Sanner et al., 1996; Tsodikov et al., 2002; Liang et al., 1998), enquanto os métodos aproximados determinam a contribuição de cada região da MS pela proporção da área de objetos geométricos conhecidos, como esferas e toróides, que elas representam (Greer e Bush, 1978).

2.3.2 *Residue depth*, *half sphere exposure* e *coordination number*

Diversas propriedades, diferentes da SAS, têm sido propostos na literatura para se aferir o grau de acessibilidade a solvente. Uma delas, denominada *Residue depth* (RD) (Chakravarty e Varadarajan, 1999), é definida como a média das *atom depth*s associadas a cada um do seus átomos, onde *atom depth* é definida como a menor distância entre o átomo e a superfície molecular (MS). Outra propriedade usada para aferir o grau de acessibilidade a solvente, denominada *half sphere exposure* (HSE) (Hamelryck, 2005), é definida utilizando uma esferra de raio igual a 13 Å, cujo centro coincide com a posição do átomo Cα. As duas meias esferas definidas pelo plano ortogonal à linha ligando os átomos Cα e Cβ são denominados *side chain half sphere* e *main chain half sphere*. O HSE, então, é definido por duas quantidades: HSEu e HSEd, onde HSEu é o número de átomos Cα contidos na *side chain half sphere* e HSEd é número de Cα contidos na *main chain half sphere*.

Além disso, somando-se os valores de HSEu e HSEd, obtém-se o *coordination number* (CN), um parâmetro que mede a densidade atômica do ambiente em que o aminoácido se encontra, ou seja, o grau de empacotamento do aminoácido.

2.3.3 Índice de planaridade

Jones e Thornton (1997a) relataram que, geometricamente, a superfície da região de interface proteína-proteína é razoavelmente plana e propuseram um índice para avaliar o quão plana é a região em que um aminoácido está inserido. Este índice, denominado índice de planaridade, considera as coordenadas xyz dos átomos Cα dos aminoácidos na região sendo avaliada e é definido como o recíproco do erro quadrático médio associado ao plano de regressão de mínimos quadrados passando por estes átomos.

2.3.4 Curvaturas sobre superfícies geométricas

Uma forma mais elegante para caracterizar a geometria da superfície molecular é a utilização do conceito de curvatura sobre superfícies geométricas. Intuitivamente, a curvatura associada a um ponto de uma curva definida em um plano (2D) indica a variação do ângulo da reta tangente a ele com relação às retas tangentes aos pontos em sua vizinhança. Formalmente, seja a curva no plano, $\gamma : \Re \rightarrow \Re^2$, parametrizada pelo comprimento de arco, ou seja, tal que $\|\dot{\gamma}(t)\| = 1$ para $\forall t \in [-\infty, \infty]$, onde a notação com ponto denota a diferenciação de γ em relação a t. O escalar $\|\ddot{\gamma}(t)\|$ define a curvatura de $\gamma(t)$ no ponto t e é denotada por $k(t)$ (Pressley, 2001). A partir desta definição, é possível mostrar que uma reta possui curvatura igual a zero e uma circunferência de raio r possui curvatura constante e igual ao inverso de seu raio, $k(t) = \frac{1}{r}$ (Pressley, 2001).

Também é possível calcular a curvatura para curvas definidas no espaço tridimensional (3D), através da seguinte equação:

$$k = \frac{\|\ddot{\gamma} \times \dot{\gamma}\|}{\|\dot{\gamma}\|^3}, \tag{2.1}$$

onde $\times$ denota produto vetorial.

Para superfícies no $\Re^3$, diferentes tipos de curvaturas são definidas. Considere uma superfície paramétrica, $S : \Re^2 \rightarrow \Re^3$, um ponto, P, sobre S e um vetor unitário, v, tangente a S em P. Considere também um plano, ω, formando um ângulo $\pi/2$ com o plano tangente à S em P, tal que $v \in \omega$, e a curva formada pela interseção de ω e S, $\gamma = \omega \bigcap S$ (Figura 2.13). Para diferentes vetores v, resultam diferentes planos ω, que produzem diferentes interseções com a superfície S, e que, por sua vez, produzem diferentes curvas γ, cada uma delas com um valor diferente de curvatura em P. Os valores máximo e mínimo dessas curvaturas são definidos como as curvaturas principais, k_1 e k_2, da superfície S em P. Duas outras curvaturas, denominadas curvatura gaussiana, K, e curvatura média, H, também são definidas para superfícies no $\Re^3$, sendo calculadas como (Pressley, 2001):

$$K = k_1 \cdot k_2 \quad e \quad H = \frac{1}{2}(k_1 + k_2) \tag{2.2}$$

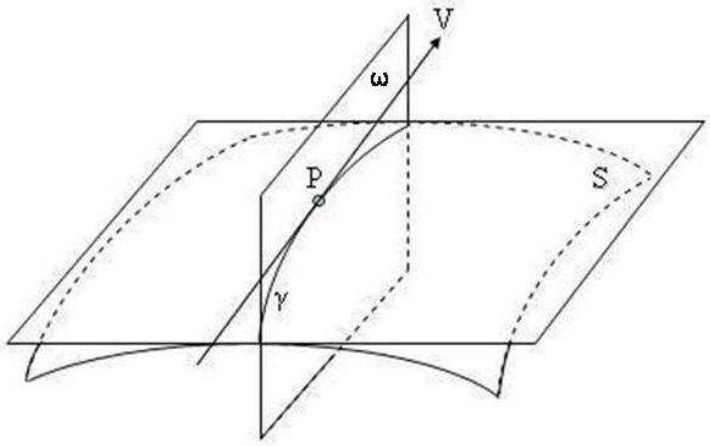

Figura 2.13: Definição de curvaturas principais sobre o ponto P de uma superfície S. Um vetor, v, tangente a P e um plano ω que definem uma curva γ sobre S são ilustrados. Adaptado de Pressley (2001)

Embora a curvatura gaussiana e média contenham a mesma informação geométrica que as curvaturas principais, elas apresentam um significado geométrico maior (Pressley, 2001). Por exemplo, a curvatura gaussiana, isoladamente, é capaz de discriminar o tipo de superfície quádrica a que um ponto pertence:

- se $K > 0$ no ponto P, então P pertence a um elipsóide;
- se $K < 0$ no ponto P, então P pertence a um hiperbolóide; e
- se $K = 0$ no ponto P, então P pertence a um parabolóide ou a um plano.

Considerando as curvaturas gaussiana, K, e média, H, em conjunto, a tabela 2.1 apresenta os tipos de superfícies que elas discriminam, de acordo com Besl e Jain (1988). A função T na Tabela 2.1 é definida por:

$$T = 1 + 3(1 + sgn_{\varepsilon}(H)) + (1 - sgn_{\varepsilon}(K)),\ onde : sgn_{\varepsilon}(x) = \begin{cases} -1, & se \quad x < \varepsilon \\ 0, & se \quad |x| \leq \varepsilon \\ 1, & se \quad x > \varepsilon \end{cases} \tag{2.3}$$

e ε é um valor arbitrário para o limiar de erro.

Além disso, índices baseados nas curvaturas principais, ou, de forma equivalente, nas curvaturas gaussiana e média, também podem ser formulados. Por exemplo, Koenderink (1990) propôs a *curvedness*, R, e o índice de forma, S, definidos como:

	$K > 0$	$K = 0$	$K < 0$
$H < 0$	*peak* $T = 1$	*ridge* $T = 2$	*saddle-ridge* $T = 3$
$H = 0$	*none* $T = 4$	*flat* $T = 5$	*minimum-surface* $T = 6$
$H > 0$	*pit* $T = 7$	*valley* $T = 8$	*daddle-valley* $T = 9$

Tabela 2.1: Interpretação de valores de curvaturas média e gaussiana (Besl e Jain, 1988).

$$R = \sqrt{\frac{k_2^2 + k_1^2}{2}} \qquad e \qquad S = -\frac{2}{\pi} \arctan\left(\frac{k_2 + k_1}{k_2 - k_1}\right), \tag{2.4}$$

onde k_1 e k_2 correspondem às curvaturas principais mínima e máxima, respectivamente. Os valores de S pertencem ao intervalo -1 e 1, sendo negativo para superfícies côncavas e positivo para superfícies convexas.

Em geral, não é possível obter uma representação paramétrica para as superfícies de moléculas de proteínas, tal que representações discretas são utilizadas, sendo a curvatura estimada através de métodos específicos. Estes podem se basear apenas na utilização do conjunto de pontos da representação discreta da superfície ou podem ter como pré-requisito a disponibilidade da informação sobre a conectividade entre esses pontos (ex: triangulação).

Uma das formas mais comuns para estimação dos valores de curvaturas para superfícies com representação discreta consiste em aproximar, localmente, essa superfície por uma função contínua e, então, calcular os valores das curvaturas para essa função. Em geral, uma função polinomial de segundo grau é considerada para aproximação da superfície e o conjunto de pontos utilizado é formado pelos pontos na vizinhança do ponto de interesse. As curvaturas da quádrica no ponto de interesse são, então, tomadas como as estimativas das curvaturas da superfície discreta.

Na sua forma mais simples, os métodos para cálculo de curvaturas de superfícies por aproximação por quádricas assumem que a superfície é descrita como $z' = ax'^2 + bx'y' + cy'^2$. Neste caso, a quádrica é descrita em um sistema de coordenadas local com o eixo z' alinhado com uma estimativa do vetor normal ao plano tangente à superfície no ponto de interesse. McIvor e Valkenburg (1997) descrevem um procedimento para cálculo de curvaturas que utiliza esta representação.

Cálculo de curvaturas por aproximação de quádricas (McIvor e Valkenburg, 1997):

1. estime o vetor normal ao plano tangente S, $\mathbf{n}$, no ponto de interesse $\mathbf{p}$. Esta estimativa pode ser feita por média simples ou ponderada dos vetores normais associados aos planos que contém

os triângulos vizinhos a $\mathbf{p}$, no caso de uma triangulação, ou pelo plano obtido através de ajuste de mínimos quadrados utilizando os pontos na vizinhança $\mathbf{p}$ para os casos em que a informação de conectividade entre pontos não está disponível;

2. defina um sistema de coordenadas local $(\mathbf{x}', \mathbf{y}', \mathbf{z}')$ no ponto $\mathbf{p}$ com o eixo de coordenadas z' alinhada ao vetor normal estimado. Para fixar o eixo de coordenadas x', McIvor e Valkenburg (1997) sugerem alinhá-lo com a projeção do eixo $\mathbf{x}$ do sistema de coordenadas global sobre o plano definido por $\mathbf{n}$. Isto resulta na matriz de rotação $\mathbf{R} = [\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}]^t$, que mapeia os pontos do sistema de coordenadas global para sistema de coordenadas local, tal que $\mathbf{r_3} = \mathbf{n}$, $\mathbf{r_1} = \frac{\mathbf{(I-nn^t)i}}{\|\mathbf{(I-nn^t)i}\|}$ e $\mathbf{r_2} = \mathbf{r_3} \times \mathbf{r_1}$, onde $\mathbf{I}$ é a matriz identidade e $\mathbf{i}$ é o vetor unitário sobre o eixo $\mathbf{x}$, $[1, 0, 0]^t$. Se a direção do vetor normal coincide com o eixo $\mathbf{x}$, outra direção deve ser escolhida, por exemplo, a do eixo $\mathbf{y}$;

3. selecione o conjunto de pontos para ajustar a quádrica, utilizando ou não informação sobre a conectividade entre os pontos;

4. mapeie os pontos selecionados, $\mathbf{x_s}$, do sistema de coordenadas global para o sistema de coordenadas local utilizando a relação $\mathbf{x'_s} = \mathbf{R}(\mathbf{x_s} - \mathbf{p})$;

5. Determine os coeficientes a, b e c que definem a quádrica, resolvendo o sistema de equações

$$\begin{pmatrix} x_1^2 & x_1 y_1 & y_1^2 \\ \dots & \dots & \dots \\ x_n^2 & x_n y_n & y_n^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} z_1 \\ \dots \\ z_n \end{pmatrix}.$$

   Este sistema, $\mathbf{Ax} = \mathbf{b}$, é sobre-determinado e sua solução é dada por $\mathbf{x} = (\mathbf{A^t A})^{-1}\mathbf{A^t b}$

6. Estime as curvaturas principais, gaussiana e média:

$$\begin{aligned} k_1 &= a + c + \sqrt{(a-c)^2 + b^2}, \\ k_2 &= a + c - \sqrt{(a-c)^2 + b^2}, \\ K &= 4ac - b^2, \\ H &= a + c. \end{aligned} \tag{2.5}$$

Estas estimativas dependem em grande parte da acuidade da estimativa do vetor normal à superfície tangente. Se a quádrica a ser ajustada, denominada quádrica estendida, é dada por $z' = ax'^2 + bx'y' + cy'^2 + dx' + ey'$, o vetor normal à superfície tangente na origem do sistema de coorde-

nadas local é dado por:

$$n(0) = \frac{1}{\sqrt{1+d^2+e^2}} \begin{pmatrix} -d \\ -e \\ 1 \end{pmatrix}.$$

Então, esta nova estimativa do vetor normal pode ser utilizada para calcular uma nova estimativa do sistema de coordenadas local, tal que uma nova quádrica pode ser ajustada nesse novo sistema de coordenadas. Este processo pode ser repetido até que algum critério de convergência pré-estabelecido seja atingido. Os coeficientes da quádrica resultante são, então, utilizados para calcular as curvaturas gaussiana e média da seguinte forma:

$$\begin{aligned} K &= \tfrac{4ac-b^2}{(1+d^2+e^2)^2}, \\ H &= \tfrac{a+c+ae^2+cd^2-bde}{(1+d^2+e^2)^{3/2}}. \end{aligned} \tag{2.6}$$

As curvaturas principais, k_1 e k_2 são obtidas resolvendo-se as equações 2.2.

Para superfícies trianguladas, uma forma alternativa para se estimar as curvaturas é o método apresentado por Meyer et al. (2002), que utiliza operadores de geometria diferencial discreta. O teorema de Gauss-Bonet (Pressley, 2001), considerando a vizinhança definida pelos triângulos conectados a $\mathbf{p}$, é utilizado para estimar a curvatura gaussiana, enquanto o operador normal da curvatura média, definido por $2H\mathbf{n}$, é utilizado para calcular a curvatura média. Dado uma região de triângulos em torno de $\mathbf{p}$, as curvaturas gaussianas, K, e média, H, são estimadas por (Meyer et al., 2002):

$$\begin{aligned} K &= \tfrac{1}{A}\left(2\pi - \textstyle\sum_j \theta_j\right) \; e \\ 2H\mathbf{n} &= \tfrac{1}{2A}\textstyle\sum_j(\cot(\alpha_j)+\cot(\beta_j))(\mathbf{p}-\mathbf{x_j}), \end{aligned} \tag{2.7}$$

onde A corresponde a uma área que contém $\mathbf{p}$ e $\mathbf{n}$ é o vetor normal à superfície em $\mathbf{p}$. θ_j, α_j e β_j são definidos conforme apresentado na Figura 2.14. Meyer et al. (2002) sugerem que a área A seja escolhida como sendo a área de Voronoi, definida em cada triângulo pelo ponto p, os pontos médios dos vértices do triângulo adjacentes a $\mathbf{p}$ e seu circuncentro (vide Figura 2.14(b)). A área de Voronoi, A_{vor}, sugerida pelos autores, é dada por $A_{vor} = 1/8\sum_j(\cot\alpha_j + \cot\beta_j)\|\mathbf{p}-\mathbf{x_j}\|^2$. Novamente, as curvaturas principais, k_1 e k_2 são obtidas resolvendo-se as equações 2.2.

2.3.5 Índice de hidrofobicidade e energia de solvatação

A propriedade de hidrofobicidade refere-se à tendência de moléculas não polares em repelir água ou, em outras palavras, à sua incapacidade de ser dissolvida em água. Ela é considerada um dos principais fatores envolvidos na estabilização tanto da estrutura tridimensional de proteínas quanto de seus complexos. Em geral, o grau de hidrofobicidade de uma molécula é avaliada experimental-

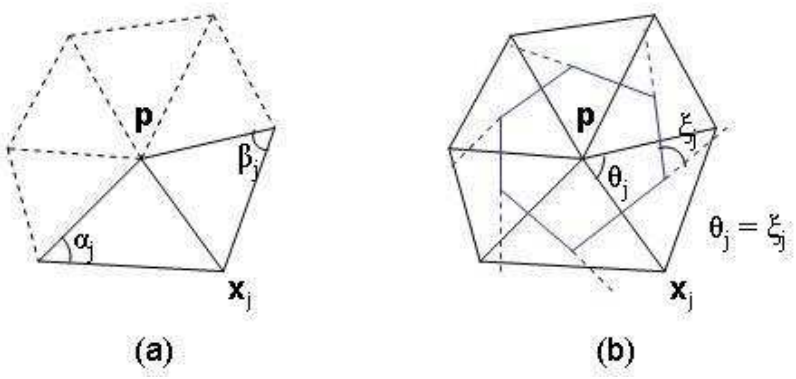

Figura 2.14: Estimativa de curvaturas utilizando operadores de geometria diferencial discreta (Meyer et al., 2002). Definições de (a) α_j e β_j e (b) Área de Voronoi e θ_j.

mente, medindo-se sua solubilidade diferencial em uma solução contendo um solvente polar e um apolar. Uma quantidade conhecida da molécula de interesse, denominada soluto, é dissolvida em uma solução contendo volumes também conhecidos dos dois tipos de solventes, em geral, água e um solvente orgânico como, por exemplo, octanol. Uma vez que os dois solventes não se misturam, é possível separar a solução em duas fases, cada uma correspondente a um dos solventes. Mede-se, então, a concentração do soluto em cada uma das fases, sendo a solubilidade diferencial do soluto em mistura solvente polar/solvente apolar expressa pelo seu coeficiente de partição, denominado logP:

$$logP = \frac{[soluto\ polar]}{[soluto\ apolar]}, \tag{2.8}$$

onde $[soluto\ polar]$ é a concentração do soluto na fase polar e $[soluto\ apolar]$ é a concentração do soluto na fase apolar.

No caso de proteínas, análises experimentais têm sido realizadas para quantificar o grau de hidrofobicidade para cada um dos 20 aminoácidos padrão em diferentes condições experimentais, resultando em diferentes índices (Fauchère e Pliska, 1983; Radzika e Wolfden, 1988; Kyte e Doolittle, 1982). Fauchère e Pliska (1983), por exemplo, utilizaram água e octanol como solventes com o pH em torno de 7 e definiram o índice de hidrofobicidade de um aminoácido como a diferença entre seu coeficiente de partição e o coeficiente de partição para o aminoácido glicina (Gly), que é o aminoácido mais simples dentre os 20 aminoácidos padrão (vide apêndice B).

A equação 2.8 pode ser relacionada à energia necessária para transferência do soluto de um meio apolar (fase do solvente apolar) para um meio polar (meio do solvente polar), através da seguinte relação:

$$\Delta G_s = -RTlogP, \tag{2.9}$$

onde R é a constante do gás ($R = 1.987 cal/mol.K$) e T a temperatura em Kelvins. ΔG_s, expresso em cal/mol, é denominado energia livre de solvatação (Nelson e Cox, 2000) e no contexto de análise de proteínas pode ser interpretado como a energia necessária para transferência de um aminoácido de seu interior para sua superfície.

Sob a suposição comum de que a energia livre de solvatação é aditiva e proporcional à área acessível a solvente (SAS), é possível determinar as contribuições individuais de cada aminoácido utilizando a razão da sua área acessível a solvente em relação à área acessível a solvente padrão do correspondente aminoácido. Esta suposição, entretanto, representa uma simplificação, uma vez que não considera a possibilidade de diferentes átomos do mesmo aminoácido apresentarem diferentes propriedades químicas, levando a regiões com características químicas distintas, como, por exemplo, o glutamato (Glu) e a lisina (Lys) que possuem caráter anfipático, apresentando, ao mesmo tempo, regiões polares e apolares.

Para contabilizar as contribuições dos diferentes tipos de átomos dos aminoácidos, Eisenberg e McLachlan (1986) e Wesson e Eisenberg (1992) propuseram a utilização de parâmetros de solvatação atômicos (ASP), $\Delta\sigma_i$, para o cálculo da energia livre de solvatação:

$$\Delta G_s = \sum_{i \in atomos} \Delta\sigma_i area(i), \tag{2.10}$$

onde $\Delta\sigma_i$, é expresso em cal/mol^2 e representa, para cada tipo de átomo, a energia livre por unidade de área acessível a solvente necessária para transferir um átomo de um meio apolar para a meio polar. A $area(i)$ representa a contribuição do átomo i para a área acessível a solvente do aminoácido e o conjunto de átomos considerado na somatória corresponde àqueles com área acessível a solvente diferente de zero. A contribuição individual de cada um desses átomos pode ser obtida das parcelas individuais da somatória na equação 2.10.

2.3.6 Potencial eletrostático

Dentre os diferentes fatores que contribuem para a energia livre de macromoléculas (proteína), as interações eletrostáticas são de especial importância por resultarem de interações à distância e à substancial carga elétrica de seus blocos básicos (aminoácidos). As interações eletrostáticas têm importante participação na determinação da estrutura e flexibilidade de macromoléculas, bem como na força de suas associações com pequenas moléculas, outras macromoléculas e membranas celulares (Baker e McCammon, 2003). Proteínas, em particular, são ricas em grupos químicos carregados, tal que a correspondente contribuição acumulada para o potencial eletrostático pode ser substancial.

A eletrostática clássica, na sua forma mais simples, considera apenas cargas elétricas distribuídas no vácuo, utilizando a equação de Poisson para descrever tais sistemas (Gilson, 2002):

$$\Delta\Phi(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\epsilon_0}, \tag{2.11}$$

onde $\rho(\mathbf{r})$ e $\Phi(\mathbf{r})$ representam a densidade de cargas elétricas e o potencial eletrostático, como funções da posição, $\mathbf{r}$, a constante ϵ_0 é a permissividade do espaço livre e Δ é o operador de Laplace. O campo elétrico associado, $E(\mathbf{r})$, é descrito por $E(\mathbf{r}) = -\nabla\Phi(\mathbf{r})$ e quando o sistema é composto apenas por cargas pontuais, $\rho(\mathbf{r})$ é representada por uma somatória de funções delta de Dirac, $\delta(.)$, tal que $\rho(\mathbf{r}) = \sum_{i=1}^{N} q_i\delta(\mathbf{r} - \mathbf{r_i})$.

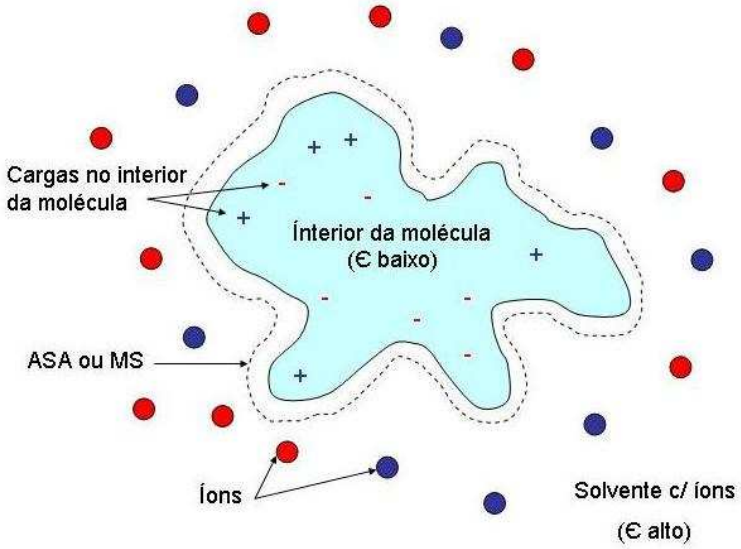

Figura 2.15: Cálculo de potencial eletrostático para moléculas de proteínas. Considera-se uma molécula de proteína em uma solução iônica.

Quando as cargas elétricas estão distribuídas em um meio dielétrico uniforme, as interações eletrostáticas entre as cargas são enfraquecidas em comparação com as interações que ocorrem no vácuo. Para modelar este fato, uma constante dielétrica, $D > 1$, é introduzida na equação 2.11, resultando em uma redução uniforme do campo elétrico em todos os pontos do espaço. Entretanto, na escala molecular, um sistema proteína/água não é bem descrito por um meio dielétrico uniforme. De fato, o interior e o exterior da molécula são caracterizados por constantes dielétricas bastante diferentes, tal que uma mudança abrupta ocorre na superfície da proteína (Figura 2.15). A equação de Poisson, neste caso, é reescrita como (Gilson, 2002):

$$\nabla(D(\mathbf{r})\nabla\Phi(\mathbf{r})) = \frac{\rho(\mathbf{r})}{\epsilon_0}. \tag{2.12}$$

Em geral, um valor de D igual a 80 é utilizado para o exterior da proteína (Baker e McCammon, 2003; Gilson, 2002). Já em seu interior, não há um consenso para o valor de D, tal que diferentes

valores são sugeridos por diferentes autores. Baker e McCammon (2003) sugerem um valor entre 2 e 20, enquanto Gilson (2002) sugere um valor entre 2 e 2.5. Uma vez que a constante dielétrica, $D(\mathbf{r})$, está dentro do escopo do operador de divergência, a descontinuidade dielétrica na superfície da proteína funciona como fonte de linhas de campos elétricos, apesar de possuir densidade de cargas elétricas (reais) igual a zero. Ou seja, o campo elétrico induzido por cargas elétricas remotas (reais), ao atravessar a descontinuidade dielétrica, induz cargas elétricas na superfície da molécula de proteína que passam, então, a atuar como fonte de linhas de campos elétricos.

Uma vez que o meio em que as proteínas se encontram é rico em íons (vide Figura 2.15), estes também precisam ser considerados como cargas elétricas no correspondente modelo eletrostático. Para isto, recorre-se a teoria de Debye-Hückel, que aborda a redistribuição de eletrólitos com mobilidade em solução, devido a interações eletrostáticas entre eles; e utiliza o fator de Boltzmann sobre os íons dissolvidos em um campo eletrostático local ($\exp(-\beta q_i \Phi(\mathbf{r}))$ para estimar a concentração iônica local ($c(r)$) com relação à sua concentração total (c_{bulk}) (Gilson, 2002). Supondo M diferentes tipos de íons, cada um com carga elétrica q_i e concentração $c_i(\mathbf{r})$, o modelo eletrostático considerando a proteína dissolvida em uma solução iônica pode ser descrito por uma equação diferencial não linear, conhecida como equação de Poisson-Boltzmann (Gilson, 2002):

$$\epsilon_0 \nabla(D(\mathbf{r})\nabla\Phi(\mathbf{r})) = \rho(\mathbf{r}) + \sum_{i=1}^{M} q_i c_i(\mathbf{r}) \exp(-\beta q_i \Phi(\mathbf{r})) \tag{2.13}$$

Para resolver a equação de Poisson-Boltzmann para formas geométricas não triviais, como é o caso da superfície da molécula de proteína, são utilizados métodos numéricos que, em geral, são computacionalmente custosos. Permanece, ainda, um tema de pesquisa o desenvolvimento de métodos mais eficientes que viabilizem a solução de sistemas contendo centenas de milhares de átomos (Baker e McCammon, 2003).

2.3.7 Grau de conservação de aminoácidos

Em geral, aminoácidos que desempenham um papel importante na manutenção da estabilidade da estrutura de uma proteína e/ou de sua função estão sujeitos a baixas taxas de mutações ao longo da evolução (Branden e Tooze, 1999) e apresentam um alto grau de conservação.

Para avaliar o grau de conservação de aminoácidos, é utilizado um conjunto de seqüências de proteínas similares, assumidas evolutivamente relacionadas, ou homólogas. Quando existe uma quantidade suficiente de seqüências de proteínas homólogas, é possível produzir um alinhamento múltiplo[4] entre elas e definir uma matriz f, de dimensão $20 \times N$, tal que cada elemento $f(i, j)$ da matriz con-

[4] Alinhamentos múltiplos de seqüências de proteínas podem ser produzidos por diferentes programas que implementam algoritmos específicos para este fim.

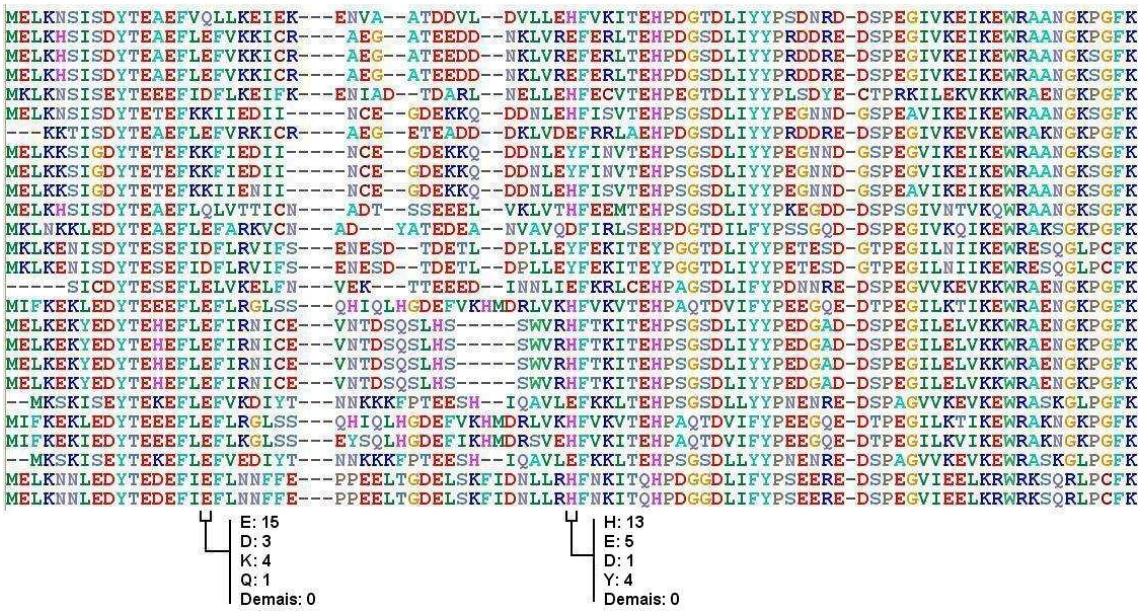

Figura 2.16: Exemplo de alinhamento múltiplo de seqüências de proteínas homólogas: arquivo PDB 7cei, cadeia A. As freqüências de ocorrência de cada tipo de aminoácido para duas posições da seqüência estão em destaque.

tenha a freqüência relativa de ocorrências do aminoácido do tipo i na posição j da seqüência sob estudo (Figura 2.16). Essa matriz é denominada perfil evolucionário da seqüência e a coluna $f(., j)$ é denominada perfil evolucionário do aminoácido na posição j. A partir do perfil evolucionário, medidas diretas do grau de conservação como, por exemplo, a entropia, podem ser derivadas para cada aminoácido da proteína sob estudo. Considerando que a entropia referente a uma coluna de f atinge seu valor máximo quando a freqüência de ocorrência de todos os tipos de aminoácidos é a mesma, sendo dado por $log(20)$, o grau de conservação de aminoácidos na posição j de uma seqüência pode ser medida por (Sander e Schneider, 1991):

$$cons(j) = \frac{\sum_{i=1}^{N} f(i,j)\log(f(i,j))}{\log(20)}. \tag{2.14}$$

Uma alternativa para levar em consideração substituições conservativas[5], consiste em agrupar aminoácidos com propriedades físico-químicas semelhantes (ex: hidrofóbicos, polares, carregados positivamente e carregados negativamente), tal que um valor de N menor que 20 (ex: $N = 4$) é utilizado na fórmula 2.14.

Além disso, uma avaliação mais elaborada do grau de conservação de aminoácidos pode ser obtida ao se considerar a relação evolutiva entre as seqüências alinhadas dada pela correspondente árvore

[5]Substituições conservativas são aquelas em que um aminoácido é substituído por outro com propriedades físico-químicas semelhantes.

filogenética (Durbin et al., 1998). Considerando, para cada posição do alinhamento, uma variável, r_j, que pondera a distância (tempo de evolução) entre os nós da árvore filogenética, Pupko et al. (2002) estimam os valores dos r_{js} por máxima verossimilhança, enquanto Mayrose et al. (2005) utilizam estimação Bayesiana.

2.4 Resumo

Neste capítulo, foram apresentados conceitos básicos sobre estruturas de proteínas. A partir destes conceitos, foi apresentado um conjunto de propriedades físico-químicas e estruturais. Estas propriedades são medidas sobre a estrutura tridimensional da molécula de proteína e são relatadas na literatura como relevantes ao estudo do fenômeno bioquímico de interação proteína-proteína, do ponto de vista da estrutura da proteínas. Nos capítulos 4 e 5, estas propriedades são utilizadas para construir os vetores de características considerados pelos preditores apresentados.

Capítulo 3

Classificação de padrões

Neste capítulo, é apresentado um conjunto de técnicas utilizadas no desenvolvimento de classificadores. Sua função é a de prover a fundamentação metodológica na construção dos preditores descritos nos capítulos 4 e 5. Assim, os leitores familiarizados com estes conceitos podem seguir diretamente para os capítulos 4 e 5 sem prejuízos à sua compreensão.

Ao longo deste capítulo, é considerado que os objetos tratados pelos classificadores são caracterizados por atributos, representados por um vetor de características, e que o conjunto de todos os possíveis vetores de características define um espaço de características. Desenvolver um classificador significa particionar este espaço de características, tal que a cada partição é associada uma classe de objetos. Assim, o objetivo final dos métodos para desenvolvimento de classificadores é realizar esta partição de forma a minimizar o erro de classificação de novos objetos.

Na parte inicial do capítulo, são apresentados métodos para classificação de padrões. Estes foram divididos em dois grupos: os métodos probabilísticos, que utilizam a teoria de decisão de Bayes para determinação da função discriminante; e aqueles que determinam diretamente a função discriminante, otimizando uma função de mérito. A apresentação desses métodos baseia-se em livros textos das áreas de classificação de padrões, análise multivariada e aprendizado de máquina. Em particular, a descrição do primeiro grupo de métodos baseia-se nos textos de Duda et al. (2001) e Mardia et al. (1979), enquanto a descrição dos métodos do segundo grupo baseia-se nos textos de Duda et al. (2001), Cristianini e Shawe-Taylor (2000), Schlkopf e Smola (2001) e Burges (1998).

Em seguida, são apresentados métodos de extração e seleção de características, cujo objetivo é obter uma representação mais compacta dos dados, reduzindo sua dimensionalidade. A apresentação dos métodos de extração de características baseia-se nos textos de Jolliffe (2002), Mardia et al. (1979) e Seber (1984), enquanto a apresentação dos métodos de seleção de características baseia-se nos textos de Webb (2002), Fukunaga (1990) e Guyon e Elisseff (2003).

Finalmente, o capítulo é encerrado com a apresentação das as técnicas mais comumente utilizadas

para avaliar o desempenho de classificadores. Esta parte do capítulo baseia-se nos textos de Webb (2002), Jain et al. (2000), Duda et al. (2001) e Fawcett (2006).

3.1 Métodos probabilísticos

Na teoria de decisão de Bayes, o problema de classificação é colocado em termos probabilísticos e todos os valores de probabilidades relevantes são considerados como conhecidos ou passíveis de estimação.

Seja $\mathbf{x}$ um vetor de características (ou padrão) definido em $\Re^d$, denominado espaço de características. Seja também c um conjunto finito de populações (ou classes), $\{\Pi_1, \ldots, \Pi_c\}$. A cada classe Π_i, está associada uma probabilidade *a priori* $P(\Pi_i)$ [1]. A função [2]

$$K(i,j) = \begin{cases} 0 & se \quad i = j \\ c_{ij} & se \quad i \neq j \end{cases} \tag{3.1}$$

representa o risco (ou perda; ou custo) associado à classificação de um padrão como pertencente a Π_i quando, de fato, ele pertence a Π_j. Se $p(\mathbf{x}|\Pi_i)$ é a função densidade de probabilidades (f.d.p) do vetor $\mathbf{x}$ condicionado a Π_i ($\mathbf{x} \in \Pi_i$), então, utilizando a fórmula de Bayes, a probabilidade *a posteriori*, $P(\Pi_j|\mathbf{x})$, é dada por:

$$P(\Pi_j|\mathbf{x}) = \frac{p(\mathbf{x}|\Pi_j)P(\Pi_j)}{p(\mathbf{x})}, \tag{3.2}$$

onde $p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\Pi_j)P(\Pi_j)$. O fator determinante da probabilidade *a posteriori* $P(\Pi_j|\mathbf{x})$ é o produto $p(\mathbf{x}|\Pi_j)P(\Pi_j)$, uma vez que $p(\mathbf{x})$ representa apenas um fator de escala que garante a validade da somatória $\sum_{j=1}^{c} P(\Pi_j|\mathbf{x}) = 1$. $P(\Pi_j|\mathbf{x})$ representa a probabilidade de Π_j ser a classe a que pertence o padrão representado pelo vetor de características observado, $\mathbf{x}$.

Assim, o risco condicional ao se classificar o padrão $\mathbf{x}$ como pertencente à classe Π_i é dado por:

$$R(\Pi_i|\mathbf{x}) = \sum_{j=1}^{c} K(i,j)P(\Pi_j|\mathbf{x}) \tag{3.3}$$

e a regra de decisão de Bayes consiste em:

1. Calcular o risco condicional $R(\Pi_i|\mathbf{x})$ ao se classificar o padrão, $\mathbf{x}$, como pertencente a cada

[1] Neste texto, a notação com letra minúscula $p(.)$ refere-se à função densidade de probabilidades, enquanto a notação com letra maiúscula $P(.)$ refere-se à função de probabilidades para variáveis discretas.

[2] A função de perda $K(i,j)$ é mais genérica do que a considerada neste texto. Em geral, considera-se um conjunto de ações, $i = 1, \ldots, k$, onde o número de ações k não é, necessariamente, igual ao número de classes c. Aqui, as ações consideradas referem-se, exatamente, à alocação de um padrão a uma das c classes e, portanto, $k = c$.

uma das classes Π_i, $i = 1, \ldots, c$.

2. Classificar o padrão, $\mathbf{x}$, como pertencente à classe Π_i que minimiza $R(\Pi_i|\mathbf{x})$.

3. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

Em particular, quando

$$K(i,j) = \begin{cases} 0 & se \quad i = j \\ 1 & se \quad i \neq j \end{cases} \tag{3.4}$$

a regra de decisão de Bayes maximiza a probabilidade a posteriori $P(\Pi_j|\mathbf{x})$, o que equivale a minimizar a probabilidade de erro de classificação. Este classificador é conhecido como classificador de máxima probabilidade *a posteriori* (MAP) ou de mínimo erro. De fato, se o padrão associado a $\mathbf{x}$ pertence a Π_i, então $R(\Pi_i|\mathbf{x}) = \sum_{j=1, j\neq i}^{c} P(\Pi_j|\mathbf{x}) = 1 - P(\Pi_i|\mathbf{x}) = P(erro|\mathbf{x})$. Mas como $P(erro) = \int_{\infty}^{\infty} P(erro, \mathbf{x}) d\mathbf{x} = \int_{\infty}^{\infty} P(erro|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$, tem-se que $P(erro)$ é minimizado ao se minimizar $P(erro|\mathbf{x})$ que, por sua vez, é minimizado ao se maximizar $P(\Pi_i|\mathbf{x})$.

Em algumas situações, pode ser vantajoso permitir que o procedimento de classificação considere as opções de aceitar ou rejeitar a tarefa de atribuir um padrão a uma classe. Se a informação contida no vetor de características é suficiente para classificar o padrão com um nível razoável de confiança, o padrão é aceito para classificação e é atribuído a uma classe, conforme apresentado acima. Caso o padrão esteja muito próximo da fronteira de decisão, o nível de confiança para realização da classificação é baixo e pode ser vantajoso rejeitar a sua classificação. Neste caso, ele não é atribuído a uma classe, podendo ser simplesmente não classificado, ser classificado manualmente pelo usuário ou ter a decisão sobre sua atribuição a uma classe postergada até que informações adicionais estejam disponíveis. Exemplos em que a utilização da opção de rejeição é vantajosa incluem aplicações em que uma decisão errada implica em um custo muito elevado ou que pode levar a uma situação de perigo.

Uma forma para se incluir a opção de rejeição no processo de classificação consiste em considerar uma classe de rejeição, Π_0 (van der Heijden et al., 2004). A função risco (equação 3.1) é estendida tal que $K(0, j), j = 1, \ldots, c$ representa o risco associado à rejeição de um padrão pertencente à classe Π_j. O classificador com a opção de rejeição é desenvolvido da mesma forma que anteriormente, com a perda esperada também dada pela equação 3.3, onde $i = 0, 1, \ldots, c$. Assim, a regra de decisão de Bayes considerando a opção de rejeição consiste em:

1. Calcular o risco condicional $R(\Pi_i|\mathbf{x})$ ao se classificar o padrão, $\mathbf{x}$, como pertencente a cada uma das classes Π_i, $i = 0, 1, \ldots, c$.

2. Classificar o padrão associado a $\mathbf{x}$ como pertencente à classe Π_i que minimiza $R(\Pi_i|\mathbf{x})$. Neste caso, a atribuição de um objeto à classe Π_0 significa rejeição.

3. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

O classificador MAP também pode ser estendido para considerar a opção de rejeição. Seja t_r o risco de rejeição independente da classe a que o padrão pertence. Também considere que os demais riscos são dados pela equação 3.4. Então, a regra de decisão para o classificador MAP considerando a opção de rejeição consiste em:

1. Calcular o erro, $P(erro_i|\mathbf{x}) = 1 - P(\Pi_i|\mathbf{x})$, de se classificar o padrão, $\mathbf{x}$, incorretamente como pertencente a cada uma das classes, Π_i, $i = 1, \ldots, c$.

2. Denotando o erro mínimo por $erro_{min} = \min_i\{P(erro_i|\mathbf{x})\}$, $i = 1, \ldots, c$, se $t_r < erro_{min}$, rejeitar o padrão.

3. Se $t_r > erro_{min}$ o padrão, $\mathbf{x}$, é classificado como pertencente à classe Π_i correspondente a $erro_{min}$.

4. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

De acordo com esta regra, a máxima probabilidade *a posteriori* é sempre maior que $\frac{1}{c}$ e, portanto, $erro_{min}$ é sempre menor que $1 - \frac{1}{c}$. Assim, a opção de rejeição está ativa somente se $t_r < 1 - \frac{1}{c}$.

3.1.1 Métodos paramétricos e a função densidade de probabilidade normal

Dentre todas as f.d.p's, a mais estudada é a normal multivariada ou gaussiana. Embora esse interesse seja em boa parte resultado da facilidade de tratamento analítico da função gaussiana, ela modela uma importante situação prática (Duda et al., 2001): aquela em que os valores de $\mathbf{x}$, correspondentes aos objetos pertencentes a cada classe Π_i, representam versões corrompidas de um vetor típico, ou protótipo, μ_i. A f.d.p normal multivariada em d dimensões é dada por (Mardia et al., 1979):

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{1/d}|\mathbf{\Sigma}|^{1/2}} \left(-\frac{1}{2}(\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right), \tag{3.5}$$

onde $\mathbf{x}$ é um vetor em $\Re^d$, μ é o vetor de médias, $\mathbf{\Sigma}$ é a matriz de covariâncias de dimensão $d \times d$, $|\mathbf{\Sigma}|$ é seu determinante e $\mathbf{\Sigma}^{-1}$ sua matriz inversa. Note que a f.d.p normal multivariada é completamente determinada pelos parâmetros μ e $\mathbf{\Sigma}$. Assim, por simplicidade, a seguinte notação é utilizada quando uma classe Π_i é caracterizada por vetores de característica d-dimensionais que seguem a distribuição normal multivariada: $p_i(\mathbf{x}) \sim N_d(\mu_i, \mathbf{\Sigma}_i)$.

Geralmente, as funções discriminantes para um classificador que utiliza a regra de Bayes (bayesiano) são definidas como $g_i(\mathbf{x}) = -R(\Pi_i|\mathbf{x}), i = 1, \ldots, c$, e, em particular, por $g_i(\mathbf{x}) = P(\Pi_i|\mathbf{x}), i = 1, \ldots, c$, quando $K(i, j)$ é dado pela equação 3.4. Contudo, este conjunto não é único, pois se $f(.)$

é uma função monótona e crescente, então, $f(g_i(\mathbf{x})), i = 1, \ldots, c$, também define um conjunto de funções discriminantes. Em particular, um conjunto de funções discriminantes bastante utilizado quando $p_i(\mathbf{x}) \sim N_d(\mu_i, \mathbf{\Sigma}_i), i = 1, \ldots, c$, é dada por:

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|\Pi_i)) + \ln(P(\Pi_i)). \tag{3.6}$$

As funções discriminantes dividem o espaço de características em um conjunto de c regiões, $\mathcal{R}_1$,. . .,$\mathcal{R}_c$, tal que a fronteira entre essas regiões é dada pelo lugar geométrico no espaço de características onde mais de uma função discriminante assume o maior valor.

Quando $p(\mathbf{x}) \sim N_d(\mu, \mathbf{\Sigma})$, as funções discriminantes definidas pela equação 3.6 são expressas como:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\mathbf{\Sigma}_i|) + \ln(\Pi_i). \tag{3.7}$$

É interessante analisar as diferentes fronteiras de decisão que resultam de diferentes restrições sobre as matrizes de covariâncias, $\mathbf{\Sigma}_i$. O caso mais simples ocorre quando $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}, i = 1, \ldots, c$, ou seja, os padrões, $\mathbf{x}$, são estatisticamente não correlacionadas e possuem a mesma variância, σ^2. Neste caso, desprezando-se os termos independentes de i na equação 3.7, as funções discriminantes são dadas por:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \mu_i\|^2}{2\sigma^2} + \ln(\Pi_i). \tag{3.8}$$

Expandindo-se o termo quadrático da equação 3.8 e desprezando-se os termos $\mathbf{x}^t\mathbf{x}$, que são iguais para todo i, tem-se que:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}, \tag{3.9}$$

onde: $\mathbf{w}_i = \frac{1}{\sigma^2}\mu_i$ e $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln(\Pi_i)$. Neste caso, as funções discriminantes são lineares, com as fronteiras de decisão representadas por hiperplanos de dimensão $d - 1$ (Figura 3.1) descritos por:

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0, \tag{3.10}$$

onde: $\mathbf{w} = \mu_\mathbf{i} - \mu_\mathbf{j}$ e $\mathbf{x}_0 = \frac{1}{2}(\mu_\mathbf{i} - \mu_\mathbf{j}) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2}\ln\left(\frac{P(\Pi_i)}{P(\Pi_j)}\right)(\mu_\mathbf{i} - \mu_\mathbf{j})$.

Considerando a mesma probabilidade *a priori* para todas as c classes, tem-se que o segundo termo da equação 3.8 pode ser desprezado e as funções discriminantes são dadas pelo negativo do quadrado da distância euclidiana entre o padrão $\mathbf{x}$ e a média μ_i, para cada classe Π_i. Assim, $\mathbf{x}$ é classificado como pertencente à classe cuja distância euclidiana entre $\mathbf{x}$ e $\mu_\mathbf{i}$ é a menor dentre todas as c classes.

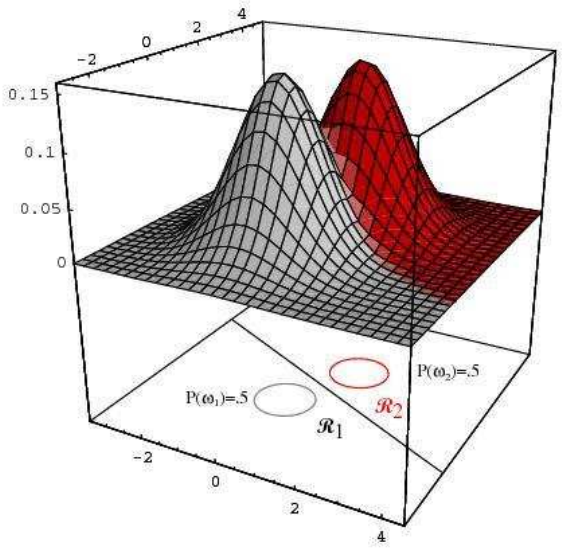

Figura 3.1: Fronteira de decisão para duas classes, Π_1 e Π_2; caso bidimensional com $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \sigma^2\mathbf{I}$. A figura ilustra as f.d.p's $p(\mathbf{x}|\Pi_i)$ e a fronteira de decisão considerando $P(\Pi_1) = P(\Pi_2)$, que corresponde a uma reta (hiperplano de dimensão 1). Adaptado de Duda et al. (2001).

O próximo caso ocorre quando $\mathbf{\Sigma}_i = \mathbf{\Sigma}, i = 1, \ldots, c$, ou seja, a matriz de covariâncias é a mesma para todas as classes. Neste caso, as funções discriminantes assumem a seguinte forma:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_i) + \ln(P(\Pi_i)). \tag{3.11}$$

Considerando as mesmas probabilidades *a priori* para as c classes, o segundo termo pode ser desprezado e a função discriminante é proporcional ao negativo do quadrado da distância de Mahalanobis entre $\mathbf{x}$ e a média μ_i, dada a matriz de covariâncias $\mathbf{\Sigma}$. De maneira análoga ao caso anterior, $\mathbf{x}$ é classificado como pertencente à classe cuja distância de Mahalanobis entre $\mathbf{x}$ e a média μ_i é a menor dentre as c classes.

Expandindo o termo quadrático na equação 3.11 e desprezando o termo $\mathbf{x}^t\Sigma^{-1}\mathbf{x}$, que é independente de i, resulta na função discriminante:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}, \tag{3.12}$$

onde: $\mathbf{w}_i = \mathbf{\Sigma}^{-1}\mu_i$ e $w_{i0} = -\frac{1}{2}\mu_i^t\mathbf{\Sigma}^{-1}\mu_i$. Novamente, como as funções discriminantes são todas lineares, o classificador também é linear (Figura 3.2) e a fronteira entre as regiões $\mathcal{R}_i$ e $\mathcal{R}_j$ é dada pela equação 3.10, mas com os valores de $\mathbf{w}$ e $\mathbf{x}_0$ dados por $\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_i - \mu_j)$ e $\mathbf{x}_0 = \frac{1}{2}(\mu_i - \mu_j) - \frac{\ln(P(\Pi_i)/P(\Pi_j))}{(\mu_i-\mu_j)^t\mathbf{\Sigma}^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)}$, respectivamente.

No caso mais geral, nenhuma restrição é imposta às matrizes de covariâncias, $\mathbf{\Sigma}_i$ = matriz ar-

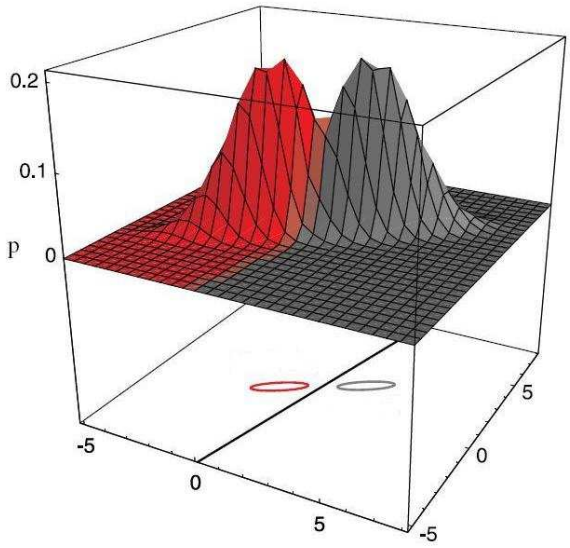

Figura 3.2: Fronteira de decisão para duas classes, Π_1 e Π_2; caso bidimensional com $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. A figura ilustra as f.d.p's $p(\mathbf{x}|\Pi_i)$ e a fronteira de decisão considerando $P(\Pi_1) = P(\Pi_2)$, que corresponde a uma reta. Adaptado de Duda et al. (2001).

bitrária[3]. Neste caso, apenas o segundo termo da equação 3.5 pode ser desprezado, resultando em funções discriminantes quadráticas (Figura 3.3):

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t + w_{i0}, \tag{3.13}$$

onde $\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}$, $\mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\mu_i$ e $w_{i0} = -\frac{1}{2}\mu_i^t\mathbf{\Sigma}^{-1}\mu_i - \frac{1}{2}\ln(|\mathbf{\Sigma_i}|) + \ln(P(\Pi_i))$.

Estimação de Parâmetros

Para construir um classificador do tipo MAP é necessário especificar as probabilidades *a priori*, $P(\Pi_i)$, e as f.d.p's condicionais, $p(\mathbf{x}|\Pi_i)$. Para isso, utiliza-se um conjunto de dados de treinamento, composto por padrões amostrados independentemente e para os quais a correspondente classe é conhecida.

Em geral, a probabilidade *a priori* para cada classe i é estimada como a freqüência relativa de padrões da classe i no conjunto de dados de treinamento. Já os parâmetros que determinam as f.d.p's $p(\mathbf{x}|\Pi_i)$ são estimados utilizando critérios estatísticos mais sofisticados. Dentre estes, o mais utilizado é o de máxima verossimilhança. Se $p_i(\mathbf{x}) \sim N_d(\mu_i, \mathbf{\Sigma}_i)$, o conjunto de parâmetros θ_i é formado pelo vetor de médias, μ_i, e a matriz de covariâncias, $\mathbf{\Sigma}_i$. Para explicitar sua dependência de θ_i, $p(\mathbf{x}|\Pi_i)$ pode ser reescrito como $p(\mathbf{x}|\Pi_i;\theta_i)$.

[3]Apesar de $\mathbf{\Sigma}_i$ ser definido de forma arbitrária, as propriedades gerais de uma matriz de covariâncias continuam a ser obedecidas, ou seja, $\mathbf{\Sigma}_i$ é simétrica e semi-definida positiva.

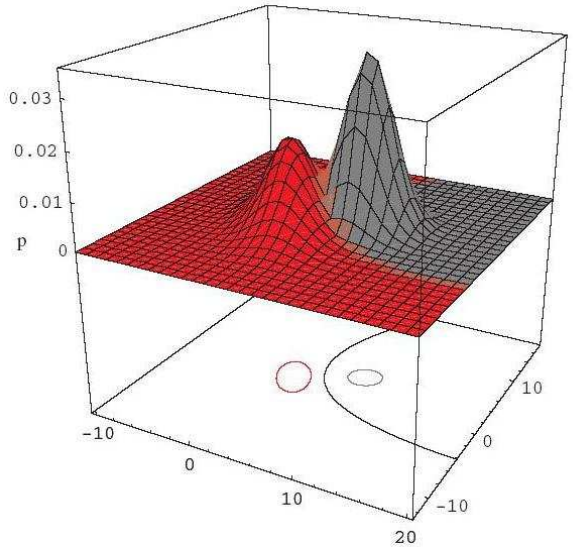

Figura 3.3: Fronteira de decisão para duas classes, Π_1 e Π_2, caso bidimensional com $\mathbf{\Sigma}_1$ e $\mathbf{\Sigma}_2$ arbitrários. A figura ilustra as f.d.p's $p(\mathbf{x}|\Pi_i)$ e a fronteira de decisão considerando $P(\Pi_1) = P(\Pi_2)$, que corresponde a uma parábola. Adaptado de Duda et al. (2001).

A função de verossimilhança para a classe Π_i é definida por Mardia et al. (1979) como:

$$L(\Pi_i; \theta_i) = \prod_{j=1}^{n_i} p(\mathbf{x}_j|\Pi_i; \theta_i). \tag{3.14}$$

onde n_i é o número de padrões em Π_i,

Então, a partir da função de verossimilhança, a função de log-verossimilhança é definida como:

$$l(\Pi_i; \theta_i) = \log(L(\Pi_i; \theta_i)) = \sum_{j=1}^{n_i} \log(p(\mathbf{x}_j|\Pi_i; \theta_i)). \tag{3.15}$$

Quando $p(\mathbf{x}|\Pi_i; \theta_i) \sim N_d(\mu_i, \mathbf{\Sigma}_i)$, utilizando a equação 3.5, a função de log-verossimilhança é dada por:

$$l(\Pi_i; \mu_i, \mathbf{\Sigma}_i) = -\frac{n_i}{2} \log(|2\pi\mathbf{\Sigma}_i|) - \frac{1}{2} \sum_{j=1}^{n_i} (\mathbf{x}_j - \mu_i)^t \mathbf{\Sigma}_i^{-1} (\mathbf{x}_j - \mu_i) \tag{3.16}$$

e tem-se:

$$l(\Pi_i; \mu_i, \mathbf{\Sigma}_i) = -\frac{n_i}{2} \log(|2\pi\mathbf{\Sigma}_i|) - \frac{1}{2} tr(\mathbf{\Sigma}_i^{-1}\mathbf{S}_i) - \frac{n_i}{2} (\bar{\mathbf{x}}_i - \mu_i)^t \mathbf{\Sigma}_i^{-1} (\bar{\mathbf{x}}_i - \mu_i), \tag{3.17}$$

onde: $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{i=1}^{n_i} \mathbf{x}_i$ e $\mathbf{S}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)^t (\mathbf{x}_j - \bar{\mathbf{x}}_i)$.

A estimativa de máxima verossimilhança determina os valores dos parâmetros, θ_i, que maximizam

a função de verossimilhança, ou, de forma equivalente, a função de log-verossimilhança. No caso da f.d.p normal multivariada, estes valores correspondem aos valores de de μ_i e $\mathbf{\Sigma}_i$. Mardia et al. (1979) demonstram que μ_i e $\mathbf{\Sigma}_i$ podem ser estimados por:

$$\hat{\mu}_i = \bar{\mathbf{x}}_i \tag{3.18}$$

e

$$\hat{\mathbf{\Sigma}}_i = \mathbf{S}_i, \tag{3.19}$$

onde $\bar{\mathbf{x}}_i$ é o vetor de média amostral e $\mathbf{S}_i$ a matriz de covariância amostral da classe Π_i.

3.1.2 Métodos não paramétricos

Em muitas aplicações de classificação, a suposição de que as f.d.p's envolvidas possuem uma forma parametrizada não é válida, tal que métodos não paramétricos para estimação de f.d.p's precisam ser utilizados. Esses métodos são utilizados para inferir, a partir de um conjunto de dados, f.d.p's cuja forma funcional não tenha sido especificada.

A idéia básica por trás destes métodos é muito simples (Duda et al., 2001; Bishop, 1997). Seja $p(\mathbf{x})$ uma f.d.p desconhecida, $\mathbf{x}$ um novo padrão amostrado a partir de $p(\mathbf{x})$ e ***R*** uma região no espaço de $\mathbf{x}$. Então, por definição, a probabilidade de que $\mathbf{x}$ esteja contido em ***R*** é dada por:

$$P = \int_R p(\mathbf{x}) d\mathbf{x}. \tag{3.20}$$

$p(\mathbf{x})$ é uma função contínua e a equação 3.20 pode ser utilizada para estimar valores específicos de $p(\mathbf{x})$ se $\mathbf{x}$ está contido em ***R***.

Considerando um conjunto de n padrões, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, amostrados de forma independente a partir de $p(\mathbf{x})$ (i.i.d), então a probabilidade de k padrões estarem contidos em R é dada pela distribuição de probabilidades binomial (Duda et al., 2001):

$$P_k = \begin{pmatrix} n \\ k \end{pmatrix} P^k (1-P)^{n-k}. \tag{3.21}$$

De acordo com a distribuição de probabilidades binomial, a proporção média de pontos contidos em R é dada por $E\left[\frac{k}{n}\right] = P$ e a variância em torno dessa média é dada por $E\left[\left(\frac{k}{n} - P\right)^2\right] = \frac{P(1-P)}{n}$. A distribuição de probabilidades binominal apresenta um pico em torno da média, tal que, à medida que o número de dados aumenta, a variância em torno desse pico tende para zero, tornando-o mais saliente. Assim, uma boa estimativa da probabilidade P é dada pela fração de padrões contidos em R:

$$P \cong \frac{k}{n}. \tag{3.22}$$

Considerando que $p(\mathbf{x})$ é uma função contínua, cujo valor praticamente não varia se $\mathbf{x}$ está contido em R, a equação 3.20 pode ser reescrita como:

$$P = \int_R p(\mathbf{x}) d\mathbf{x} \cong p(\mathbf{x}')V, \tag{3.23}$$

onde V é o hipervolume de R e $\mathbf{x}'$ representa um ponto no interior de R. Combinando 3.22 e 3.23, obtém-se a seguinte estimativa para $p(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{V}\frac{k}{n}. \tag{3.24}$$

A validade das equações acima pressupõe algumas condições relacionadas à escolha da região R. Enquanto a precisão da equação 3.22 é garantida quando R é relativamente grande, tal que ele contenha um grande número de pontos, a precisão da equação 3.23 é garantida fazendo R tão pequeno quanto possível. Já para a equação 3.24, considerando $p_n(\mathbf{x})$, k_n e V_n dependentes de n, as seguintes condições garantem a convergência de $p_n(\mathbf{x})$ para $p(\mathbf{x})$ a medida que $n \to \infty$ (Duda et al., 2001):

- $\lim_{n\to\infty} V_n = 0$;
- $\lim_{n\to\infty} k_n = \infty$;
- $\lim_{n\to\infty} \frac{k_n}{n} = 0$.

Método de Parzen

O método de Parzen consiste em estimar $p(\mathbf{x}|\Pi)$ fixando o hipervolume da região R e permitindo que o número de pontos na região, k, varie. Se a região R é definida como um hipercubo d-dimensional de lado h e centrado em x, então o hipervolume de R é dado por $V = h^d$. É possível, então, obter uma expressão analítica que permite computar o número de pontos contidos em R utilizando a seguinte função janela:

$$\Phi(\mathbf{u}) = \begin{cases} 1 & se\ |u_j| \leq \frac{1}{2}, \quad j = 1, 2, \ldots, d; \\ 0 & em\ caso\ contrário. \end{cases} \tag{3.25}$$

Para todo ponto $\mathbf{y}$, $\Phi((\mathbf{x}-\mathbf{y})/h)$ é igual a 1 se $\mathbf{y}$ está contido na região delimitada pelo hipercubo centrado em $\mathbf{x}$ e zero, caso contrário. Portanto, o número total de pontos contidos no hipercubo é dado por:

$$k = \sum_{j=1}^{n} \Phi\left(\frac{\mathbf{x} - \mathbf{x}_j}{h}\right). \tag{3.26}$$

Utilizando a equação 3.24, obtém-se a seguinte estimativa para $p(\mathbf{x})$:

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{V^j} \Phi\left(\frac{\mathbf{x} - \mathbf{x}_j}{h}\right). \tag{3.27}$$

A equação 3.27 também pode ser vista como uma sobreposição de hipercubos centrados nos n padrões de treinamento. Ela também sugere a utilização de uma classe mais geral de funções $\Phi(\mathbf{u})$. De fato, qualquer f.d.p constitui uma função $\Phi(\mathbf{u})$ válida ou, mais precisamente, a validade da função $\Phi(\mathbf{u})$ é garantida se as seguintes condições são satisfeitas (Duda et al., 2001):

- $\Phi(\mathbf{u}) \geq 0$ e
- $\int \Phi(\mathbf{u}) d\mathbf{u} = 1$.

Uma escolha comum para a função $\Phi(\mathbf{u})$ é a f.d.p normal multivariada. Neste caso, a equação 3.27 assume a seguinte forma:

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{(2\pi h^2)^{\frac{d}{2}}} \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_j\|}{2h^2}\right). \tag{3.28}$$

Finalmente, uma propriedade importante do método de Parzen pode ser obtida ao ser calcular a esperança da f.d.p estimada, $E[\hat{p(\mathbf{x})}]$, em um ponto $\mathbf{x}$, obtida em relação às diferentes possibilidades de seleção de padrões do conjunto de dados. Utilizando a equação 3.27 tem-se que:

$$E[\hat{p(\mathbf{x})}] = \frac{1}{n} \sum_{j=1}^{n} E\left[\frac{1}{V^n} \Phi_n\left(\frac{\mathbf{x} - \mathbf{x}_j}{h_n}\right)\right] = \int \frac{1}{V^n} \Phi_n\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) p(\mathbf{y}) d\mathbf{y} = \int \delta_n(\mathbf{x} - \mathbf{y}) p(\mathbf{y}) d\mathbf{y}, \tag{3.29}$$

onde $\delta_n(\mathbf{x} - \mathbf{y}) = \frac{1}{V^n} \Phi_n\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)$.

A equação 3.29 mostra que a esperança da f.d.p estimada, $E[\hat{p(\mathbf{x})}]$, é a convolução [4] da f.d.p verdadeira, mas desconhecida, $p(\mathbf{x})$, e a função janela, $\Phi_n(\mathbf{u})$. O parâmetro h desempenha a função de suavização da f.d.p estimada. Se o número de padrões, n, é tal que $n \to \infty$, à medida que $h \to 0$, então, $\Phi_n(\mathbf{u})$ aproxima a função delta de Dirac e $\hat{p}(\mathbf{x})$ aproxima a f.d.p verdadeira (Duda et al., 2001). Contudo, se n é finito, a escolha de um h muito pequeno leva a uma aproximação ruidosa para a f.d.p, que no limite corresponderá a um conjunto de funções delta de Dirac, cada uma centrada em um padrão.

[4] A convolução entre duas função $f(u)$ e $g(u)$ é definida como $\int f(u - y) g(y) dy$.

Método dos K vizinhos mais próximos

Um dos problemas potenciais do método de Parzen é a determinação do valor ótimo do parâmetro h. Se o valor de h é muito elevado, a estimativa da f.d.p em algumas regiões pode ser excessivamente suavizada, tal que detalhes sobre o seu comportamento nesta região são perdidos. Já se o valor de h é muito pequeno, a f.d.p estimada em uma região de baixa densidade de dados pode se tornar muito ruidosa. Portanto, a escolha do valor ótimo de h depende da região do espaço de características que contém $\mathbf{x}$.

O método dos k vizinhos mais próximos (ou kNN) tenta minimizar este problema fixando o número de padrões k contidos na região R. Ao invés de fixar o hipervolume V associado à R como no método de Parzen, o método kNN considera uma hiperesfera centrada em $\mathbf{x}$ contendo exatamente k padrões e estima a f.d.p para o ponto $\mathbf{x}$ utilizando a equação 3.24, onde V é o hipervolume da hiperesfera.

Da mesma forma que os demais métodos para estimação de f.d.p's, o método kNN também pode ser utilizado para estimar $p(\mathbf{x}|\Pi_i)$ para cada classe Π_i, tal que um classificador pode ser construído utilizando o teorema de Bayes. Entretanto, também é possível obter uma regra de decisão específica para o método kNN.

Considere um conjunto de dados com c classes, $\Pi_1, \ldots, \Pi_c$, tal que cada classe contenha n_i padrões e $\sum_{i=1}^{c} n_i = n$. Considere também um novo padrão $\mathbf{x}$ e uma hiperesfera com centro em $\mathbf{x}$ contendo em seu volume k pontos pertencentes a quaisquer das classes consideradas. Seja V o hipervolume dessa hiperesfera e k_i o número de padrões da classe Π_i nela contidos. Utilizando a equação 3.24, obtém-se que $p(\mathbf{x}|\Pi_i) = \frac{k_i}{n_i V}$ e $p(\mathbf{x}) = \frac{k}{nV}$. Já as probabilidades *a priori* para cada classe podem ser estimadas como $P(\Pi_i) = \frac{n_i}{n}$. Então, pelo teorema de Bayes:

$$P(\Pi_i|\mathbf{x}) = \frac{p(\mathbf{x}|\Pi_i)P(\Pi_i)}{p(\mathbf{x})} = k_i/k \tag{3.30}$$

A equação 3.30 mostra que dados os k vizinhos mais próximos de um padrão, a probabilidade *a posteriori*, $P(\Pi_i|\mathbf{x})$, pode ser estimada diretamente da proporção de padrões pertencentes à classe Π_i dentre os k padrões considerados. Portanto, um novo padrão $\mathbf{x}$ é classificado com a menor probabilidade de erro se ele é considerado como pertencente à classe Π_i, correspondente à máxima razão k_i/k. Isto resulta no seguinte procedimento, conhecido como a regra de decisão dos k vizinhos mais próximos:

1. Determine os k padrões mais próximos de $\mathbf{x}$.
2. Determine a razão k_i/k para cada classe $\Pi_i, \quad i = 1, \ldots, c$.
3. Atribua o padrão $\mathbf{x}$ à classe Π_i que apresenta a maior razão k_i/k.

4. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

3.2 Métodos baseados na determinação de funções discriminantes

O objetivo final dos métodos de classificação é dividir o espaço de características de acordo com o conjunto de classes conhecidas, tal que novos padrões são classificados de acordo com essa divisão. Para isso, os classificadores probabilísticos utilizam um conjunto de dados de treinamento para estimar as f.d.p's condicionais para cada classe e o teorema de Bayes para obter as funções discriminantes que determinam a divisão do espaço de características. Uma segunda abordagem para classificação de padrões consiste em expressar as funções discriminantes através de uma forma paramétrica conhecida e estimar os valores de seus parâmetros diretamente do conjunto de dados de treinamento.

3.2.1 Funções discriminantes lineares

Uma das classes de funções discriminantes mais importante é formada pelas funções discriminantes lineares. Esse tipo de função discriminante possui uma série de propriedades analíticas interessantes e, em alguns casos, pode representar o classificador ótimo. Mesmo nos casos em que ele não é ótimo, os classificadores lineares ainda são atraentes, pois sua simplicidade, muitas vezes, compensa uma eventual perda de acuidade do classificador. Além disso, na ausência de informações que sugiram um tipo de função discriminante diferente, funções discriminantes lineares são uma alternativa atraente para uma primeira tentativa de classificação.

Uma função discriminante linear é uma combinação linear de componentes de $\mathbf{x}$, podendo ser escrita da seguinte forma:

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0, \tag{3.31}$$

onde $\mathbf{w}$ representa um vetor de pesos e w_0 um limiar (ou *bias*).

Quando $c = 2$, apenas uma função discriminante é especificada e a regra de decisão consiste em atribuir $\mathbf{x}$ à classe Π_1 se $g(\mathbf{x}) > 0$ e à classe Π_2 em caso contrário. Se $g(\mathbf{x}) = 0$, a regra é indefinida e $\mathbf{x}$ pode ser atribuído a qualquer das duas classes indiferentemente.

Neste caso, a superfície de decisão representa o lugar geométrico do espaço de características que separa os padrões classificados como pertencentes às classes Π_1 e Π_2. Sejam $\mathbf{x}_1$ e $\mathbf{x}_2$ tal que $g(\mathbf{x}_1) = g(\mathbf{x}_2) = 0$, ou seja, ambos estão sobre a superfície de decisão. Então, $\mathbf{w}^t\mathbf{x}_1 + w_0 = \mathbf{w}^t\mathbf{x}_2 + w_0 = 0$, ou seja:

$$\mathbf{w}^t(\mathbf{x}_1 - \mathbf{x}_2) = 0. \tag{3.32}$$

A equação 3.32 mostra que a superfície de decisão é um hiperplano, aqui denotado por H, no espaço de características e que o vetor de pesos $\mathbf{w}$ é ortogonal a ele. Ela também mostra que o vetor $\mathbf{w}$ está direcionado para a região correspondente ao padrão $\mathbf{x}_1$, Π_1.

Seja, $\mathbf{x}_p$ a projeção ortogonal de $\mathbf{x}$ sobre H e r a distância algébrica (considerando o sinal) entre $\mathbf{x}$ e H. Então, $\mathbf{x}$ pode ser expresso como:

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}. \tag{3.33}$$

Mas como $g(\mathbf{x}) = \mathbf{w}^t\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0$, tem-se que:

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}. \tag{3.34}$$

A equação 3.34 evidencia uma propriedade da função discriminante linear com duas classes: a função discriminante fornece uma medida algébrica da distância de um padrão $\mathbf{x}$ para o hiperplano que representa a superfície de decisão, positiva se $\mathbf{x}$ pertence a Π_1 e negativa se $\mathbf{x}$ pertence a Π_2.

No caso geral, quando o número de classes é maior que dois, $c > 2$, define-se uma função discriminante para cada classe:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}, \quad i = 1, \ldots, c \tag{3.35}$$

e a regra de decisão consiste em classificar $\mathbf{x}$ como pertencente a Π_i se $g_i(\mathbf{x}) > g_j(\mathbf{x})$, para todo $j \neq i$. Em caso de empate, a decisão é considerada indefinida. Se o padrão $\mathbf{x}$ está sobre a superfície de decisão entre $\mathcal{R}_i$ e $\mathcal{R}_j$, $g_i(\mathbf{x}) = g_j(\mathbf{x})$, tem-se que:

$$(\mathbf{w}_i - \mathbf{w}_j)^t\mathbf{x} + (w_{i0} - w_{j0}) = 0. \tag{3.36}$$

Assim, da mesma forma que no caso com duas classes, a superfície de decisão é um hiperplano, aqui denotado por H_{ij}, e o vetor de diferença de pesos $(\mathbf{w}_i - \mathbf{w}_j)$ é ortogonal a H_{ij}. Além disso, a distância algébrica de $\mathbf{x}$ para H_{ij} é dada por:

$$r = \frac{g_i(\mathbf{x}) - g_j(\mathbf{x})}{\|\mathbf{w}_i - \mathbf{w}_j\|}. \tag{3.37}$$

Discriminante linear de Fisher

O critério de classificação de Fisher se aplica a problemas de duas classes e consiste em encontrar a direção de projeção que maximize a razão entre as dispersões (ou soma de quadrados das variâncias) interclasses e intraclasses. O classificador que resulta é conhecido como discriminante linear de

Fisher (ou análise discriminante linear - LDA), sendo um dos exemplos mais conhecidos de classificador linear.

Seja $y = \mathbf{w}^t\mathbf{x}$ a projeção do vetor $\mathbf{x}$ na direção definida por $\mathbf{w}$. Então, a seguinte relação entre as médias amostrais de y e $\mathbf{x}$ é válida:

$$\tilde{\mu}_i = \mathbf{w}^t\hat{\mu}_i, \quad i = 1,\ 2, \tag{3.38}$$

onde $\tilde{\mu}_i$ é a média amostral de y e $\hat{\mu}_i$ é o vetor de média amostral de $\mathbf{x}$, ambos relativos à classe Π_i.

Seja a dispersão intraclasse definida como:

$$\tilde{s}_1^2 + \tilde{s}_2^2, \tag{3.39}$$

onde

$$\tilde{s}_i^2 = \sum_{y|y=\mathbf{w}^t\mathbf{x}\ e\ \mathbf{x}\in\Pi_i} (y - \tilde{\mu}_i)^2 \tag{3.40}$$

representa a dispersão dos padrões projetados para cada classe $\Pi_i, i = 1, 2$.

Utilizando as equações 3.38 e 3.40, obtém-se que $\tilde{s}_i^2 = \mathbf{w}^t\mathbf{S}_i\mathbf{w}$, onde $\mathbf{S}_i$ representa a matriz de dispersão para cada classe $\Pi_i, i = 1, 2$, dado pelo produto externo $\mathbf{S}_i = (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^t$. Assim, a dispersão intraclasse total é dada por:

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t\mathbf{S}_w\mathbf{w}, \tag{3.41}$$

onde $\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2$.

De maneira similar, uma vez que a distância entre as duas médias projetadas é dada por $|\tilde{\mu}_1 - \tilde{\mu}_2| = |\mathbf{w}^t(\hat{\mu}_1 - \hat{\mu}_2)|$, tem-se que:

$$|\tilde{\mu}_1 - \tilde{\mu}_2|^2 = \mathbf{w}^t\mathbf{S}_B\mathbf{w}, \tag{3.42}$$

onde $\mathbf{S}_B = (\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)^t$ representa a dispersão entre classes considerando os padrões não projetados.

Finalmente, a partir das equações 3.41 e 3.42, o critério de Fisher consiste em maximizar a seguinte função:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_w\mathbf{w}}. \tag{3.43}$$

Considerando que $\mathbf{S}_w$ seja inversível, o valor de $\mathbf{w}$ que maximiza $J(\mathbf{w})$ é dado por (Mardia et al., 1979; Duda et al., 2001):

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \mathbf{S}_w^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \tag{3.44}$$

e a regra de decisão para o discriminante linear de Fisher consiste em:

1. Determinar $\mathbf{w}$ que maximiza $J(\mathbf{w})$.
2. Decida por Π_1 se $|\mathbf{w}^t(\mathbf{x} - \hat{\mu}_1)| < |\mathbf{w}^t(\mathbf{x} - \hat{\mu}_2)|$; e por Π_2 se $|\mathbf{w}^t(\mathbf{x} - \hat{\mu}_1)| > |\mathbf{w}^t(\mathbf{x} - \hat{\mu}_2)|$.
3. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

Ou seja, a regra de decisão para o discriminante linear de Fisher consiste em atribuir o padrão, $\mathbf{x}$, à classe cuja média projetada, $\tilde{\mu}_i$, esteja mais próxima de sua própria projeção, y, de acordo com a distância euclidiana.

3.2.2 Funções discriminantes não lineares

Métodos baseados em funções discriminantes não lineares buscam relaxar a restrição de linearidade da função discriminante. Uma das estratégias mais bem sucedidas para obter funções discriminantes não lineares consiste em utilizar uma função discriminante linear em um espaço de características correspondente a uma projeção não linear dos padrões originais. Esta estratégia pode ser representada por uma função discriminante linear generalizada, dada por (Duda et al., 2001):

$$g(\mathbf{x}) = \sum_{i=1}^{d'} w_i \psi_i(\mathbf{x}) + b, \tag{3.45}$$

onde $\psi_i(\mathbf{x}), i = 1, \ldots, d'$, representam funções não lineares arbitrárias, $\mathbf{x} = (x_1, \ldots, x_d)$ é um vetor d-dimensional, $\mathbf{w} = (w_1, \ldots, w_d')$ é um vetor de pesos d'-dimensional e b, $b \in \Re$, representa um termo de *bias*. O ponto crucial ao se utilizar esta abordagem é a determinação das funções não lineares $\psi_i(\mathbf{x})$. Os parâmetros da função discriminante linear generalizada e a não linearidade apropriada devem ser determinados simultaneamente (Duda et al., 2001).

A regra de decisão para métodos baseados em funções discriminantes lineares generalizadas, com duas classes, é dada por:

1. Decida por Π_1 se $\mathbf{w}^t\psi(\mathbf{x}) + b > 0$, onde ψ representa um mapeamento não linear $\psi : d \rightarrow d'$, $\mathbf{w}$ representa os parâmetros do classificador e $\mathbf{x}$ é o padrão a ser classificado.
2. Da mesma forma, decida por Π_2 se $\mathbf{w}^t\psi(\mathbf{x}) + b < 0$.
3. Em caso de empate, a regra é indefinida e a atribuição de $\mathbf{x}$ é realizada de forma arbitrária.

Máquinas de vetores-suporte

Classificadores baseados em máquinas de vetores-suporte (SVM) são classificadores lineares generalizados. Eles realizam uma classificação linear em um espaço de características de alta dimensionalidade, induzido por uma função não linear associada a uma função especial, denominada *kernel*. Essa estratégia baseia-se em um resultado obtido por Cover (1965), que afirma que "um problema de classificação de padrões projetado em um espaço de alta dimensão, através de uma função não linear, possui uma maior probabilidade de ser separável linearmente que em um espaço de baixa dimensão".

Os classificadores SVM podem ser entendidos como métodos não paramétricos, uma vez que nenhuma suposição sobre a forma da distribuição dos dados é assumida. Uma vez que o treinamento do classificador é realizado com base em um conjunto de dados finito, existe o risco de ocorrência de um fenômeno conhecido como sobre-ajuste, em que o classificador se ajusta ao conjunto de dados específico utilizado no treinamento e apresenta uma capacidade de generalização baixa. Para controlar este risco e garantir uma boa generalização os classificadores SVM implementam o princípio de minimização do risco estrutural, oriundo da teoria de aprendizado estatístico (Vapnik, 1997). Este princípio envolve a minimização de limitantes superiores do erro de generalização, o que implica na maximização da capacidade de generalização do classificador. Além disso, o algoritmo de treinamento de classificadores SVM é formulado como um problema de otimização convexo[5], que apresenta a conveniente propriedade de possuir um único ótimo (global).

A formulação mais popular de um classificador SVM utiliza classificadores de máxima margem. Seja um conjunto de padrões de treinamento $\mathbf{X} = (\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)$, onde $\mathbf{x}_i \in \Re^d$ e $y_i \in \{-1, +1\}$. Considerando que os dados em $\mathbf{X}$ são separáveis por um hiperplano H definido pelos parâmetros $\mathbf{w}$ e b, define-se a margem (funcional) de um padrão de treinamento $(\mathbf{x}_i, y_i)$ com respeito ao hiperplano H como sendo a quantidade

$$\gamma_i = y_i(\mathbf{w}^t\mathbf{x}_i + b). \tag{3.46}$$

Além disso, se H representa a superfície de decisão de um classificador linear, o padrão $\mathbf{x}$ é classificado corretamente se $\gamma_i > 0$. A margem (funcional) de H é definida como a menor margem associada a um padrão de treinamento, $\mathbf{x}_i$.

Se o hiperplano H é normalizado, ou seja, definido pelos parâmetros $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ e $\frac{b}{\|\mathbf{w}\|}$, diz-se que a margem (equação 3.46) é geométrica e mede a distância euclidiana entre os padrões e o hiperplano H (vide também equação 3.34). Por fim, a margem associada ao conjunto de treinamento $\mathbf{X}$ é definida como a margem geométrica máxima ao considerar todos os possíveis hiperplanos de separação. Este hiperplano é denominado hiperplano de máxima margem e define um classificador linear de máxi-

[5]Um problema de otimização convexo possui função objetivo convexa e restrições que definem uma região factível convexa.

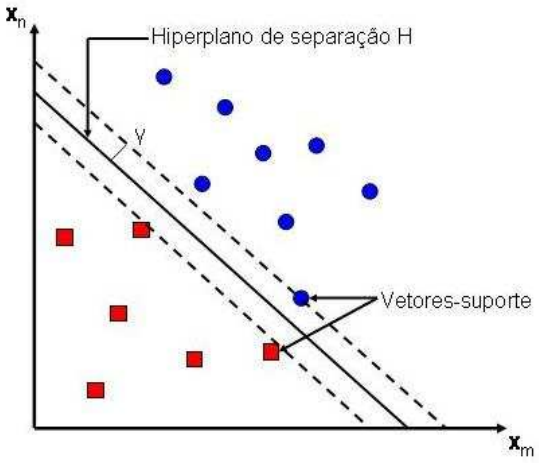

Figura 3.4: Hiperplano de separação, vetores-suporte e margem de separação, γ. Caso separável linearmente.

ma margem, enquanto os vetores que a definem são denominados vetores-suporte do hiperplano H (Figura 3.4).

Nas duas próximas seções, o classificador de máxima margem é visto como um classificador linear generalizado. Considera-se um mapeamento $\mathbf{z} = \varphi(\mathbf{x})$, onde $\mathbf{x}$ é um padrão de treinamento e $\mathbf{z}$ é o padrão no espaço de características induzido pelo mapeamento não linear $\varphi(.)$.

Caso separável linearmente: Seja o hiperplano H que define a fronteira de decisão de um classificador linear de máxima margem, definido por:

$$H : g(\mathbf{z}) =< \mathbf{w}, \mathbf{z} > +b = 0, \tag{3.47}$$

onde $< ., . >$ representa um produto interno. Se os padrões de treinamento $\mathbf{Z} = (\mathbf{z}_1, y_1)..., (\mathbf{z}_n, y_n)$ são separáveis linearmente com margem γ, então, para todo padrão $\mathbf{z}_i$ pertencente a $\mathbf{Z}$, tem-se que $y_i g(\mathbf{z}_i) \geq \gamma, i = 1, \ldots, n$. Além disso, tem-se que:

$$\gamma_i \frac{g(\mathbf{z}_i)}{\|\mathbf{w}\|} \geq \frac{\gamma}{\|\mathbf{w}\|} = \gamma_g, \quad i = 1, \ldots, n, \tag{3.48}$$

onde γ_g representa a margem geométrica de $\mathbf{Z}$. Qualquer conjunto de parâmetros $(\lambda\mathbf{w}, \lambda b)$, tal que λ pertença a $\Re^+$, também define um hiperplano de separação (equação 3.47) e representa um grau de liberdade sobre o valor da margem funcional γ. Quando a representação de H é tal que a margem funcional, γ, é igual a 1, a representação que resulta é denominada canônica. Utilizando esta representação, a margem geométrica de Z, γ_g, depende apenas do valor de $\mathbf{w}$, sendo dada por:

$$\frac{1}{\|\mathbf{w}\|} = \gamma_g. \tag{3.49}$$

A equação 3.49 sugere que maximizar a margem geométrica, γ_g, é equivalente a minimizar a norma do vetor $\mathbf{w}$, o que, por sua vez, sugere a seguinte formulação do problema de otimização correspondente ao treinamento do classificador SVM:

$$\begin{array}{ll} \min_{\mathbf{w},b} & \frac{1}{2} < \mathbf{w}, \mathbf{w} > \\ s.a & y_i(< \mathbf{w}, \mathbf{z}_i > +b) \geq 1, \quad i = 1, \ldots, n, \end{array} \tag{3.50}$$

onde o fator $\frac{1}{2}$ na função objetivo foi acrescentado apenas por conveniência. O problema de otimização 3.50 representa um problema de otimização quadrático e, portanto, um problema convexo, para o qual existe um único mínimo (global) (Bazaraa et al., 1993).

Entretanto, para fins de treinamento de classificadores SVM, é mais conveniente resolver o problema 3.50 utilizando sua formulação dual. Assim, seja a função lagrangeana:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} < \mathbf{w}, \mathbf{w} > - \sum_{i=1}^{n} \alpha_i \left\{ y_i(< \mathbf{w}, \mathbf{z}_i > +b) - 1 \right\}, \tag{3.51}$$

onde $\alpha_i \geq 0, i = 1, \ldots, n$ são os multiplicadores lagrangeanos.

Diferenciando $L(\mathbf{w}, b, \alpha)$ em relação a $\mathbf{w}$ e b e impondo a condição de estacionariedade, obtém-se as seguintes relações:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{z}_i, \tag{3.52}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0. \tag{3.53}$$

Agora, substituindo as relações 3.52 e 3.53 em 3.51, obtém-se:

$$L(\mathbf{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j < \mathbf{z}_i, \mathbf{z}_j > . \tag{3.54}$$

Se o valor de b é fixo, a relação 3.53 não é considerada (ex: se $b = 0$, o problema se restringe a hiperplanos que passam pela origem) e o número de graus de liberdade do problema é reduzido de um.

A forma dual do problema 3.50 consiste em maximizar 3.54 sujeito à restrição de positividade de $\alpha_i, i = 1, \ldots, n$, que acrescido da restrição 3.53 resulta na formulação dual para treinamento de um classificador SVM para o caso em que os dados são separáveis linearmente:

$$\begin{array}{ll} \max_{\alpha} & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j < \mathbf{z}_i, \mathbf{z}_j > \\ s.a & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, \ldots, n. \end{array} \tag{3.55}$$

Os valores de $\mathbf{w}$ e da margem geométrica, γ_g, são obtidos através das relações 3.52 e 3.49, respectivamente. O valor de b envolve a utilização de restrições do problema primal, uma vez que ele não aparece na formulação dual. Uma possibilidade consiste em utilizar a seguinte relação (Cristianini e Shawe-Taylor, 2000):

$$b = -\frac{1}{2} \left(max_{y_i=-1} < \mathbf{w}, \mathbf{z}_i > \quad + \quad min_{y_i=+1} < \mathbf{w}, \mathbf{z}_i > \right). \tag{3.56}$$

Outra consiste em utilizar o seguinte conjunto de relações, que fazem parte das condições de otimalidade de Karush-Kuhn-Tucker (KKT) (Bazaraa et al., 1993):

$$\alpha_i (y_i (\mathbf{w}^t \mathbf{z}_i + b) - 1) = 0, \quad i = 1, \ldots, n. \tag{3.57}$$

Neste caso, considera-se o valor médio obtido através da solução de b para cada uma das equações em que $\alpha_i > 0$ (Burges, 1998). No problema 3.55, existe um multiplicador lagrangeano para cada padrão de treinamento. A satisfação da condição de KKT, expressa pela relação 3.57, implica que apenas os padrões, $\mathbf{z}_i$, mais próximos do hiperplano de decisão ($\gamma = 1$) estão associados a multiplicadores lagrangeanos diferentes de zero, $\alpha_i \neq 0$, e definem os vetores-suporte do hiperplano de separação. Um classificador SVM treinado com um conjunto de padrões de treinamento composto apenas pelos vetores-suporte, possui o mesmo hiperplano de separação que um classificador SVM treinado com o conjunto completo de padrões. Finalmente, utilizando as relações 3.47 e 3.52, a representação dual do hiperplano de separação é dada por:

$$H : g(\mathbf{z}) = \sum_{i \in SV}^{n} y_i \alpha_i < \mathbf{z}_i, \mathbf{z} > + b, \tag{3.58}$$

onde SV denota o conjunto de vetores-suporte associado ao conjunto de padrões de treinamento.

Caso não separável linearmente: O classificador de máxima margem apresentado acima é ilustrativo dos conceitos envolvidos na formulação de um classificador SVM. Entretanto, a suposição de que os dados são separáveis linearmente é um limitante. Contudo, numa situação real, é esperado que os dados contenham algum nível de ruído, podendo não ser separáveis linearmente. Assim, esta seção introduz o conceito de margem suave (Figura 3.5) para o caso em que os dados não são separáveis linearmente.

A utilização de uma margem suave permite que o classificador de máxima margem tolere alguns

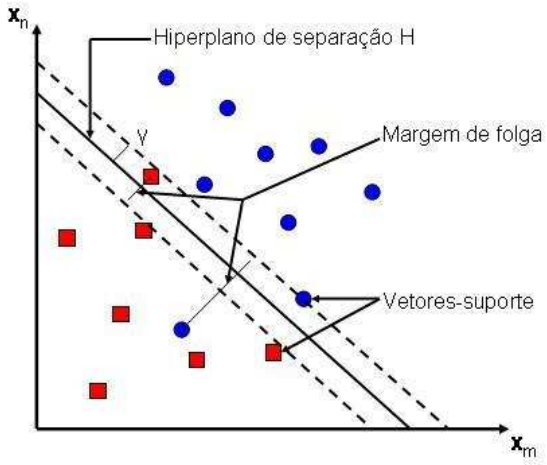

Figura 3.5: Definições de margem de folga. Caso não separável linearmente.

padrões a uma distância inferior à margem considerada, podendo, inclusive, conter dados com rótulos incorretos. Para acomodar essa situação, a formulação do problema primal (problema 3.50) deve ser alterada de forma a relaxar o conjunto de restrições. Formalmente, isso é feito pela introdução de variáveis de folga, $\xi_i \geq 0, i = 1, \ldots, n$, tal que as restrições passam a ser expressas da seguinte forma: $y_i(< \mathbf{w}, \mathbf{z}_i > +b) \geq 1 - \xi_i, \geq 0, i = 1, \ldots, n$. Quando há a ocorrência de um erro, a variável de folga assume um valor maior que 1. Assim, a quantidade $\sum_i \xi_i$ é um limitante superior do número de erros de treinamento. Isto sugere uma forma para atribuir custo (ou penalidade) aos erros: $C\left(\sum_i \xi_i\right)^k$, onde $C > 0$ é um parâmetro (definido pelo usuário) que pondera a penalidade e k é um valor inteiro e positivo. Considerando $k = 1$, o problema primal 3.50 pode ser reescrito como:

$$\begin{array}{ll} \min_{\mathbf{w},b} & \frac{1}{2} < \mathbf{w}, \mathbf{w} > +C \sum_i \xi_i \\ s.a & y_i(< \mathbf{w}, \mathbf{z}_i > +b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{array} \tag{3.59}$$

Novamente, o problema que resulta, 3.59, é um problema de otimização quadrático[6].

De forma análoga ao caso separável linearmente, a função lagrangeana é dada por:

$$L(\mathbf{w}, b, \alpha, \xi, \rho) = \frac{1}{2} < \mathbf{w}, \mathbf{w} > +C \sum_i \xi_i - \sum_{i=1}^{n} \alpha_i \left\{ y_i(< \mathbf{w}, \mathbf{z}_i > +b) - 1 + \xi_i \right\} - \sum_{i=1}^{n} \rho_i \xi_i, \tag{3.60}$$

onde $\alpha_i \geq 0, \rho_i \geq 0, i = 1, \ldots, n$, são os multiplicadores lagrangeanos.

Diferenciando $L(\mathbf{w}, b, \alpha, \xi, \rho)$ com respeito a $\mathbf{w}$, α e b e impondo a condição de estacionariedade,

[6] A outra escolha possível para k, que resulta em um problema de otimização quadrático é $k = 2$.

obtém-se as seguintes relações:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{z}_i \tag{3.61}$$

$$C - \alpha_i - \rho_i = 0, \qquad i = 1, \ldots, n \tag{3.62}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0. \tag{3.63}$$

Substituindo as relações 3.61, 3.62 e 3.63 na equação 3.60, obtém-se:

$$L(\mathbf{w}, b, \alpha, \xi, \rho) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j < \mathbf{z}_i, \mathbf{z}_j > . \tag{3.64}$$

A restrição 3.62 em conjunto com as restrições $\rho_i = 0, i = 1, \ldots, n$, implica que $\alpha_i \leq C, i = 1, \ldots, n$. Além disso, as condições de complementaridade de KKT são dadas por:

$$y_i(< \mathbf{w}, \mathbf{z}_i > + b) - 1 + \xi_i = 0, \quad i = 1, \ldots, n \qquad e \tag{3.65}$$

$$\xi_i(\alpha_i - C) = 0, \quad i = 1, \ldots, n, \tag{3.66}$$

tal que as variáveis de folga assumem valores diferentes de zero somente se $\alpha_i = C$. Assim, o problema de otimização dual associado ao treinamento de um classificador SVM com margem suave é dado por:

$$\begin{array}{ll} \max_{\alpha} & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j < \mathbf{z}_i, \mathbf{z}_j > \\ s.a & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n. \end{array} \tag{3.67}$$

A única diferença deste problema para o problema de otimização 3.55 é que os multiplicadores lagrangeanos, α_i, são limitados superiormente pelo valor da constante C. Novamente, o valor de b não aparece na formulação do problema, sendo determinado indiretamente da mesma forma que no problema separável linearmente. A representação dual do hiperplano de separação é dada por 3.58.

A função *kernel*: A implementação de uma função discriminante linear generalizada, utilizada por classificadores SVM, é equivalente a utilizar uma função não linear, $\varphi(.)$, para mapear o vetor de entrada, x, para um espaço de alta dimensão (o espaço de características), e, então, realizar a discriminação linear neste espaço de alta dimensão.

Além disso, tanto no caso separável linearmente quanto no caso não separável linearmente, o

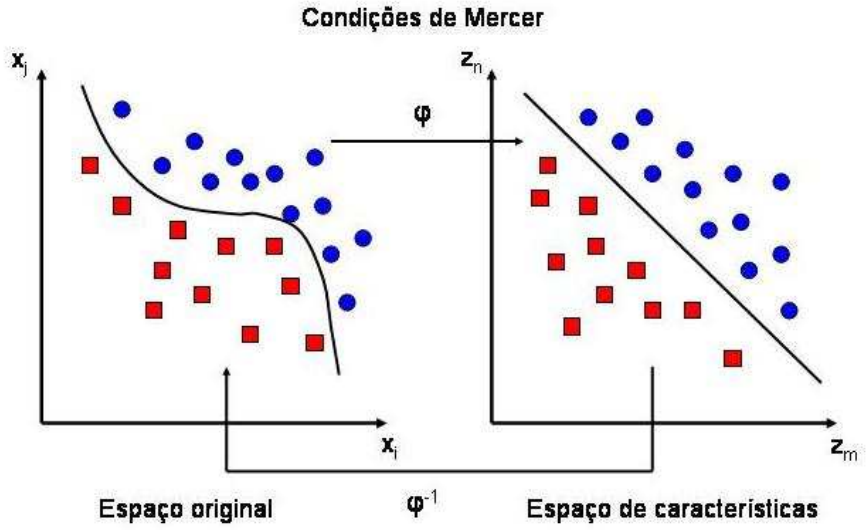

Figura 3.6: Transformação do espaço original para o espaço de características (de maior dimensão), induzida pela função *kernel*. O conjunto de padrões não é separável linearmente no espaço original, mas é separável linearmente no espaço de características.

problema de otimização que expressa o treinamento de classificadores SVM (equações 3.55 e 3.67, respectivamente), utiliza os dados apenas para computar um produto interno $(< \mathbf{z}_1, \mathbf{z}_2 > = < \varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2) >)$.

A função *kernel*, $K(.,.)$, é definida como:

$$K(\mathbf{x}_1, \mathbf{x}_2) =< \varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2) >, \tag{3.68}$$

onde $\varphi(.)$ é um mapeamento do espaço original para o espaço de características. Ela computa um produto interno, $< \varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2) >$, definido no espaço de características, através de uma função definida no espaço orignal dos dados, o que permite a fusão das duas etapas necessárias para a computação de uma função discriminante linear generalizada. Assim, pode-se dizer que um classificador SVM implementa uma função discriminante linear generalizada em um único passo (Figura **??**).

Considerando uma função *kernel*, $K(.,.)$, o problema 3.67 pode ser reescrito como:

$$\begin{array}{ll} \max_\alpha & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ s.a & \sum_{i=1}^n y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, \ldots, n. \end{array} \tag{3.69}$$

O problema 3.69 está expresso em função de valores escalares, C, y_i e α_i, e padrões, $\mathbf{x}_i$, definidos no espaço original, embora o hiperplano de separação utilizado para classificação seja definido no espaço de características. Analogamente, a superfície de decisão, correspondente ao hiperplano de

decisão no espaço de características (equação 3.58), é expressa no espaço original como:

$$g(\mathbf{x}) = \sum_{i \in SV} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{3.70}$$

A caracterização da função *kernel*, tal que a propriedade 3.68 seja satisfeita, é provida pelo ramo da matemática denominado análise funcional, através do teorema de Mercer, que tem a relação 3.68 como um caso especial. As condições do teorema, também conhecidas como condições de Mercer, podem ser enunciadas da seguinte forma (Haykin, 1999): "Seja $K(\mathbf{x}, \mathbf{x}')$ uma função contínua e simétrica, definida sobre uma região compacta $\chi \subset \Re^d$. Seja a expansão de $K(\mathbf{x}, \mathbf{x}')$ dada pela série $K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{x}')$ com coeficientes $\lambda_i \geq 0$, para todo i. Essa expansão é válida e converge absoluta e uniformemente se e somente se $\int_{\chi \times \chi} K(\mathbf{x}, \mathbf{x}') \Psi(\mathbf{x}) \Psi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$ é válido para todo $\Psi(.)$, onde $\int_{\chi} \Psi^2(\mathbf{x}) d\mathbf{x} < \infty$". As funções $\varphi_i(\mathbf{x})$ são denominadas autofunções da expansão e os números λ_i seus autovalores. A condição de positividade dos autovalores implica que $K(x, x')$ seja definida positiva. Assim, para um conjunto de dados finito, a condição de positividade é equivalente a requerer que a matriz $\mathbf{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n}$ seja definida positiva.

As condições de Mercer não indicam como construir uma função *kernel*, mas apenas como verificar se uma dada função é admissível como *kernel*. Entretanto, é possível obter novas funções que se qualifiquem como *kernels*, a partir de algumas funções *kernel*, previamente, conhecidas. Seja $K_1(.,.)$ um *kernel* (ex: o produto interno $< .,. >$), $exp(.)$ a função exponencial e $p(.)$ um polinômio com coeficientes positivos. Então, as seguintes funções são exemplos de funções *kernels* (Cristianini e Shawe-Taylor, 2000; Schlkopf e Smola, 2001):

- $K(\mathbf{x}, \mathbf{z}) = p(K_1(\mathbf{x}, \mathbf{z}))$ (polinomial);
- $K(\mathbf{x}, \mathbf{z}) = \exp(K_1(\mathbf{x}, \mathbf{z}))$ (exponencial);
- $K(\mathbf{x}, \mathbf{z}) = \tanh(\kappa K_1(\mathbf{x}, \mathbf{z}) + \upsilon), \quad \kappa > 0, \upsilon < 0$ (sigmóide); e
- $K(\mathbf{x}, \mathbf{z}) = \exp(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\sigma^2}), \quad \sigma > 0$ (gaussiano).

Além disso, se $K_1(.,.)$, $K_2(.,.)$ e $K_3(.,.)$ são funções *kernels*, α uma constante real positiva, $f(.)$ uma função real, $\varphi(.)$ um mapeamento entre dimensões finitas e B uma matriz simétrica semi-definida positiva, então, as seguintes funções também se qualificam como *kernels* (Cristianini e Shawe-Taylor, 2000):

- $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) + K_2(\mathbf{x}, \mathbf{z})$;
- $K(\mathbf{x}, \mathbf{z}) = \alpha K_1(\mathbf{x}, \mathbf{z})$;

- $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$;
- $K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$;
- $K(\mathbf{x}, \mathbf{z}) = K_3(\varphi(\mathbf{x}), \varphi(\mathbf{z}))$;
- $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^t\mathbf{B}\mathbf{z}$.

Comentário sobre a teoria de aprendizado estatístico: Para completar a introdução aos classificadores SVM, faz-se necessário um breve comentário sobre a teoria de aprendizado estatístico (Vapnik, 1997). Esta teoria, que vem sendo desenvolvida ao longo dos últimos 30 anos, tem por objetivo desenvolver técnicas de aprendizado de máquina que maximizem a capacidade de generalização de classificadores e regressores, ajustados a partir de um conjunto de dados finito.

Seja um conjunto de dados $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ independentes, amostrados de acordo com uma função densidade de probabilidades desconhecida $p(\mathbf{x}, y)$. Uma máquina de aprendizado é definida pelo conjunto de mapeamentos $MA = \{M : \mathbf{x} \longrightarrow f(\mathbf{x}, \alpha)\}$, onde α representa um conjunto de parâmetros ajustáveis de $f(.,.)$. Supõe-se que a máquina de aprendizado é determinística, tal que um elemento de MA, especificado pelo valor de α, determina uma máquina treinada. Quando a máquina de aprendizado é um classificador, essa definição é equivalente a inferir a função discriminante que resulta no menor erro de classificação de padrões não observados (padrões de teste), a partir de um conjunto de padrões de treinamento. O erro de teste esperado para uma máquina de aprendizado treinada é dado por:

$$R(\alpha) = \int \frac{1}{2}|y - f(\mathbf{x}, \alpha)|p(\mathbf{x}, y)d\mathbf{x}dy. \tag{3.71}$$

A quantidade $R(\alpha)$ é denominada risco esperado e mede o grau de generalização da máquina de aprendizado treinada. Entretanto, sua determinação depende do conhecimento de $p(\mathbf{x}, y)$ que é suposto desconhecido. Outra medida de desempenho de uma máquina de aprendizado, denominada risco empírico, é dada por:

$$R_{empir}(\alpha) = \frac{1}{2n}\sum_{i=1}^{n}|y_i - f(\mathbf{x_i}, \alpha)|. \tag{3.72}$$

O risco empírico, $R_{empir}(\alpha)$, depende de α e do conjunto de treinamento, mas não de $p(\mathbf{x}, y)$. A quantidade $|y_i - f(\mathbf{x_i}, \alpha)|$ é denominada perda e no caso de classificação assume valores 0 ou 1.

Seja η tal que $0 \leq \eta \leq 1$, então, com probabilidade $1 - \eta$, o risco esperado da função $f(.,.)$ que minimiza $R_{empir}(\alpha)$ satisfaz à seguinte relação (Vapnik, 1997):

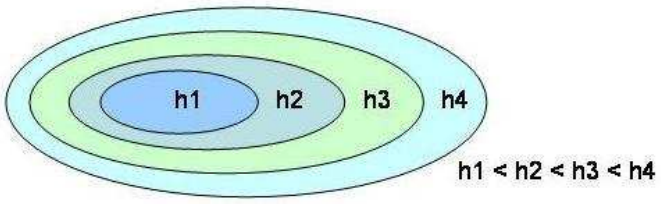

Figura 3.7: Princípio de minimização do risco estrutural. O conjunto de funções MC_i é estruturado em subconjuntos aninhados, de acordo com sua dimensão VC.

$$R(\alpha) \leq R_{empir}(\alpha) + \frac{\varepsilon(n)}{2}\left(1 + \sqrt{1 + \frac{4R_{empir}(\alpha)}{\varepsilon(n)}}\right), \quad onde \quad \varepsilon(n) = 4\frac{h\log(\frac{2n}{h}+1) - \log(\frac{\eta}{4})}{n}, \tag{3.73}$$

n representa o número de dados no conjunto de treinamento e o parâmetro h denota um número inteiro e não negativo, denominado dimensão *Vapnik-Chervonenkis* (VC). A dimensão VC representa uma medida da complexidade ou capacidade de contorção associada a uma classe de funções. Esta relação não depende de $p(\mathbf{x}, y)$ e representa um limitante para $R(\alpha)$ em função de $R_{empir}(\alpha)$ e h. Seja o conjunto de funções $MC = \{f(\mathbf{x}, \alpha) : \mathbf{x} \longrightarrow \{-1, 1\}\}$. MC é o conjunto de todas as funções que definem classificadores binários. Se um conjunto contendo l pontos de mesma dimensão que $\mathbf{x}$ pode ser rotulado de todas as 2^l possíveis maneiras, sendo que para cada uma dessas possibilidades, existe um membro de MC que atribui esses rótulos corretamente, diz-se que este conjunto de pontos é particionável (*shattered*) pelo conjunto de funções MC. Define-se a dimensão VC como a cardinalidade do maior conjunto de dados particionável por MC.

A relação 3.73 sugere a existência de uma combinação dos parâmetros $R_{empir}(a)$ e h que minimiza o risco esperado, $R(\alpha)$. Este fato, leva à proposição do princípio de minimização do risco estrutural para treinamento de uma máquina de aprendizado (SRM). O princípio SRM impõe uma estrutura ao conjunto de funções MC, dividido-o em subconjuntos aninhados de funções, tal que para todo i e j, $MC_i \subset MC_j \Leftrightarrow h_i < h_j$ (Figura 3.7). Observe que a aplicação do princípio SRM requer o conhecimento das dimensões VC para os subconjuntos de funções de MC ou de um conjunto de correspondentes limitantes. Ao utilizar o princípio SRM, uma série de máquinas de aprendizado é treinada, uma para cada subconjunto MC_i, utilizando o princípio de minimização do risco empírico, $R_{empir}(\alpha)$. Desta série, escolhe-se a maquina de aprendizado treinada que resulte no menor limitante para o risco esperado, $R(\alpha)$.

Outro resultado oriundo da teoria de aprendizado estatístico diz que se $\mathbf{x}$ está contido em uma hiperesfera de raio R, o conjunto de hiperplanos de separação com margem γ tem a dimensão VC limitada pela desigualdade (Vapnik, 1997):

$$h \leq \min\left(\frac{R^2}{\gamma^2}, d\right) + 1, \tag{3.74}$$

onde d é a dimensão do espaço onde o hiperplano reside. Portanto, o valor da margem, γ, pode ser utilizado para controlar o valor da dimensão VC.

Assim, de posse desses conceitos, a formulação do classificador SVM apresentada nas seções anteriores é justificada pelo princípio SRM que, por sua vez, minimiza o risco esperado (ou erro de generalização). No caso separável linearmente, é fácil observar que o risco esperado é minimizado para $R_{empir}(\alpha) = 0$ (separação das classes) e h mínimo (máxima margem). No caso não separável linearmente, a justificativa é um pouco mais sutil (Vapnik, 1997): dado um valor apropriado de margem, γ, e, portanto, da dimensão VC, h, é possível determinar o hiperplano de separação correspondente ao menor risco empírico, resultando no risco esperado mínimo. Na formulação de classificadores SVM, apresentada acima, existe um valor para o parâmetro C (vide equação 3.69) que leva a um valor apropriado de margem, γ, e, portanto, da dimensão VC, h, que resulta na minimização do risco esperado.

Comentário o parâmetro (C) e a função *kernel*: Apesar da proposição de classificadores SVM ser solidamente fundamentada na teoria de aprendizado estatístico e implementar o conceito de minimização do risco estrutural, o desempenho do classificador depende da escolha do parâmetro de regularização, C, e da seleção da função *kernel*. Contudo, métodos para determinação automática desses parâmetros ainda permanecem como temas de pesquisa e, na prática, estas escolhas são realizadas pelo usuário.

Na falta de um procedimento de consenso, é conveniente utilizar recomendações práticas como a de Chang e Lin (2001), que sugerem a utilização do *kernel* gaussiano e um procedimento de validação cruzada com uma estratégia de busca em *grid* para determinação dos parâmetros C e σ. Lima (2004) pondera que, para a maioria das situações encontradas na prática, esse tipo de procedimento tende a continuar competitivo em relação a métodos teóricos bem fundamentados que possam vir a ser propostos.

Estimativa de probabilidade *a posteriori* para classificadores SVM: A probabilidade *a posteriori* é muito importante em situações em que o classificador é parte de um processo mais amplo de decisão como, por exemplo, na construção de combinações de classificadores (Webb, 2002) ou na análise ROC (Fawcett, 2006). Contudo, os classificadores baseados em SVM produzem um valor (vide equação 3.70) que não representa a probabilidade *a posteriori* (Platt, 2000). Por isso, diferentes métodos têm sido desenvolvidos para calibrar a saída produzida por classificadores SVM para representar a probabilidade *a posteriori*. Em particular, Platt (2000) propôs um procedimento de pós-

processamento, onde o valor de saída produzido pelo SVM é calibrado como probabilidade, através de uma função do tipo sigmóide:

$$P_{A,B}(g) \equiv \frac{1}{1 + \exp(Ag + B)}, \tag{3.75}$$

onde g é a saída produzida pelo SVM. Os parâmetros A e B são estimados do conjunto de dados de treinamento, minimizando a correspondente função de máxima log-verossimilhança:

$$L = -\sum_i t_i log(p_i) + (1 - t_i) log(1 - t_i), \tag{3.76}$$

onde t_i é a probabilidade desejada, definida como $t_i = \frac{y_i + 1}{2}$ e $p_i = P(y_i = 1|g_i)$.

3.3 Seleção e extração de características

Existem diversas razões para se manter a dimensão do espaço de características tão baixa quanto possível. Dentre elas, estão a redução do custo associado à medição do conjunto original das características e o combate à maldição da dimensionalidade, uma vez que uma dimensão elevada do espaço de características reduz o desempenho do classificador (Jain et al., 2000).

Usualmente, os métodos de redução de dimensionalidade são divididos em duas categorias (Jain et al., 2000; Webb, 2002):

- seleção de características (ou seleção de variáveis); e
- extração de características.

Os métodos de extração de características tem como objetivo construir um novo conjunto de características combinando as características originais e, por essa razão, também são referenciados como métodos de construção de características (Guyon e Elisseff, 2003).

Já os métodos para seleção de características tem como objetivo selecionar um subconjunto das características ou variáveis originais. Kohavi e John (1997), Blum e Langley (1997) e Guyon e Elisseff (2003) descrevem 3 abordagens para seleção de características:

Filtro: a seleção das características é realizada como um procedimento de pré-processamento e independe do classificador utilizado, podendo ser realizado por um procedimento que ordena as características para, então, selecionar um subconjunto.

***Wrapper*:** um subconjunto de características é selecionado utilizando um procedimento de busca que considera o desempenho de um classificador, definido previamente, como critério de avaliação

dos subconjuntos testados. Esta abordagem resulta num conjunto de características que depende do método de classificação utilizado.

Embutidos: neste caso, o processo de seleção das características está embutido no processo de treinamento do classificador (ex: árvores de decisão).

Nas duas próximas seções são apresentados dois dos métodos mais conhecidos para extração de características no contexto de classificação de padrões (Análise de componentes principais (PCA) e Análise discriminante múltipla (MDA)). Então, na seção seguinte, são apresentados alguns dos algoritmos mais simples para seleção de características no contexto de classificação de padrões. Métodos do tipo embutidos não são abordados.

3.3.1 Análise de componentes principais

A análise de componentes principais (PCA) é um dos métodos mais utilizados para se obter uma representação mais compacta de um conjunto de dados. Componentes principais são obtidos através de combinações lineares das características originais, tal que a maior parte da variância contida nos dados esteja concentrada nos primeiros componentes. Mais especificamente, o primeiro componente principal é definido como a combinação linear normalizada das caracterísitcas originais que apresenta variância máxima, onde 'normalizada' significa que a soma dos quadrados dos coeficientes da combinação linear é igual a 1; o segundo componente principal é a combinação linear normalizada das características originais que apresenta a segunda maior variância e não é correlacionada com o primeiro componente principal. De maneira geral, se $2 < k \leq d$, onde d é o número de características originais o k-ésimo componente principal é a combinação linear normalizada das características originais que apresenta a k-ésima maior variância e não é correlacionada com quaisquer dos $k-1$ primeiros componentes principais.

Como o número de componentes principais é igual ao número de características originais, para reduzir a dimensionalidade de um conjunto de dados, através da análise de componentes principais, apenas os primeiros componentes principais são utilizados para representar os dados. A idéia é que o conjunto de dados seja descrito por um número pequeno de características (os componentes principais) que preserve, tanto quanto possível, a variância contida nos dados (Jolliffe, 2002). Essa estratégia é conhecida como sumarização parcimoniosa (Mardia et al., 1979).

Seja $\{\mathbf{x}\}$ um conjunto de vetores aleatórios de dimensão d e $\mathbf{S}$ sua matriz de covariâncias, assumida como semi-definida positiva ($\mathbf{\Sigma} \geq \mathbf{0}$). Seja também um vetor de coeficientes, α_1, assumido desconhecido, tal que o primeiro componente principal é dado por $z_1 = \alpha_1^t \mathbf{x}$. Tem-se que $\alpha_1^t \alpha_1 = 1$ e a variância de z_1 é dada por $var(z_1) = var(\alpha_1^t \mathbf{x}) = \alpha_1^t var(\mathbf{x}) \alpha_1 = \alpha_1^t \mathbf{\Sigma} \alpha_1$. Então, o primeiro componente principal é determinado resolvendo-se o seguinte problema de otimização:

$$\begin{array}{ll} \max_{\alpha_1} & \alpha_1^t \mathbf{\Sigma} \alpha_1 \\ s.a & \alpha_1^t \alpha_1 = 1. \end{array} \tag{3.77}$$

A função lagrangeana associada ao problema 3.77 é dada por $L(\alpha_1, \lambda) = \alpha_1^t \mathbf{\Sigma} \alpha_1 - \lambda(\alpha_1^t \alpha_1 - 1)$, onde λ é um multiplicador lagrangeano. Diferenciando $L(\alpha_1, \lambda)$ com respeito a α_1 e impondo a condição de estacionariedade, obtém-se que $\mathbf{\Sigma}\alpha_1 = \lambda\alpha_1$ e, portanto, λ é um autovalor de $\mathbf{\Sigma}$ e α_1 é seu correspondente autovetor. Uma vez que a matriz de covariâncias, $\mathbf{\Sigma}$, é simétrica, todos os autovalores de $\mathbf{\Sigma}$, λ, são reais. Além disso, a variância associada ao primeiro componente principal (função objetivo do problema 3.77) é dada por $\alpha_1^t \mathbf{\Sigma} \alpha_1 = \alpha_1^t \lambda \alpha_1 = \lambda$. Se $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ são os d autovalores distintos de $\mathbf{\Sigma}$, então, a solução do problema 3.77 é dada pelo autovetor α_1 e o correspondente valor da função objetivo é dado pelo autovalor λ_1.

O segundo componente principal, z_2, é determinado de forma análoga a z_1 com a restrição adicional de que z_1 e z_2 são não correlacionados. Portanto, z_2 é caracterizado pelo seguinte problema de otimização:

$$\begin{array}{ll} \max_{\alpha_2} & \alpha_2^t \mathbf{\Sigma} \alpha_2 \\ s.a & \alpha_2^t \alpha_2 = 1 \\ & \alpha_1^t \alpha_2 = 0, \end{array} \tag{3.78}$$

onde α_1 designa o vetor solução do problema 3.77 e a segunda restrição refere-se à imposição de não correlação entre o primeiro e o segundo componentes principais, $cov(z_1, z_2) = \alpha_1^t \mathbf{\Sigma} \alpha_2 = \lambda \alpha_1^t \alpha_2 = 0$.

A função lagrangeana associada ao problema 3.78 é dada por $L(\alpha_2, \lambda, \mu) = \alpha_2^t \mathbf{\Sigma} \alpha_2 - \lambda(\alpha_2^t \alpha_2 - 1) - \mu \alpha_2^t \alpha_1$, onde λ e μ são os multiplicadores lagrangeanos. Diferenciando $L(\alpha_2, \lambda, \mu)$ com respeito a α_2 e impondo a condição de estacionariedade, obtém-se que $\mathbf{\Sigma}\alpha_2 - \lambda\alpha_2 - \mu\alpha_1 = 0$, que pré-multiplicado por α_1^t resulta em $\alpha_1^t \mathbf{\Sigma} \alpha_2 - \lambda \alpha_1^t \alpha_2 - \mu \alpha_1^t \alpha_1 = 0$. Como as restrições do problema impõem que os dois primeiros termos desta expressão sejam iguais a zero e que, a partir do problema 3.77, $\alpha_1^t \alpha_1 = 1$, resulta que $\mu = 0$. Portanto, $\mathbf{\Sigma}\alpha_2 - \lambda\alpha_2 = 0$ ou $\mathbf{\Sigma}\alpha_2 = \lambda\alpha_2$ e, de maneira análoga ao problema 3.77, o valor de α_2 que maximiza $var(z_2)$ é dado pelo autovetor correspondente ao segundo maior autovalor de $\mathbf{\Sigma}$, λ_2. Cabe observar que o autovetor associado ao autovalor λ_1 não é uma solução para o problema 3.78, pois isto implicaria em $\alpha_2 = \alpha_1$, o que viola a restrição $cov(z_1, z_2) = 0$.

De forma análoga, para todo k, tal que $k \leq d$, o k-ésimo componente principal é dado pelo k-ésimo autovetor e o correspondente valor da função objetivo é dado pelo k-ésimo maior autovalor de $\mathbf{\Sigma}$. Uma demonstração completa pode ser encontrada em (Anderson, 2003).

Sejam as matrizes $\mathbf{B} = (\alpha_1, \ldots, \alpha_d)$, formada pela justaposição dos vetores (coluna) α_k, $k = 1, \ldots, d$ e $\mathbf{\Lambda} = \{\lambda_{ij} | \lambda_{ii} = \lambda_i \quad e \quad \lambda_{ij} = 0 \quad se \quad i \neq j\}$, $i, j = 1, \ldots, d$. Então, a partir das

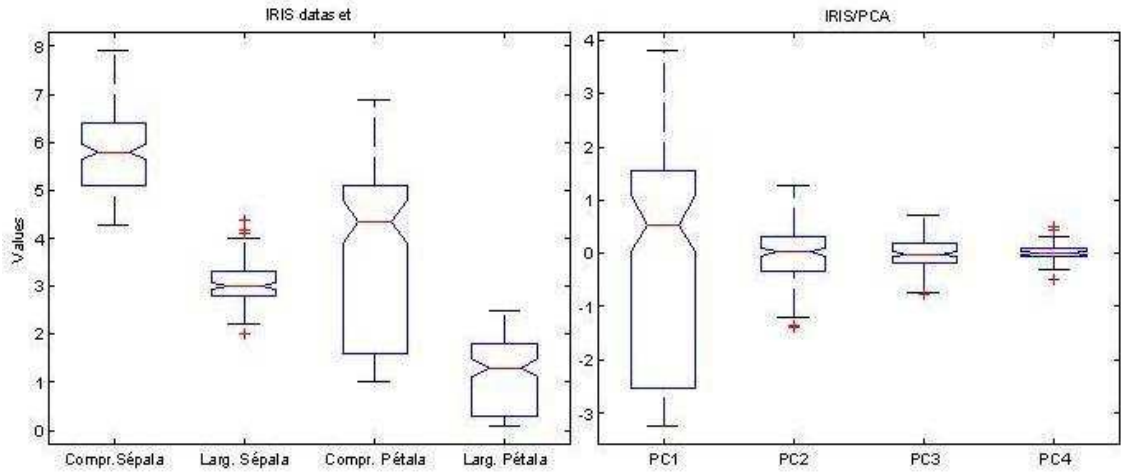

Figura 3.8: Boxplot das variáveis do conjunto de dados IRIS (à esquerda) e seus componentes principais (à direita).

equações $\mathbf{\Sigma}\alpha_k = \lambda_k\alpha_k,\ \alpha_r^t\alpha_r = 1$ e $\alpha_r^t\alpha_s = 0$, $r \neq s$, obtém-se que:

$$\mathbf{B}^t\mathbf{\Sigma}\mathbf{B} = \mathbf{B}^t\mathbf{\Lambda}\mathbf{B} = \mathbf{\Lambda}. \tag{3.79}$$

Pode-se mostrar que a transformação 3.79 preserva tanto a variância generalizada [7] quanto a soma das variâncias[8] contida nos dados, ou seja, $|\mathbf{\Sigma}| = |\mathbf{\Lambda}|$ e $tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda})$.

A Figura 3.8 apresenta o *boxplot* para as variáveis originais do conjunto de dados IRIS e para os correspondentes componentes principais. Pode-se observar que a dispersão nos componentes principais se apresenta inversamente ordenado, estando a maior parte concentrada nos primeiros componentes princiapis.

Finalmente, cabe ressaltar que, também, é possível utilizar PCA considerando a matriz de correlação. Contudo, os resultados obtidos não são equivalentes aos obtidos através da utilização da matriz de covariâncias, uma vez que PCA é uma técnica sensível à escala.

3.3.2 Análise discriminante múltipla

O objetivo da análise discriminante múltipla (MDA) é encontrar projeções cujos componentes maximizem o critério de discriminação de Fisher e sejam ortogonais entre si. Quando $c = 2$, tem-se o discriminante linear de Fisher, apresentado na seção 3.2.1.

Seja a matriz de dispersão intraclasse, $\mathbf{S}_W = \sum_{i=1}^{c}(n_i - 1)\mathbf{S}_i$, onde, como antes a dispersão e a média por classes são dados por $\mathbf{S}_i = \sum_{\mathbf{x}\in\Pi_i}(\mathbf{x} - \hat{\mu}_i)^t(\mathbf{x} - \hat{\mu}_i)$, $i = 1, \ldots, c$ e $\hat{\mu}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\Pi_i}\mathbf{x}$,

[7] A variância generalizada é definida como o determinante da matriz de covariâncias (Mardia et al., 1979).
[8] A soma de variâncias ou variância total é definida como o trace da matriz de covariâncias (Mardia et al., 1979).

respectivamente, onde n_i é o número de padrões no conjunto de dados pertencentes à classe Π_i. Seja também a dispersão total, $\mathbf{S}_T$, definida como $\mathbf{S}_T = \sum_{\mathbf{x}}(\mathbf{x} - \hat{\mu})(\mathbf{x} - \hat{\mu})^t$, onde $\hat{\mu} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i\hat{\mu}_i$ é a média considerando todo o conjunto de dados. A partir desta equação, tem-se que $\mathbf{S}_T = \mathbf{S}_W + \sum_{i=1}^{c} n_i(\hat{\mu_i} - \hat{\mu})(\hat{\mu_i} - \hat{\mu})^t$. Então, definindo a matriz de dispersão entre classes como $\mathbf{S}_B = \sum_{i=1}^{c} n_i(\hat{\mu_i} - \hat{\mu})(\hat{\mu_i} - \hat{\mu})^t$, obtém-se que $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$.

A função que se deseja maximizar, $J(\mathbf{w})$, é dada por:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} \tag{3.80}$$

Considerando que $\mathbf{S}_W$ seja definida positiva, utilizando a decomposição de Cholesky (Mardia et al., 1979), obtém-se que $\mathbf{S}_W = \mathbf{Z}^t\mathbf{Z}$. Então, considerando $\mathbf{b} = \mathbf{Z}\mathbf{w}$, obtém-se que $J(\mathbf{w}) = \frac{\mathbf{b}^t(\mathbf{Z}^t)^{-1}\mathbf{S}_B\mathbf{Z}^{-1}\mathbf{b}}{\mathbf{b}^t\mathbf{b}} = \mathbf{a}^t\mathbf{A}\mathbf{a}$, onde $\mathbf{A} = (\mathbf{Z}^t)^{-1}\mathbf{S}_B\mathbf{Z}^{-1}$ e $\mathbf{a} = \frac{\mathbf{b}}{\|\mathbf{b}\|}$, tal que $\|\mathbf{a}\| = 1$. Pode-se observar que esta restrição é satisfeita por diversos valores de $\mathbf{b}$ (e $\mathbf{w}$).

Portanto, o problema de se encontrar o primeiro componente da projeção pode ser formulado como:

$$\begin{array}{ll} \max_{\mathbf{w}} & J(\mathbf{w}) = \mathbf{a^t}\mathbf{A}\mathbf{a} \\ s.a & \|\mathbf{a}\| = 1 \end{array} \tag{3.81}$$

Este problema é formalmente equivalente ao problema de se encontrar o primeiro componente principal com a matriz de covariâncias, $\mathbf{\Sigma}$, substituída pela matriz $\mathbf{A}$. Portanto, a solução é dada pelo autovetor correspondente ao maior autovalor da matriz $\mathbf{A}$. Analogamente, as demais projeções são aquelas que maximizam $J(\mathbf{w})$ e são ortonormais às soluções anteriores. Elas são determinadas pelos autovetores de $\mathbf{A}$, com os correspondentes autovalores ordenados em ordem decrescente (Seber, 1984). Em geral, $\mathbf{A}$ possui k autovalores diferentes de zero com $k = min(d, c-1)$, o que resulta em uma projeção linear dos dados, a partir de um espaço de dimensão d, em um sub-espaço de dimensão k (Mardia et al., 1979).

Seja o r-ésimo autovalor de $\mathbf{A}$, λ_r, e o correspondente autovetor, $\mathbf{a}_r$, $r \leq k$, tal que $\mathbf{A}\mathbf{a}_r = \lambda_r\mathbf{a}_r$. Então, pré-multiplicando esta equação por $\mathbf{Z}^{-1}$ e utilizando as definições de $\mathbf{A}$, $\mathbf{Z}$, $\mathbf{a}$ e $\mathbf{b}$, obtém-se que $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w}_r = \lambda_r\mathbf{w}_r$, ou seja, a matriz $\mathbf{S}_W^{-1}\mathbf{S}_B$ possui os mesmos autovalores que a matriz $\mathbf{A}$ com autovetores, $\mathbf{w}$ e, portanto, também representa uma solução para o problema 3.81. A matriz $\mathbf{W}$ de dimensão $k \times d$, formada pela justaposição dos vetores (coluna) $\mathbf{w}_i$, $i = 1, \ldots, k$, tal que $\mathbf{y}_j = \mathbf{W}^t\mathbf{x}_j$, $j = 1, \ldots, n$, define uma transformação para um sistema de coordenadas conhecido como coordenadas discriminantes (Seber, 1984) ou variáveis canônicas.

Além disso, se as matrizes de covariância de todas as classes são iguais, $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_c$, tal que a variância total, $\mathbf{\Sigma}_t$, é dada por $\mathbf{\Sigma}_t = \mathbf{S}_w/(n-c)$, então, cada elemento (i, j) da matriz de covariâncias das variáveis canônicas, $\mathbf{\Sigma}_y$, é dada por $\mathbf{\Sigma}_y(i,j) = \mathbf{w}_i^t\mathbf{\Sigma}_t\mathbf{w}_j = \frac{\mathbf{w}_i^t\mathbf{S}_W\mathbf{w}_j}{n-c} = \frac{(\mathbf{Z}\mathbf{w}_i)^t(\mathbf{Z}\mathbf{w}_j)}{n-c} =$

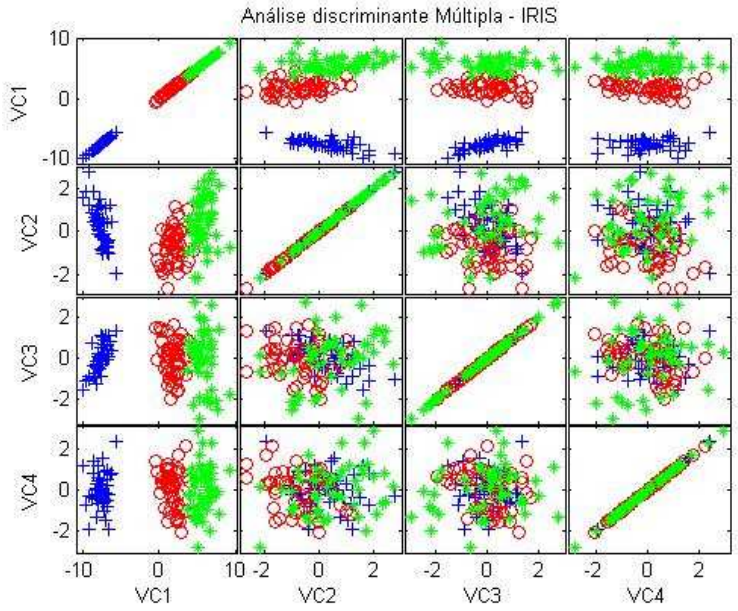

Figura 3.9: Análise discriminante múltipla para o conjunto de dados IRIS.

$\frac{\mathbf{b}^t\mathbf{b}}{n-c}$. As variáveis canônicas são não correlacionadas e para que elas tenham variância unitária, $\mathbf{\Sigma}_y = \mathbf{I}$, é preciso definir $\mathbf{b}$, tal que $\|\mathbf{b}\| = \sqrt{n-c}$ e os autovetores, $\mathbf{w}$, sejam dados por $\mathbf{w} = \sqrt{n-c}\mathbf{Z}^{-1}\mathbf{a}$.

A Figura 3.9 apresenta os gráficos para todos os pares de variáveis canônicas correspondentes ao conjunto de dados IRIS. Pode-se observar que, praticamente, toda a informação discriminante presente nos dados está concentrada na primeira variável canônica.

3.3.3 Seleção de características

Seja $\mathcal{F}$ o conjunto de características associado a um problema de classificação de padrões, tal que $|\mathcal{F}| = n$, e $\mathcal{S}$ é um subconjunto de $\mathcal{F}$. Seja também $f(\mathcal{S})$ um critério para avaliação da discriminação que resulta da utilização de $\mathcal{S}$. O problema de seleção de características pode ser formulado de duas formas: na primeira, o número de características do conjunto $\mathcal{S}$, m, é especificado *a priori* enquanto que na segunda deseja-se determinar o menor conjunto de características, tal que o valor máximo de $f(\mathcal{S})$ não exceda um determinado limiar t. Elas são equivalentes, sendo formalmente expressas pelos seguintes problemas de otimização.

Formulação com $\|\mathcal{S}\| = m$ pré-determinado	Formulação que minimiza $\|\mathcal{S}\|$
$\min_{\mathcal{S}} \quad J(\mathcal{S}) = f(\mathcal{S})$	$\min_{\mathcal{S}} \quad J(\mathcal{S}) = \|\mathcal{S}\|$
s.a $\quad \mathcal{S} \in \mathcal{F}, \|\mathcal{S}\| = m, m < n$	s.a $\quad \mathcal{S} \subseteq \mathcal{F}, f(\mathcal{S}) < t$

Diferentes critérios podem ser utilizados na definição do critério $J(.)$, incluindo (Webb, 2002):

- o erro de classificação associado a um classificador ajustado para cada um dos subconjuntos de características testados (*wrappers*); e

- uma estimativa da sobreposição entre as distribuições a partir dos quais os dados foram obtidos, tal que o valor de $J(.)$ é maior quanto maior for a discriminabilidade (filtros). Em geral, são utilizadas distâncias entre distribuições como as de Chernoff ou de Bhattacharyya (Duda et al., 2001).

Uma vez definido o critério $J(.)$, é preciso resolver o problema de otimização combinatória que resulta. Contudo, este é um problema de complexidade computacional *NP-hard* (Garey e Johnson, 1979), o que torna inviável a utilização de uma busca exaustiva à medida que o número de características, n, aumenta. Embora, algoritmos de *Branch & Bound* (Fukunaga, 1990) possam ser utilizados para encontrar a solução ótima quando n não é muito grande, a maior parte dos algoritmos utilizados para seleção de características são sub-ótimos. Eles têm como objetivo encontrar subconjuntos de $\mathcal{F}$, tais que $J(\mathcal{F})$ aproximem o valor ótimo e, em sua maioria, implementam uma busca seqüencial utilizando uma heurística gulosa.

Algoritmos seqüenciais *Forward* (SFS), *Backward* (SBS) e suas formas generalizadas (GSFS) e (GSBS)

No algoritmo SFS (Fukunaga, 1990; Webb, 2002), a busca é iniciada com o conjunto de características selecionadas vazio, tal que uma nova característica é acrescentada a cada interação. Se no passo k, o conjunto de características selecionadas é $\mathcal{S}_k$, a próxima característica a ser acrescentada, x_i pertencente a $\mathcal{F} - \mathcal{S}_k$, é escolhida tal que $J(\mathcal{S}_k \bigcup \{x_i\}) > J(\mathcal{S}_k)$ e $J(\mathcal{S}_k \bigcup \{x_i\}) > J(\mathcal{S}_k \bigcup \{x_j\})$, para todo x_j pertencente a $\mathcal{F} - \mathcal{S}_k$ e diferente de x_i, ou seja, a característica acrescentada é escolhida de tal modo que o subconjunto de características resultante maximize o incremento de $J(.)$. O algoritmo termina quando nenhuma característica pode ser encontrada em $\mathcal{F} - \mathcal{S}_k$, tal que o suconjunto de características resultante apresente um valor de $J(.)$ maior que $J(\mathcal{S}_k)$ ou quando o número máximo de características é atingido.

De maneira similar, o algoritmo SBS (Fukunaga, 1990; Webb, 2002) considera inicialmente o conjunto de características original, eliminando a cada iteração a característica que resulta no maior valor de $J(.)$.

Ambos os métodos, SFS e SBS, podem ser generalizados se a cada iteração r características, $r > 1$, são analisadas para inclusão (ou remoção) do subconjunto de características selecionadas. Essas generalizações são denominadas de SFS generalizado (GSFS) e SBS generalizado (GSBS), respectivamente (Webb, 2002). Embora envolvendo um custo computacional mais elevado, GSFS e

GSBS apresentam a vantagem de levar em consideraração a possibilidade da existência de características relacionadas.

Algoritmo seqüencial de seleção acrescente l - remova r e suas formas generalizadas

O algoritmo acrescente l - remova r, denominado PTA(l,r) (Webb, 2002), acrescenta algum *backtracking* ao procedimento de busca sequencial, podendo ser executado tanto como um procedimento *top-down* quanto como *bottom-up*. Se $l < r$, o procedimento é *bottom-up*. A cada iteração, l características são acrescentadas ao subconjunto de características selecionado, utilizando o algoritmo SFS, seguido pela remoção de r características, utilizando o algoritmo SBS, até que o número de características requerido seja atingido. De forma análoga, se $l > r$ o procedimento é *top-down*. Começando com o conjunto de características original, a cada iteração r características são removidas e l acrescentadas até que o número requerido seja atingido.

Uma forma simples de generalização do algoritmo PTA(l,r) é a utilização dos algoritmos GSFS e GSBS considerando os valores de l e r, respectivamente. Uma forma alternativa de generalização menos custosa em termos computacionais consiste em subdividir l e r em diversos componentes, $l_i = 1, \ldots, n_l$ e $r_j = 1, \ldots, n_r$, tal que $0 \leq l_i \leq l$, $0 \leq r_j \leq r$, $\Sigma l_i = l$ e $\Sigma r_j = r$. Neste caso, ao invés de aplicar o algoritmo GSFS em um único passo, denotado por GSFS(l), ele é aplicado n_l vezes de maneira sucessiva para $i = 1, \ldots, n_l$, ou seja, aplica-se GSFS(l_i) para $i = 1, \ldots, n_l$. De maneira análoga, GSBS(r) é substituído pela aplicação sucessiva de GSBS(r_j) para $j = 1, \ldots, n_r$ (Webb, 2002).

Algoritmos de busca seqüencial com seleção flutuante (SFFS) e (SFBS)

Os algoritmos de busca seqüencial com seleção flutuante, *forward* (SFFS) e *backward* (SFBS), podem ser considerados como uma alternativa de generalização do algoritmo PTA(l,r), onde os valores de l e r são flutuantes, no sentido de variarem ao longo da execução do algoritmo (Pudil et al., 1994).

O algoritmo SFFS retém em memória os melhores subconjuntos de características, $\mathcal{S}_1, \ldots, \mathcal{S}_m$, de cardinalidades $1, \ldots, m$, respectivamente, onde m é a máxima cardinalidade considerada pelo algoritmo. Seja o correspondente valor de $J(.)$ para cada subconjunto, $\mathcal{S}_i$, $J_i = J(\mathcal{S}_i), \quad i = 1, \ldots, m$.

Algoritmo SFFS:

1. Inicializar: $k = 0$; $\mathcal{S}_0 \leftarrow \emptyset$;

2. Execute o algoritmo SFS até que um conjunto $\mathcal{S}_2$ seja obtido; $k \leftarrow 2$.

3. Execute os passos abaixo até que nenhuma melhora seja obtida:

   (a) Selecione a característica x_j de $\mathcal{F} - \mathcal{S}_k$ executando uma iteração do algoritmo SFS e atualize o subconjunto de características selecionadas: $\mathcal{S}_{k+1} \leftarrow \mathcal{S}_k \bigcup \{x_j\}$.

   (b) Encontre a característica x_r no subconjunto de características selecionadas, $\mathcal{S}_{k+1}$, executando uma iteração do algoritmo SBS. Se $x_r = x_j$, faça $J_{k+1} \leftarrow J(\mathcal{S}_{k+1})$, incremente k e vá para o passo 3a. Caso contrário faça $\mathcal{S'}_k \leftarrow \mathcal{S}_{k+1} - \{x_r\}$.

   (c) Continue a remover características do subconjunto $\mathcal{S'}_k$, obtendo $\mathcal{S'}_{k-1}$ enquanto $J(\mathcal{S'}_{k-1}) > J_{k-1}$; $k \leftarrow k - 1$; ou até que $k = 2$. Então, vá para o passo 3a.

De forma similar, o algoritmo SFBS é inicializado com $k = n$, $\mathcal{S} \leftarrow \mathcal{F}$ e executa o algoritmo SBS até que $\mathcal{S}_{n-2}$ seja obtido. As iterações são realizadas de forma a primeiro remover e, posteriormente, acrescentar características do subconjunto selecionado.

3.4 Avaliação de desempenho

Uma vez que um classificador tenha sido construído, é preciso avaliar se seu desempenho satisfaz às necessidades da aplicação para a qual ele se destina ou se modelos e/ou caracteríticas alternativas precisam ser avaliadas. O parâmetro mais importante ao se avaliar um classificador é sua taxa de erro de classificação (Jain et al., 2000). O erro de um classificador pode ser divido em três partes: o erro Bayesiano, que é inerente ao problema e não pode ser reduzido; o erro que resulta da escolha inadequada do modelo de classificação; e o erro que resulta da insuficiência dos dados para estimar os parâmetros do classificador. O ideal é que os dois últimos sejam solucionados durante o desenvolvimento do classificador, tal que o erro de classificação aproxime o erro Bayesiano.

Para algumas famílias de f.d.p's é possível obter um limitante adequado para o erro Bayesiano (Jain et al., 2000) como, por exemplo, os limitantes de Chernoff e Battacharyya para a f.d.p normal (Duda et al., 2001). Entretanto, para a maior parte das aplicações, a taxa de erro de classificação precisa ser estimada a partir do conjunto de dados disponível, usualmente, dividindo-o em dois conjuntos de dados independentes. Um deles, o conjunto de dados de treinamento, é utilizado para estimar os parâmetros do classificador, enquanto o outro, o conjunto de dados de teste, é utilizado para estimar o erro de classificação, através da porporção de padrões de teste classificados incorretamente. A confiabilidade desta estimativa dependende dos tamanhos dos conjuntos de dados de treinamento e teste, bem como da independência entre os mesmos (Jain et al., 2000). Visando obter melhores estimativas da taxa de erro, dadas as diferentes condições determinadas pela quantidade de dados disponível, diferentes estratégias para divisão do conjunto de dados de treinamento e teste têm sido propostos. Aquelas mais comumente utilizadas incluem (Jain et al., 2000; Duda et al., 2001; Webb, 2002):

- Ressubstituição: o mesmo conjunto de dados é utilizado tanto para treinamento quanto para teste, tal que a taxa de erro é estimada pela de taxa erro de treinamento. A taxa de erro obtida também é denominada taxa de erro aparente. Sendo os conjuntos de dados de treinamento e teste não independentes, este método é inapropriado para medir o desempenho de um classificador, e resulta em uma estimativa tendenciosamente otimista. Entretanto, este método pode ser utilizado quando o número de padrões no conjunto de dados, n, é muito grande. De fato, à medida que n tende para infinito, todas as estratégias aqui apresentadas levam à mesma estimativa de erro (Jain et al., 2000).

- *Holdout*: neste caso, o conjunto de dados é dividido em duas partes mutuamente exclusivas, uma utilizada para estimar os parâmetros do classificador e a outra para estimar a taxa de erro. O conjunto de dados é sub-utilizado e a estimativa que resulta é tendenciosamente pessimista.

- Validação cruzada: este método, às vezes denominado rotação ou validação cruzada k-*fold*, consiste em dividir o conjunto de dados em $k \leq n$ partes e, então, para cada bloco de dados $i, i = 1, \ldots, k$, treinar o classificador com um subconjunto de dados formado pela exclusão do bloco i e utilizar os dados deste bloco para teste. A taxa de erro é estimada considerando as predições realizadas sobre os padrões pertencentes ao bloco excluído do conjunto de treinamento. Diversas variantes são possíveis em função da escolha de k e da forma como o conjunto de dados é divido. Se $k = n$, o método é conhecido como *leave-one-out* e se $n = 2$, tem-se uma versão do método de *Holdout* em que as duas partições são utilizadas alternadamente para treinamento e teste. Este método fornece uma boa estimativa da taxa de erro, apesar de ser computacionalmente muito mais caro que os métodos apresentados nos dois itens anteriores, especialmente à medida que o valor de k cresce.

- *Bootstrap*: este método estima a taxa de erro esperada utilizando a taxa de erro aparente combinada com um fator de correção de *bias*, estimado por um procedimento de *bootstrap*. Usualmente, B conjuntos de dados, de tamanho n, são construídos por re-amostragem com reposição do conjunto de dados original. Cada subconjunto de dados $i, i = 1, \ldots, B$ é utilizado como um conjunto de treinamento diferente. Uma vez que os B subconjuntos de dados são construídos através de amostragem com reposição, existe uma probabilidade de que alguns dados sejam amostrados mais de uma vez enquanto que outros nem sejam amostrados. Em uma de suas variantes, denominada estimador 0.632, o erro estimado é uma combinação linear do erro aparente, e_A, e o erro obtido do estimador *bootstrap*, e_0: $e_{0.632} = 0.368e_A + 0.632e_0$. Para estimar e_0, o conjunto de dados original é utilizado como conjunto de teste, sendo e_0 dado pela taxa de erro considerando-se o conjunto de padrões que não aparece nas amostras de *bootstrap*. O número de padrões classificados erroneamente para todas as amostras *bootstrap* é somada e

dividida pelo número total de padrões que não fazem parte das amostras *bootstrap*. Embora este método seja o de maior custo computacional, ele também fornece as melhores estimativas da taxa de erro.

Em algumas aplicações também é interessante avaliar o desempenho do classificador utilizando outros tipos de medidas. Por exemplo, um classificador pode considerar a opção de rejeição, que implica em não classificar padrões próximos da fronteira de decisão, segundo um dado limiar, ou postergar a sua classificação até que mais informação esteja disponível (Webb, 2002). Nesse caso, mensurar a taxa de rejeição, ou a proporção dos dados de teste rejeitados pelo classificador, fornece uma informação útil sobre o comportamento do classificador. Esta medida indica o nível de confiabilidade com que as classificações são realizadas.

Por vezes, também pode ser interessante observar a matriz de confusão, que fornece informações sobre a maneira como a taxa de erro se decompõe. Sejam c classes, Π_i, $i = 1, \ldots, c$. A matriz de confusão é uma matriz de dimensão $c \times c$ onde cada elemento (i, j) indica a freqüência absoluta de padrões da classe i classificadas como pertencentes à classe j. Idealmente, a matriz de confusão é uma matriz diagonal, com todos os elementos fora da diagonal principal iguais a zero e, por conseqüência, com taxa de erro também igual a zero.

Quando o problema de classificação possui apenas duas classes, $c = 2$, denominando uma das classes como positiva e a outra como negativa, os elementos da matriz de confusão podem ser identificados individualmente como: positivo verdadeiro (TP), falso positivo (FP), negativo verdadeiro (TN) e falso negativo (FN). A partir desses valores, é possível definir uma série de parâmetros de desempenho do classificador (Baldi et al., 2000) como, por exemplo, a taxa de detecção, taxa de *hits*, cobertura ou sensitividade que, definida como

$$taxa\ de\ hits = \frac{TP}{TP + FN}, \tag{3.82}$$

estima a probabilidade de classificar um padrão positivo corretamente; a especificidade ou precisão, que definida como

$$especificidade = \frac{TP}{TP + FP}, \tag{3.83}$$

estima a probabilidade de que a classificação de um padrão como positivo esteja correta; a taxa de falso positivos, taxa de falsa aceitação ou falso alarme, que definida como

$$falso\ alarme = \frac{FP}{FP + TN}, \tag{3.84}$$

estima a probabilidade de que uma classificação de um padrão como positivo esteja incorreto; a

especificidade negativa ou taxa de falsa rejeição, que definida como

$$falsa\ rejeicao = \frac{FN}{FN + TP}, \tag{3.85}$$

estima a probabilidade de que a classificação de um padrão como negativo esteja incorreto; e o coeficiente de correlação entre os dados observados e os preditos, que para o caso de duas classes é definido como

$$coeficiente\ de\ correlacao = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}. \tag{3.86}$$

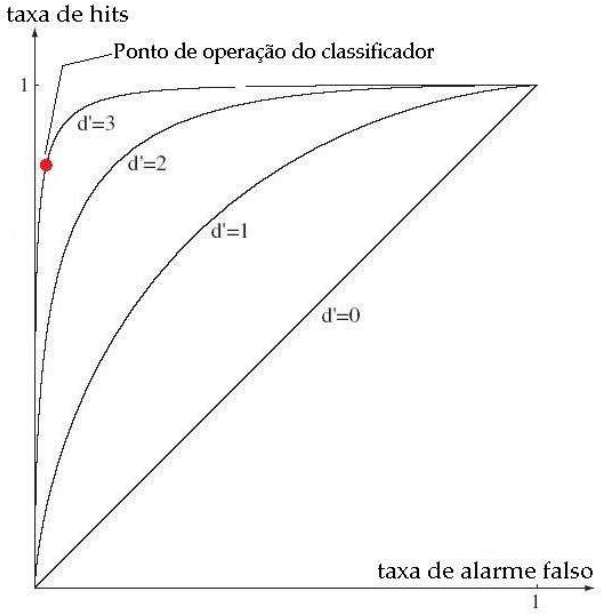

Figura 3.10: Curva ROC. Valores maiores de d' indicam maior capacidade de discriminação do classificador. O limiar de decisão indica o ponto de operação do classificador. Adaptado de (Duda et al., 2001).

Além disso, a taxa de positivo verdadeiros (taxa de *hits*) e a taxa de falsos positivos (falso alarme) podem ser utilizados para construir uma curva denominada *receiver operating characteristic* (ROC) (Fawcett, 2006) (Figura 3.10). A curva ROC representa os diferentes valores de taxa de positivos verdadeiros e falsos positivos para os quais um classificador pode ser, simultameamente, calibrado. Ela independe tanto das probabilidades *a priori* de cada classe quanto do custo de erro de classificação. Os diferentes valores de taxa de positivo verdadeiros e taxa de falso positivos são controlados por um limiar, geralmente, um *score* (ex: probabilidade *a posteriori*) associado à classe considerada como positiva. Assim, utilizando este limiar, o usuário pode calibrar o classificador para

o ponto de operação mais apropriado à sua aplicação. Além disso, o desempenho do classificador também pode ser sumarizado através de um escalar, denominado *area under ROC curve* (AUC). Ele representa a probabilidade de um classificador dar preferência a um padrão positivo, escolhido aleatoriamente, em relação a um padrão negativo, também escolhido aleatoriamente, quando ordenando-os de forma descrescente, de acordo com a probabilidade *a posteriori* do padrão ser positivo (Hanley e McNeil, 1982).

3.5 Considerações finais

Neste capítulo foram apresentadas técnicas para o desenvolvimento de classificadores, incluindo métodos para classificação, extração e seleção de características, e critérios de avaliação de desempenho de classificadores. Contudo, deve ser enfatizado que existe uma diversidade enorme de técnicas descritas na literatura de classificações de padrões (Duda et al., 2001; Webb, 2002; Bishop, 1997; Jain et al., 2000) e que não foram abordadas. Por exemplo, com relação à técnicas de classificação, as seguintes técnicas também são muito utilizadas:

Discriminante logístico Os discriminantes logísticos (caso de duas classes) pressupõem que a diferença entre os logaritmos das funções de densidades de probabilidades condicionais é uma função linear do vetor de características. Essa é uma suposição válida para muitas famílias de distribuição de probabilidades, o que faz com este classificador seja aplicável a diversas situações em que os dados não seguem a distribuição normal (Webb, 2002).

Redes neurais artificiais (RNA) As RNA modelam funções de decisão não lineares e são inspiradas no funcionamento do cérebro. Seu processamento depende tanto do modelo do neurônio quanto da topologia que os interliga, ou seja, de sua arquitetura. Em geral, as arquiteturas de RNAs podem ser classificadas em três grandes grupos (Haykin, 1999): *feed-forward* de camada simples, *feed-forward* de múltiplas camadas e redes recorrentes. Diversos modelos de RNA com arquiteturas variadas são de interesse para reconhecimento de padrões, sendo a arquitetura *feed-forward* de múltiplas camadas, ou Perceptron de múltiplas camadas (MLP), a mais utilizada em aplicações práticas.

Árvore de decisão As árvores de decisão são exemplos de procedimentos de decisão multi-estágio, em que diferentes subconjuntos de características são utilizados para tomada de decisão em diferentes níveis da árvore. Elas modelam superfícies de decisão não lineares e são facilmente interpretáveis, facilitando a tarefa de compreensão da estrutura contida nos dados (Webb, 2002).

Combinações de classificadores Classificadores também podem ser construídos através da combinação de classificadores simples (Webb, 2002) (Kittler et al., 1998), utilizando regras simples para combinação (ex: voto, min, max, mean, product e median), procedimentos para geração de diferentes classificadores a partir do mesmo conjunto de dados (ex: Bagging (Breiman, 1996), Boosting (Freund e Schapire, 1999) e Random Forest (Breiman, 2001)) ou procedimentos de aprendizado que combinam diferentes componentes classificadores utilizando uma função de *gating* (ex: mistura de especialistas (Jacobs et al., 1991)).

Também existem muitas outras técnicas para extração e seleção de características como, por exemplo, *multidimensional scaling* (MDS) (Mardia et al., 1979), mapas auto-organizáveis (SOM) (Haykin, 1999) e *kernel*-PCA (Schlkopf e Smola, 2001), e algoritmos baseados em meta-heurísticas como algoritmos genéticos (GA) (Jain e Zongker, 1997).

O escopo considerado neste capítulo compreendeu apenas aquelas técnicas utilizadas no desenvolvimento dos preditores de região de interface e de *hot spots*, apresentados nos capítulos 4 e 5, respectivamente. Ao leitor interessado em uma descrição detalhada dessas técnicas, recomenda-se a leitura das referências indicadas.

Capítulo 4

Predição de regiões de interface proteína-proteína

Neste capítulo, é apresentado um preditor de regiões de interface proteína-proteína, baseado em medidas de propriedades físico-químicas e estruturais, considerando os aminoácidos da superfície da molécula de proteína como as unidades básicas para classificação. Este preditor elimina a restrição de regiões de interface de formato circular associada à utilização de *patches*, uma evolução sugerida por Bradford e Westhead (2005). No seu desenvolvimento, o modelo de interface proposto por Lo Conte et al. (1999), Chakrabarti e Janin (2002), Bahandur et al. (2003) e De et al. (2005) é explorado através de uma estratégia de classificação em dois estágios.

Inicialmente, é apresentada uma revisão bibliográfica sobre o tema visando posicionar o preditor desenvolvido em relação à literatura. Em seguida, é apresentada a estratégia em dois estágios utilizada para construção do preditor. Então, utilizando as técnicas de classificação de padrões introduzidas no capítulo 3, é apresentada a metodologia de desenvolvimento do preditor, e os experimentos realizados para avaliar seu desempenho. As propriedades utilizadas na formação do vetor de características foram apresentadas no capítulo 2.

4.1 Trabalhos relacionados

Os estudos para caracterização estrutural da interface de interação proteína-proteína iniciaram-se há pelo menos 30 anos (Chothia e Janin, 1975), tendo se concentrado basicamente na análise da composição de aminoácidos na região de interface, comparando-a com a composição do restante da superfície da proteína e a de seu interior. De forma geral, tem sido relatado que a região de interface apresenta uma maior propensão por aminoácidos hidrofóbicos que o restante da superfície da proteína, porém menor que a de seu interior (Lo Conte et al., 1999; Chakrabarti e Janin, 2002;

Bahandur et al., 2003). A região de interface proteína-proteína também tem sido analisada através de propriedades estruturais medidas sobre sua superfície. Jones e Thornton (1997a), por exemplo, analisaram propriedades relacionadas à geometria e à facilidade de solvatação dos aminoácidos na superfície da proteína. Outra propriedade importante é o grau de conservação de aminoácidos (Valdar e Thornton, 2001), que é relatado como mais elevado para aminoácidos na região de interface, embora alguns estudos questionem este resultado (Caffrey et al., 2004).

Um dos principais resultados relativos à caracterização da região de interface é que ela não constitui uma região uniforme. Lo Conte et al. (1999), Chakrabarti e Janin (2002), Bahandur et al. (2003) e De et al. (2005) analisaram a perda de área acessível a solvente na formação do complexo. Eles propuseram um modelo onde a região de interface é formada por (a) um núcleo composto por aminoácidos que são acessíveis a solvente quando a proteína está isolada, mas que quando em complexo se tornam completamente obstruídos; e (b) um anel periférico composto por aminoácidos que não são totalmente obstruídos quando em complexo. O núcleo foi relatado como apresentando uma composição de aminoácidos mais próxima da composição do interior da proteína, sendo esta tendência mais acentuada para o caso de interações do tipo homodímeros (Bahandur et al., 2003), enquanto para o anel externo foi observada uma composição mais próxima à do restante da superfície da proteína. Este modelo para a região de interação proteína-proteína apresenta semelhanças com o modelo proposto por Bogan e Thorn (1998), denominado *O-ring*, baseado em dados experimentais de mutagênese sítio-dirigido, onde um agregado de aminoácidos no centro da interface, denominado *hot spots*, apresentam maior contribuição para a energia de interação. Ambos os modelos apresentam uma composição de aminoácidos similar, com maiores propensões para aminoácidos do tipo Trp, Arg e Tyr (vide anexo B).

O problema de predição de regiões de interface consiste em determinar quais aminoácidos pertencem à região de interface quando os dados sobre a proteína com a qual ela interage não estão disponíveis. Preditores de regiões de interface proteína-proteína podem ser desenvolvidos utilizando o conhecimento adquirido e as propriedades analisadas em estudos de caracterização estrutural da interface de interação proteína-proteína. Esse tipo de preditor provê informações úteis para a formulação de hipóteses sobre a função da proteína, identifica aminoácidos prioritários para análise experimental e pode ser utilizado para guiar o processo de busca em algoritmos de *molecular docking* (Halperin et al., 2002). *Molecular docking* refere-se a métodos para predição da orientação preferencial de uma molécula com relação a uma segunda na formação de um complexo molecular estável.

Diversos preditores de regiões de interface proteína-proteína foram propostos na literatura, baseados no conhecimento da seqüência de aminoácidos ou da estrutura tridimensional da proteína. Ofran e Rost (2003b) e Yan et al. (2004) utilizaram uma codificação para os 20 tipos de aminoácidos que compõem as proteínas e construíram o vetor de características para cada aminoácido da estrutura

primária da proteína concatenando o seu código com os dos n aminoácidos precedentes e dos n subseqüentes, onde n é igual a 4 (vizinhança seqüencial). O primeiro utilizou uma rede neural do tipo *feed-forward* como classificador e relatou uma acuidade de 70%, enquanto o segundo utilizou SVM como classificador e relatou uma acuidade de 72%, considerando apenas aminoácidos na superfície da proteína. Nguyen et al. (2007) também consideraram apenas informações sobre o tipo de aminoácido e realizaram a predição utilizando um *Hidden Markov Model* (HMM), relatando uma cobertura de 61% e precisão de 66%. Reš et al. (2005) e Ofran e Rost (2006) utilizaram um conjunto de seqüências de proteínas homólogas à seqüência de interesse, a partir do qual um alinhamento múltiplo é construído e a freqüência relativa de ocorrência de cada um dos 20 tipos de aminoácidos padrão em cada posição do alinhamento é extraída (perfil evolucionário). O vetor de características para cada aminoácido da seqüência foi formado pela concatenação do seu perfil evolucionário com o dos n aminoácidos precedentes e n subseqüentes. Reš et al. (2005) utilizaram um SVM para realizar a prdição e relataram uma acuidade de 64%, enquanto Ofran e Rost (2006) utilizaram uma rede neural do tipo *feed-forward* e relataram que para um nível de acuidade de 61%, pelo menos um aminoácido da região de interface foi identificado em mais de 90% das proteínas testadas. Fariselli et al. (2002) consideraram os aminoácidos da superfície da estrutura tridimensional da proteína e construíram o vetor de características para cada aminoácido na superfície da proteína concatenando o perfil evolucionário para o aminoácido e os 10 aminoácidos de superfície mais próximos (vizinhança estrutural). Eles utilizaram uma rede neural do tipo *feed-forward* como classificador, relatando uma acuidade de 73%. Wang et al. (2006) utilizaram o mesmo conjunto de dados e um vetor de características formado pelo perfil evolucionário com vizinhança estrutural acrescido do grau de conservação de aminoácidos. Eles avaliaram duas medidas diferentes de grau de conservação de aminoácidos e utilizaram SVM como classificador. Na melhor configuração obtida pelos autores, foi relatada uma acuidade de 63.7%, correspondendo a uma cobertura de 53.75% e uma precisão de 47.5%. Zhou e Shan (2001) também utilizaram um vetor de características formado pelo perfil evolucionário com vizinhança estrutural, acrescido de SAS, e relataram uma acuidade de 70%. Da mesma forma, Koike e Takagi (2004) também utilizaram o perfil evolucionário para construir o vetor de características, mas utilizaram SVM como classificador. Eles testaram o desempenho do classificador ao acrescentar a SAS, o índice de planaridade e a proporção de aminoácidos de interface na proteína como características, bem como a definição de vizinhança seqüencial e estrutural. O melhor desempenho foi obtido ao se utilizar vizinhança estrutural e as propriedades adicionais SAS e proporção de aminoácidos na proteína, correspondendo a uma acuidade de 73.5%. Bordner e Abagyan (2005) utilizaram SVM como classificador e o perfil evolucionário com vizinhança estrutural para construir o vetor de características. Eles também avaliaram o desempenho do classificador ao se incluir as propriedades hidrofobicidade, proporção de tipos de aminoácidos e grau de conservação de aminoácidos. Uma me-

lhora no desempenho do classificador foi obtida apenas quando as duas últimas propriedades foram acrescentadas. Os autores relataram uma acuidade de 80%, correspondendo a uma cobertura de 57% e uma precisão de 39%. Li et al. (2007) utilizaram o perfil evolucionário com vizinhança estrutural, acrescido de SAS e grau de conservação de aminoácidos para formar o vetor de características e *Conditional Random Fields* (CRF) (Sutton e McCallum, 2006) para realizar a classificação, com a informação contextual obtida da estrutura primária da proteína. Foi relatada uma acuidade de 75%, correspondente a uma cobertura de 30.4% e uma precisão de 51.6%. Os autores também compararam estes resultados com os obtidos ao se utilizar SVM sobre o mesmo conjunto de dados, concluindo que CRF apresenta um desempenho compatível com o de SVM, mas é mais robusto a variações na razão entre exemplos positivos e negativos. Chung et al. (2006) acrescentaram ao vetor de características formado pelo perfil evolucionário com vizinhança estrutural, SAS e uma medida de conservação estrutural de aminoácidos. Utilizando SVM como classificador , eles relataram que em 52% dos casos de teste, os aminoácidos de interface foram precisamente preditos (mais de 70% dos aminoácidos da região de interface foram identificados), 77% dos aminoácidos de interface foram corretamente preditos (mais de 50% dos aminoácidos da região de interface foram identificados) e em 21% dos casos de teste, os aminoácidos de interface foram parcialmente preditos (os aminoácidos identificados representam menos de 50% do total dos aminoácidos de interface).

Já Jones e Thornton (1997b), Bradford e Westhead (2005) e Neuvirth et al. (2004) consideraram propriedades físico-químicas e estruturais medidas sobre *patches*, uma região contínua de forma aproximadamente circular, formada por um conjunto de aminoácidos amostrados sobre a superfície da proteína. Jones e Thornton (1997b) construíram o vetor de características utilizando um conjunto de seis propriedades, utilizadas em um estudo anterior (Jones e Thornton, 1997a): potencial de solvatação, propensão do aminoácido para interface, hidrofobicidade, planaridade, índice de protrusão e área acessível a solvente. Os autores relataram que o preditor foi capaz de localizar a região de interface em 66% dos casos de teste. Bradford e Westhead (2005) utilizaram um conjunto de propriedades mais extenso projetados na superfície da proteína para formar o vetor de características: grau de conservação de aminoácidos, potencial eletrostático na superfície da proteína, índice de hidrofobicidade, propensão de aminoácidos para a interface, área acessível a solvente e duas medidas relacionadas à forma geométrica da superfície, baseadas na curvatura, o índice de forma e a *curvedness*. Utilizando um classificador do tipo SVM, os autores relataram uma taxa de sucesso de 76%. Posteriormente, os mesmos autores relataram uma taxa de sucesso de 82%, utilizando os mesmos dados e um classificador do tipo näive Bayes (Bradford et al., 2006). Neuvirth et al. (2004) utilizaram uma estratégia bayesiana empírica para estimar a probabilidade de um *patch* representar uma região de interface. O conjunto de propriedades utilizadas para formar o vetor de características inclui a composição química das regiões de interface (propensão de aminoácidos para a interface, distribuição de pares de

Preditor	Informação	Método de classificação	Desempenho
Ofran e Rost (2003b)	S, aa	RNA	Ac=70%
Yan et al. (2004)	S, aa	SVM	Ac=72%
Nguyen et al. (2007)	S, aa	HMM	Cob=61% e Prec=66%
Reš et al. (2005)	S, aln	SVM	Ac=64%
Ofran e Rost (2006)	S, aa	RNA	Ac=61% com 1 verdadeiro positivo em 90% das proteínas
Fariselli et al. (2002)	E, aln	RNA	Ac=73%
Wang et al. (2006)	E, aln	SVM	Ac=63.7%, Cob=53.75% e Prec=47.5%
Zhou e Shan (2001)	E, aln	RNA	Ac=70%
Koike e Takagi (2004)	E, aln	SVM	Ac=73.5%
Bordner e Abagyan (2005)	E, aln	SVM	Ac=80%, Cob=57% Prec=39%
Li et al. (2007)	E, aln	CRF	Ac=75%, Cob=30.4% e Prec=51.6%
Chung et al. (2006)	E, aln	SVM	52% precisos, 77% corretos 21% parcialmente corretos
Jones e Thornton (1997b)	E, fqs	empírico	TxSuc=66%
Bradford e Westhead (2005)	E, fqs	SVM	TxSuc=76%
Bradford et al. (2006)	E, fqs	Näive Bayes	TxSuc=82%
Neuvirth et al. (2004)	E, fqs	Bayes empírico	TxSuc1=70%
Higa e Tozzi (2008b)	E, fqs	Linear	TxSuc=82.1%

Tabela 4.1: Preditores de regiões de interface. S: considera apenas a seqüência dos aminoácidos; E: considera a estrutura tridimensional da proteína. aa: vetor de características formado a partir da codificação do aminoácido na posição analisada e de seus vizinhos; aln: vetor de características basicamente formado a partir de um alinhamento múltiplo de proteínas homólogas, podendo incluir alguma medida adicional como conservação de aminoácidos ou área acessível a solvente; fqs: vetor de características formado por um conjunto de propriedades físico-químicas e estruturais. RNA: Redes neurais artificiais do tipo *feed-forward*; SVM: Máquinas de vetores-suporte. HMM: modelos ocultos de Markov; CRF: campos condicionais aleatórios. Ac: acuidade; Cob: cobertura; Prec: precisão; TxSuc: Taxa de sucesso em localizar a região de interface com base em valores mínimos de precisão e cobertura; TxSuc1: Taxa de sucesso em localizar a região de interface com base em valores mínimos de precisão.

aminoácidos estruturalmente vizinhos, grau de conservação de aminoácidos), propriedades geométricas (estrutura secundária, comprimento de estruturas secundárias não regulares, distância seqüencial) e informações específicas obtidas do experimento de difração por raios X (fator de temperatura e número normalizado de moléculas d'água). Uma taxa de sucesso de 70% foi relatada pelos autores.

A tabela 4.1 sumariza as principais características dos preditores de regiões de interface apresentados acima. A última linha da tabela refere-se ao resultado apresentado neste capítulo.

A questão abordada neste capítulo tem origem no fato dos trabalhos, descritos na literatura, para predição de regiões de interface, baseados em propriedades físico-químicas e estruturais, utilizarem o conceito de *patches*, tal que as regiões a serem classificadas como interface ou não interface são amostradas *a priori*, de forma arbitrária.

Neste capítulo, é proposto um preditor de regiões de interface, baseado em propriedades físico-químicas e estruturais, que utiliza o aminoácido da superfície da proteína como unidade básica para classificação. Considerando que a superfície da proteína é dividida de forma mais natural com base em seus aminoácidos, a idéia é que a região de interface predita aproxime melhor a região de interface real utilizando esta estratégia. Nesta proposta, as propriedades físico-químicas e estruturais são utilizadas para predizer os aminoácidos de interface, sendo a região de interface, propriamente dita, formada pela justaposição desses aminoácidos.

4.2 Estratégia de predição proposta

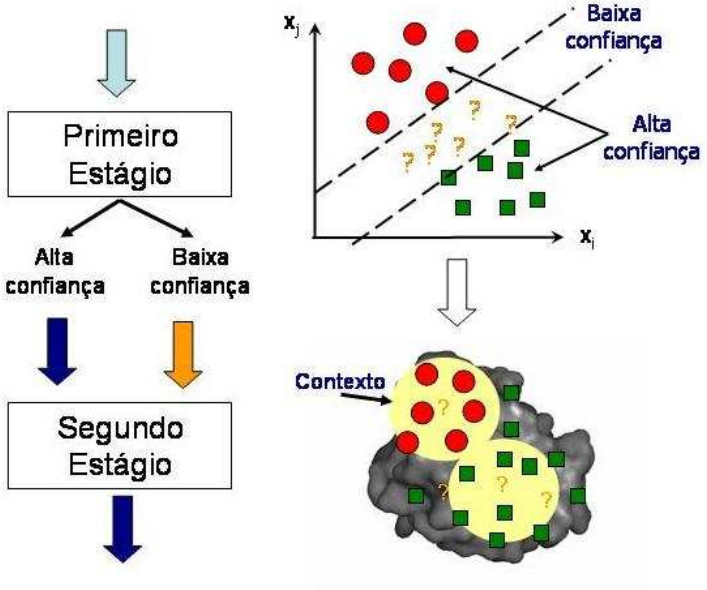

Figura 4.1: Estratégia de predição em dois estágios.

A estratégia proposta para o desenvolvimento do preditor é uma abordagem em dois estágios (Figura 4.1). No primeiro estágio, um classificador é utilizado para identificar com alta confiança os aminoácidos pertencentes ou não à interface. Aqueles aminoácidos que não podem ser classificados

com alta confiança têm sua atribuição a uma das duas classes postergada para o segundo estágio.

O vetor de características utilizado neste estágio consiste de um conjunto de medidas de propriedades físico-químicas e estruturais empregadas na análise estrutural de proteínas (vide capítulo 2). Ele possui 56 dimensões, com 28 deles correspondendo a propriedades físico-químicas e estruturais, medidas sobre o aminoácido, e 28 correspondendo à média dessas mesmas propriedades considerando o conjunto de aminoácidos em sua vizinhança. É assumido que a vizinhança de um aminoácido é formada pelo conjunto de aminoácidos obtidos da estrutura de dados *alpha shapes* (Liang et al., 1998), utilizada para o cálculo de SAS e MS (vide capítulo 2). Desta forma, o vetor de características compreende:

x_1, x_2**:** os dois índices (Hagerty et al., 1999), derivados da base de dados Aaindex (Kidera et al., 1985), utilizados para representar os vinte tipos padrões de aminoácidos.

x_3, x_4, x_5**:** a área acessível a solvente (SAS), a superfície molecular (MS) e a área acessível a solvente relativa (rSAS).

$x_6, \ldots, x_{11}$**:** a energia de solvatação por aminoácido considerando três conjuntos diferentes de ASPs (Eisenberg e McLachlan, 1986; Wesson e Eisenberg, 1992) e as respectivas energias de solvatação por unidade de área.

$x_{12}, \ldots, x_{16}$**:** o *residue depth* (RD), a *half sphere exposure* (HSE) e o *coordination number* (CN).

$x_{17}, \ldots, x_{23}$**:** as curvaturas principais (k_1 e k_2), as curvaturas média e gaussiana (H e K), o índice de forma (S) e a *curvedness* (R).

x_{24}, x_{25}**:** o potencial eletrostático e a contribuição eletrostática para a energia de solvatação.

x_{26}, x_{27}, x_{28}**:** as medidas de grau de conservação de aminoácidos conteúdo de informação, entropia relativa e pressão evolucionária.

As medidas $x_{29}, ..., x_{56}$ do vetor de características correspondem aos valores médios das propriedades x_1 a x_{28} na vizinhança de cada aminoácido. Todas as propriedades, com exceção do tipo de aminoácido, são normalizadas por proteína, calculando o correspondente z-*score*.

No segundo estágio os aminoácidos, cujas classificações foram postergadas no estágio anterior, são classificados observando a vizinhança na qual eles estão inseridos (dependência de contexto). Dado que os aminoácidos de interface formam uma região, eles estão, necessariamente, próximos uns dos outros. Assim, um aminoácido localizado numa região da superfície da proteína, cuja vizinhança possui muitos aminoácidos pertencentes à região de interface, tem boas chances de também pertencer à região de interface. A idéia é que aqueles aminoácidos não classificados com alta confiança no

primeiro estágio, mas cercados por aminoácidos identificados como pertencentes à região de interface também sejam classificados como pertencentes à interface. Os demais são classificados como não pertencentes à região de interface.

Em adição, visando comparar o desempenho do preditor proposto com o dos preditores propostos por Bradford e Westhead (2005) e Bradford et al. (2006), os aminoácidos de interface identificados são separados em grupos e ordenados de acordo com o número de aminoácidos identificados no primeiro estágio. Como Bradford e Westhead (2005) e Bradford et al. (2006) relatam como resultado os *patches* preditos com maior confiança, a idéia é compará-los com os grupos com maior número de aminoácidos identificados com alta confiança pelo preditor proposto.

4.3 Metodologia de desenvolvimento

4.3.1 Conjunto de dados

Para o desenvolvimento do preditor de regiões de interface, é necessário um conjunto de estruturas de complexos de proteínas, tal que os aminoácidos pertencentes à região de interface possam ser identificados através de um critério estrutural. Este conjunto deve ser formado a partir das estruturas de complexos de proteínas depositadas no PDB (vide capítulo 2), removendo aquelas que não satisfaçam a critérios específicos como comprimento mínimo da estrutura primária, alta resolução e tipo de interação. Além disso, o conjunto de dados deve ser não redundante, evitando a inclusão de estruturas de proteínas muito similares.

Conforme apresentado na seção 4.1, existem muitos métodos para predição de regiões de interface relatados na literatura. Em todos os casos, o conjunto de dados utilizado para desenvolvimento do preditor é derivado do PDB (vide capítulo 2), mas os critérios utilizados são muito variados, o que resulta em conjuntos de dados distintos.

Por isso, neste trabalho, optou-se por utilizar o conjunto de dados compilado por Bradford e Westhead (2005) para o desenvolvimento do preditor de regiões de interface. Este conjunto de dados é representativo da diversidade de tipos de interações proteína-proteína, incluindo interações permanentes e transientes. Apesar de existirem conjuntos de dados alternativos, compilados de forma automática ou semi-automática, este conjunto de dados tem como vantagem o fato de incluir apenas estruturas para as quais a ocorrência da interação *in vivo* foi verificada na literatura pelos autores. Além disso, todas as estruturas consideradas possuem resolução superior a 3Å e o critério considerado para remoção de redundâncias (máximo de 20% de identidade entre qualquer par de proteínas no conjunto de dados) está entre os mais restritivos dentre os relatados na literatura sobre preditores de regiões de interface proteína-proteína. Esta escolha permitiu não só economizar o tempo necessário

para compilar um conjunto de dados próprio, como também viabilizou a comparação direta dos resultados obtidos neste trabalho com aqueles relatados por Bradford e Westhead (2005) e Bradford et al. (2006). Das 180 proteínas do conjunto de dados originalmente compilado por Bradford e Westhead (2005), oito foram descartadas por não ter sido possível calcular para seus aminoácidos as medidas de conservação. Assim, o conjunto de dados efetivamente utilizado possui 172 estruturas de proteínas não redundantes.

Para rotular os dados, o seguinte critério foi utilizado: os aminoácidos, cuja área acessível a solvente é maior que zero (SAS > 0), são considerados como expostos na superfície da proteína. Cada aminoácido exposto na superfície é considerado como pertencente ou não à região de interface, de acordo com a perda de área acessível a solvente (ΔSAS = $SAS_{isolado}$ - $SAS_{complexo}$) no processo de formação do complexo. Se ΔSAS é maior que zero, o aminoácido é rotulado como "interface"; e em caso contrário, como "não interface". Isso resultou em 37.758 aminoácidos expostos a solvente (na superfície da molécula), dos quais 8.642 (22.8%) foram rotulados como "interface" e 29.116 (77.2%) foram rotulados como "não interface".

4.3.2 Cálculo das propriedades físico-químicas e estruturais

As propriedades físico-químicas e estruturais que formam o vetor de características utilizado pelo preditor de regiões de interface foram apresentadas no capítulo 2. Para o cálculo dessas propriedades, sempre que possível, foi utilizado um software de domínio público. Nos casos em que não foi possível utilizar uma implementação pública, o algoritmo para cálculo da propriedade foi implementado utilizando a linguagem de programação Python (Lutz, 1996).

Para o cálculo da área acessível a solvente (SAS) e da superfície molecular (MS) foi utilizado o programa *volbl*, incluído no pacote de software *alpha shapes* (Liang et al., 1998), considerando um probe de raio 1.4 Å e o conjunto de raios de átomos fornecido pelo pacote. A opção "*outside fringe*" foi usada para evitar a possibilidade de descontinuidades na região de interface, devido a cavidades introduzidas na formação do complexo, como pode acontecer quando limiares arbitrários de distâncias ou de ΔSAS são utilizados. Um programa Python foi desenvolvido para calcular a área acessível a solvente relativa (rSAS), utilizando os valores de SAS para cada aminoácido em seu estado estendido, de acordo com os valores relatados por Ahmed et al. (2004).

Um programa Python foi desenvolvido para calcular a energia de solvatação por átomo. Seu cálculo consiste, basicamente, em multiplicar o SAS do átomo pelo correspondente parâmetro de solvatação atômico (ASP). Três diferentes conjuntos de ASPs foram considerados pelo programa (Eisenberg e McLachlan, 1986; Wesson e Eisenberg, 1992). É assumido que a contribuição por átomo é aditiva, tal que a energia de solvatação por aminoácido é calculada através da soma das contribuições de seus correspondentes átomos. Além disso, para cada conjunto de ASP considerado,

o programa também calcula a energia de solvatação por unidade de área.

Para calcular o *residue depth* (RD), *half sphere exposure* (HSE) e o *coordination number* (CN) foi utilizado o módulo Bio.PDB do pacote biopython (Hamelryck e Manderick, 2003).

Para descrever a geometria de um aminoácido na superfície da proteína, considera-se o conjunto de átomos formado pelos átomos do aminoácido que estão expostos na superfície da proteína; e todos os átomos de outros aminoácidos que estão expostos na superfície da proteína e a uma distância inferior a 10 Å de qualquer dos átomos do aminoácido. As coordenadas xyz destes átomos foram utilizadas para calcular os seguintes parâmetros geométricos: curvatura gaussiana, curvatura média, curvaturas principais, *curvedness*, índice de forma e índice de planaridade. Um programa Python foi desenvolvido para calcular estes parâmetros, conforme descrito nas seções 2.3.3 e 2.3.4 do capítulo 2.

Para calcular o potencial eletrostático e a contribuição eletrostática para a energia de solvatação foi utilizado o software *apbs* (Baker et al., 2001) com átomos de hidrogênio adicionados à estrutura da proteína através do software pdb2pqr.py (Dolinsky et al., 2004). Considerou-se uma constante dielétrica de 78.54 para o solvente e de 2.0 para o interior da proteína, com uma concentração iônica de 150 mM. O potencial eletrostático por aminoácido foi calculado como a média do potencial eletrostático correspondente ao ponto do *grid* mais próximo de cada átomo na superfície da proteína. De maneira similar, a contribuição eletrostática para a energia de solvatação por aminoácido foi calculada somando-se as correspondentes energias por átomo, fornecidas pelo software *apbs* (Baker et al., 2001).

Para medir o grau de conservação de aminoácidos, o seguinte procedimento foi implementado em Python: primeiro o software Blast (Altschul et al., 1997), com matriz de substituição BLOSUM62 e "expect value" = 0.1, foi utilizado contra a base de dados Swissprot/Uniprot knowledgebase, release 9.6 (Apweiler et al., 2004), para encontrar seqüências similares de proteínas. Em seguida, as seqüências encontradas foram filtradas, de acordo com o limiar HSSP (Rost, 1999), tal que apenas as seqüências homólogas foram mantidas. Neste processo de filtragem, oito proteínas do conjunto de dados original foram eliminadas por não possuírem o número mínimo de cinco seqüências homólogas. O conjunto de seqüências homólogas resultante, para cada proteína, foi utilizado para construir um alinhamento múltiplo de seqüências (MAS), através do software ClustalW (Thompson et al., 1994), utilizando a série de matrizes de substituição BLOSUM, "*gapopen*" = 3.0 e "*gap ext*" = 0.1. Finalmente, três medidas de grau de conservação de aminoácidos foram calculadas: o conteúdo de informação, conforme implementado no módulo Bio.PDB de biopython, a entropia relativa (Rost, 1999) e a pressão evolucionária (Pupko et al., 2002).

4.3.3 Critério de avaliação de desempenho

O critério de avaliação escolhido tem por objetivo permitir a comparação direta com o dos preditores propostos por Bradford e Westhead (2005) e Bradford et al. (2006) e consiste em medir a taxa de sucesso em localizar as regiões de interface, onde o sucesso para uma proteína é definido em função das medidas de precisão e cobertura. Especificamente, uma predição é considerada como sucesso se ao menos um dos grupos de aminoácidos preditos apresenta cobertura $\geq 20\%$ e precisão $\geq 50\%$.

4.3.4 Desenvolvimento do preditor

O primeiro estágio é implementado como um classificador com opção de rejeição. Para escolha do método de classificação utilizado foram testados os classificadores bayesiano linear, bayesiano quadrático e do vizinho mais próximo (1-NN). Métodos de classificação mais complexos como k vizinhos mais próximos (k-NN, k > 1), Parzen e SVM (vide capítulo 3) foram evitados devido ao alto custo computacional para treinar estes classificadores utilizando um conjunto de treinamento de aproximadamente 38.000 aminoácidos. Dentre os métodos avaliados, aquele que apresentou o melhor desempenho foi o bayesiano linear. Embora, a questão de seleção do método de classificação mereça uma investigação mais profunda, neste momento, optou-se por adotar o classificador bayesiano linear para o restante do desenvolvimento do preditor e postergar uma investigação mais profunda a respeito da seleção do método de classificação para uma oportunidade futura.

Os parâmetros do classificador são estimados assumindo-se distribuições gaussianas multivariadas com uma matriz de covariâncias comum (classificador linear).

Uma matriz de custo é utilizada tanto para implementar a opção de rejeição (vide capítulo 3) quanto para ajustar o custo de erro de classificação. Estes valores foram ajustados através de um procedimento empírico, sendo que os seguintes fatores foram levados em consideração:

- o conjunto de dados utilizado para treinamento possui um número de exemplos negativos bem maior que o número de exemplos positivos. Assim, um *bias* em favor dos exemplos positivos é introduzido, mantendo-se o custo de atribuir um aminoácido pertencente à classe negativa para a classe positiva sempre maior que o custo de atribuir um aminoácido pertencente à classe positiva para a classe negativa; e

- o mesmo custo de rejeição é considerado para ambas as classes positiva e negativa.

O critério considerado pelo procedimento empírico para seleção desses valores foi a taxa de sucesso (vide 4.3.3). Ao final da execução desse procedimento empírico, os seguintes valores de custo foram obtidos:

- zero para o custo de atribuir um aminoácido para a classe correta;

- dois para o custo de atribuir um aminoácido pertencente à classe positiva para a classe negativa;
- três para o custo de atribuir um aminoácido pertencente à classe negativa para a classe positiva;
- um para o custo de rejeição de um aminoácido, independente de sua classe verdadeira.

Também foi avaliada a utilização de duas técnicas de redução de dimensionalidade, PCA (vide seção 3.3.1) e *Sequential Backward Selection* (vide seção 3.3.3). Neste caso, o objetivo foi verificar se o classificador utilizado no primeiro estágio não estava sujeito ao fenômeno de *peaking* (Jain et al., 2000), tal que fosse necessário proceder uma redução da dimensionalidade dos dados. Em ambos os casos, não foi observado um ganho de desempenho ao se utilizar técnicas para redução da dimensionalidade do vetor de características. Assim, optou-se por manter o vetor de características original na seqüência do desenvolvimento do preditor.

O segundo estágio foi implementado por um procedimento empírico executado em três passos. É assumido que os aminoácidos detectados no primeiro estágio pertencem à região núcleo da interface e que, portanto, formam grupos no espaço xyz. O primeiro passo consiste em identificar esses grupos. Para isso é utilizado o algoritmo *k-means* (vide apêndice A), considerando as coordenadas xyz dos aminoácidos como variáveis e inicializando o centróide de cada grupo aleatoriamente. A coordenada xyz do aminoácido é definido pelo centróide dos seus átomos que estão expostos na superfície da proteína. Para determinar o número ótimo de grupos, o valor médio da *silhouette value* (vide apêndice A) foi utilizado como critério. Foi determinado, previamente, que o número de grupos para a maioria das proteínas varia de 1 a 3. Dessa forma, os valores de *silhouette value* considerando até 3 grupos são avaliados. Ao avaliar os aminoácidos, cuja classificação foi postergada no primeiro estágio, considera-se que ele está próximo de um dos grupos de aminoácidos identificados pelo algoritmo *k-means* se a distância entre suas coordenadas xyz e o centróide do grupo é menor que um limiar pré-especificado. Assim, o segundo passo do procedimento consiste em estimar este limiar. Para isso, a área da região de interface da proteína em teste é estimada a partir de sua área de superfície total, através de um procedimento de regressão linear ajustado para as proteínas no conjunto de treinamento. O limiar foi definido, empiricamente, como 47% do raio da área circular correspondente à área estimada para a região de interface, sendo sua utilização similar à forma como Bradford e Westhead (2005), Bradford et al. (2006) e Jones e Thornton (1997b) estimam o tamanho dos *patches*. No terceiro passo do procedimento, os aminoácidos, cuja classificação foram postergadas no primeiro estágio, são aceitos como pertencentes à região de interface se eles estão próximos de um dos grupos de aminoácidos identificados no primeiro passo do procedimento. Um aminoácido é considerado próximo de um grupo de aminoácidos se sua distância para o centróide do grupo é inferior ao limiar calculado no passo dois e ele é vizinho de pelo menos dois aminoácidos do grupo. Ambos os parâmetros utilizados neste procedimento, o limiar de distância do centróide e o número mínimo de vizinhos,

foram ajustados empiricamente, variando-se seus valores e usando como crítério de avaliação a taxa de sucesso.

As predições são relatadas como grupos (um, dois ou três) de aminoácidos pertencentes à região de interface, ordenadas pelo número de aminoácidos de interface identificados no primeiro estágio.

Os dois estágios do classificador foram implementados em Matlab (Mathworks, 2004), sendo que no primeiro estágio foi utilizado o *toolbox* de reconhecimento de padrões PRTools (Duin et al., 2004).

4.4 Experimento

Para estimar a taxa de sucesso (vide seção 4.3.3) para o preditor desenvolvido, foi utilizado um procedimento de validação cruzada *leave-one(protein)-out* com dez repetições. Esta estratégia de repetições foi utilizada para estimar a variância da taxa de sucesso, devido à inicialização aleatória dos centróides pelo algoritmo *k-means*, utilizado no segundo estágio.

A título de ilustração, o preditor desenvolvido foi utilizado para identificar as regiões de interface de duas proteínas incluídas entre as 172 proteínas do conjunto de dados utilizado em seu desenvolvimento. Em ambos os casos, o preditor é treinado com as 171 proteínas que restam no conjunto de dados ao se remover a proteína utilizada na ilustração. Isto equivale a uma rotação do processo de validação cruzada utilizado para estimar a taxa de sucesso alcançada pelo preditor.

4.5 Resultados

O preditor desenvolvido apresentou uma taxa de sucesso de 82.1% com um desvio padrão de 0.25%, indicando um resultado próximo ao obtido por Bradford et al. (2006), que utilizam o mesmo conjunto de dados e critério de avaliação.

Bradford et al. (2006) relatam os *patches* preditos como interface, ordenando-os de acordo com a nível de confiança obtido do classificador. Eles, então, analisam como a taxa de sucesso se decompõem em função da ordenação dos *patches* e do tipo de interação proteína-proteína. Para comparar os resultados obtidos com os reportados por Bradford et al. (2006), os grupos preditos como interface foram ordenados de acordo com o número de aminoácidos preditos com alta confiança, ou seja, aqueles detectados no primeiro estágio. A correspondente decomposição dos resultados é apresentada na Tabela 4.2. Os valores apresentados correspondem a uma execução do processo de validação cruzada *leave-one(protein)-out*. As linhas da tabela referem-se aos tipos de interação. Para cada tipo de interação, a segunda coluna indica o número total de proteínas, a terceira indica número de proteínas, cuja interface foi predita com sucesso, e a coluna "Ordem" indica a posição do grupo satisfazendo a

Tipo de Interação	Número de Exemplos	Número de Sucessos	Ordem		
			1	2	3
Enzima-Inibidor	31	23	17	6	-
NEIT	29	24	16	6	2
Sub-Total	60	47	33	12	2
Homodímero	85	72	46	22	4
Heterodímero	27	23	17	6	-
Sub-Total	112	95	63	28	4
Total	172	142	96	40	6

Tabela 4.2: Detalhes sobre o desempenho do preditor. NEIT significa transiente-não inibidor-enzima. Os grupos de aminoácidos estão ordenados de acordo com o número de aminoácidos detectados no primeiro estágio. A coluna "Ordem" indica a posição do grupo satisfazendo à condição de sucesso. Adaptado de (Higa e Tozzi, 2008b).

condição de sucesso. O número de sucessos resultante da avaliação de cada um dos três grupos de aminoácidos preditos como interface é apresentado na tabela.

Com relação à ordenação dos grupos preditos, 55.8% (96/172) das predições bem sucedidas do preditor desenvolvido resultam da avaliação do grupo na primeira posição, enquanto que Bradford et al. (2006) relatam que 52.2% (94/180) das predições bem sucedidas resultam da avaliação do *patch* posicionado na primeira posição.

Com relação às predições por tipo de interação, o preditor proposto apresenta uma taxa de sucesso de 84.8% (95/112) para as iterações permanentes e de 78.3% (47/60) para as interações transientes. Estes resultados são bastante compatíveis com os relatados por Bradford et al. (2006): 84.2% (96/114) para as interações permanentes e 75.3% (52/69) para as interações transientes.

Para ilustrar sua utilização, o preditor desenvolvido foi utilizado para identificar as regiões de interface de duas proteínas incluídas entre as 172 proteínas do conjunto de dados utilizado em seu desenvolvimento. A primeira é a sub-unidade da proteína 3-hydroxy-methylglutaryl-CoA (HMG-CoA) redutase (Figura 4.2), uma enzima envolvida na biossíntese de colesterol em mamíferos (Tabernero et al., 1999). Dois grupos de aminoácidos são preditos para esta proteína, um apresentando cobertura de 38.13% e precisão de 79.1% e outro apresentando cobertura de 25.9% e precisão de 66.67%. Os dois grupos, em conjunto, se sobrepõem à maior parte da área de interface da proteína, conforme pode ser observado na Figura 4.2.

A segunda proteína apresentada é a barstar (Figura 4.3), o inibidor natural de uma proteína da família RNase chamada barnase, uma enzima extracelular do organismo *Bacyllus amyloliquefaciens* (Sevcik et al., 1998). Neste caso, três grupos de aminoácidos são preditos, mas apenas um deles (o segundo na ordenação dos grupos) satisfaz às condições de sucesso. Ele apresenta cobertura de 26.32% e precisão de 100%. Embora os outros dois grupos não satisfaçam à condição de sucesso,

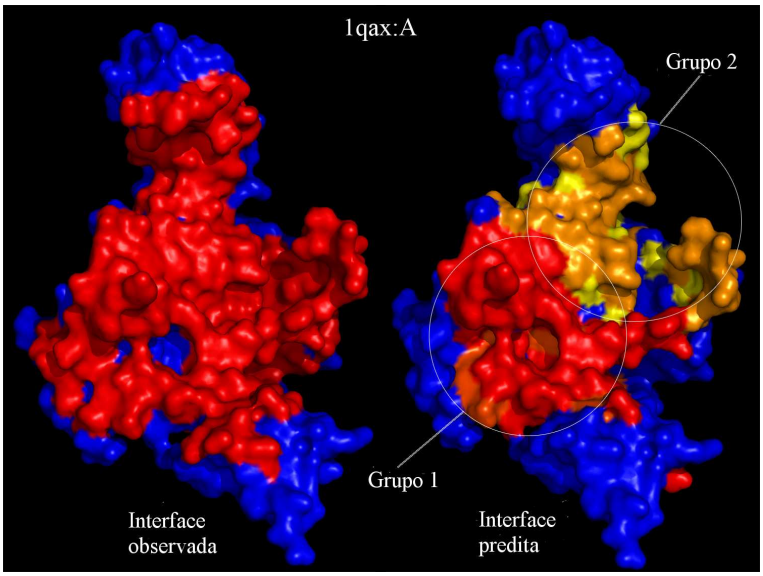

Figura 4.2: Região de interface observada (esquerda) e predita (direita) para a proteína 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) redutase (PDB ID - 1qax:A). Adaptado de (Higa e Tozzi, 2008b).

quando os três grupos são considerados, em conjunto, eles se sobrepõem à maior parte da área da região de interface da proteína. Esta aplicação mostra que, mesmo não satisfazendo às condições de sucesso, os grupos de aminoácidos preditos como pertencentes à região de interface podem fornecer uma boa indicação de sua localização quando analisados em conjunto com os demais grupos.

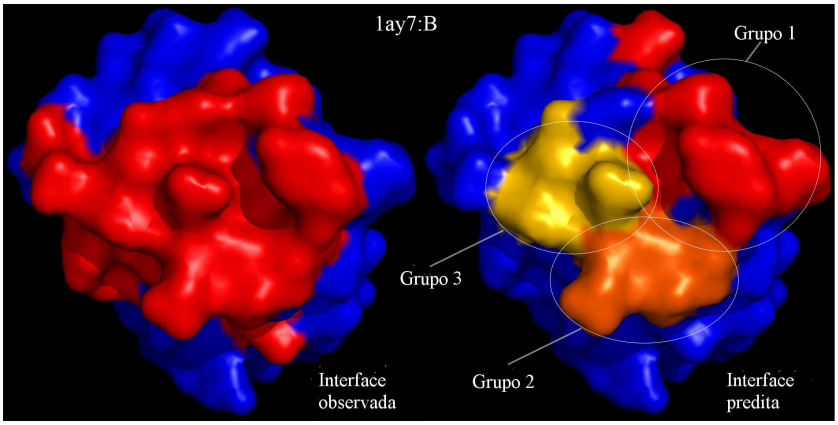

Figura 4.3: Região de interface observada (esquerda) e predita (direita) para a proteína barstar (PDB ID - 1ay7:B). Adaptado de (Higa e Tozzi, 2008b).

4.6 Discussão

Em função das escolhas realizadas ao longo do desenvolvimento deste preditor, é conveniente discutir as similaridades e diferenças entre o preditor proposto e os trabalhos de Bradford e Westhead (2005) e Bradford et al. (2006), em termos de metodologia e resultados. Ambos os preditores utilizam um conjunto de propriedades físico-químicas e estruturais para realizar as predições, relatando os resultados como um conjunto de aminoácidos. A principal diferença entre os preditores refere-se ao objeto descrito pelos vetores de características. Bradford e Westhead (2005) e Bradford et al. (2006) seguem a metodologia de análise de *patch*, introduzida por Jones e Thornton (1997b). O conjunto de propriedades é projetado sobre pontos distribuídos ao longo da superfície molecular e *patches* circulares são amostrados nesta superfície. O vetor de características é formado pelas médias e desvios padrões das medidas sobre os *patches*. Por fim, cada *patch* é classificado como "interface" ou "não interface" e as três predições com maior nível de confiança são consideradas.

Já o preditor desenvolvido, considera o aminoácido da superfície da proteína como a unidade básica a ser classificada. A classificação é realizada por um processo formado por dois estágios. No primeiro estágio, um classificador linear com opção de rejeição considera um vetor de características formado por um conjunto de propriedades físico-químicas e estruturais. No segundo estágio, novos aminoácidos são incorporados ao conjunto de aminoácidos preditos como de interface, utilizando informação contextual. De maneira similar a Bradford e Westhead (2005) e Bradford et al. (2006), os aminoácidos preditos como de interface são relatados como grupos de aminoácidos. Contudo, neste caso, eles não formam, necessariamente, uma região de formato circular (*patches*).

Com relação ao método de classificação utilizado, apesar de Bradford e Westhead (2005) utilizarem um método de classificação teoricamente mais poderoso (SVM), é interessante notar que o desempenho que eles obtiveram foi inferior ao alcançado pelo preditor proposto. Também é interessante notar que, posteriormente, os mesmos autores (Bradford et al., 2006) relataram um desempenho similar ao obtido pelo preditor desenvolvido (taxa de sucesso igual a 82%) utilizando um classificador teoricamente mais simples, o classificador näive bayes (Duda et al., 2001).

4.7 Considerações finais

Neste capítulo foi apresentado um preditor de regiões de interface, onde a restrição de utilização do conceito de *patches*, utilizada por preditores que se baseiam em medidas físico-químicas e estruturais, é relaxada. O preditor desenvolvido é formado por dois estágios. No primeiro estágio, aminoácidos são identificados, com alta confiança, como pertencentes à região de interface; utilizando um classificador bayesiano com opção de rejeição. No segundo estágio, os aminoácidos,

cujas classificações foram postergadas (rejeitadas) no primeiro estágio, são classificados através de um procedimento empírico que considera os aminoácidos na sua vizinhança (dependência de contexto). A taxa de sucesso alcançada pelo preditor desenvolvido é de 82.1% das proteínas testadas, um desempenho bastante competitivo comparado com o melhor desempenho relatado na literatura quando tanto o conjunto de dados quanto o critério de avaliação de desempenho foram os mesmos (taxa de sucesso = 82%) (Bradford e Westhead, 2005; Bradford et al., 2006).

No próximo capítulo é apresentado um classificador para predizer *hot spots* dentre os aminoácidos da região de interface. Considerando que, do ponto de vista energético, os *hot spots* são os aminoácidos mais importantes da região de interface, a idéia é discriminá-los dentre aqueles preditos como pertencentes à região de interface, complementando a informação fornecida pelo preditor apresentado neste capítulo.

Capítulo 5

Predição de *hot spots* dentre os aminoácidos da região de interface

Neste capítulo, é apresentado um preditor de *hot spots* dentre os aminoácidos da região de interface. Diferente de outros preditores de *hot spots* descritos na literatura, o preditor desenvolvido baseia-se na utilização de um conjunto de propriedades físico-químicas e estruturais, cujo cálculo não depende do conhecimento da estrutura do complexo formado pela proteína de interesse e sua parceira de interação. Isto permite a sua utilização em complemento ao preditor de regiões de interface apresentado no capítulo 4, qualificando os aminoácidos preditos como de região de interface quanto à sua contribuição para a energia de ligação (*hot spots* ou não *hot spots*).

Inicialmente, uma breve revisão bibliográfica sobre o tema é apresentada, visando posicionar o preditor proposto em relação à literatura. Em seguida, é apresentada a estratégia utilizada para construção do preditor. Então, utilizando as técnicas de classificação de padrões introduzidas no capítulo 3, é apresentada a metodologia de desenvolvimento do preditor, e os experimentos realizados para avaliar seu desempenho. As propriedades utilizadas na formação do vetor de características foram apresentadas no capítulo 2.

5.1 Trabalhos relacionados

Estudos experimentais utilizando uma técnica de mutação sítio-dirigido, denominada *alanine scanning*, revelaram que a energia de interação proteína-proteína não se distribui uniformemente entre os aminoácidos da região de interface (Clackson e Wells, 1995). De fato, uma significativa fração desta energia deve-se a um subconjunto dos aminoácidos da região de interface, conhecidos como *hot spots* (Moreira et al., 2007; Bogan e Thorn, 1998).

Dado à importância dos *hot spots* e o custo envolvido em sua determinação experimental, a uti-

lização de métodos computacionais para a sua predição dentre os aminoácidos da região de interface aparece como uma alternativa atraente, pois permite focar os procedimentos experimentais naqueles aminoácidos com maiores chances de serem *hot spots* (Darnell et al., 2007).

A maioria dos métodos computacionais para predição de *hot spots* utiliza um modelo físico que avalia o impacto sobre a energia de interação, devido a mutações específicas dentro da região de interface (Kortemme e Baker, 2002; Moreira et al., 2006). Uma segunda linha de investigação analisa propriedades físico-químicas e estruturais com o objetivo de diferenciar *hot spots* dentre os aminoácidos da região de interface. Bogan e Thorn (1998) relataram que os *hot spots* apresentam uma tendência de se aglutinarem na região central da interface, formando uma estrutura que eles denominaram de *O-ring*, caracterizada por um grupo de aminoácidos polares protegidos por um anel formado por aminoácidos hidrofóbicos. Eles também analisaram a propensão de diferentes tipos de aminoácidos de serem ou não *hot spots*, concluindo que o triptofano, a tirosina e a argenina são os tipos de aminoácidos que apresentam as maiores propensões. Outra propriedade importante associada a *hot spots* é o grau de conservação de aminoácidos. Hu et al. (2000) caracterizaram *hot spots* como aminoácidos polares conservados, enquanto Ma et al. (2003) os caracterizaram como aminoácidos estruturalmente conservados. A partir deste estudo, Li et al. (2004) também analisaram a organização geométrica de aminoácidos estruturalmente conservados, concluindo que a maioria dos *hot spots* são encontrados em regiões da interface caracterizadas por uma reentrância em uma das proteínas complementada por uma parte protuberante de um aminoácido da segunda proteína.

Já Guney et al. (2007) predizem *hot spots* usando o grau de conservação de aminoácidos e ΔSAS, utilizando um método desenvolvido de forma empírica, enquanto Ban et al. (2006) aplicam um método puramente geométrico, predizendo *hot spots* como aminoácidos localizados em partes da região de interface protegidas da periferia. Contudo, apenas recentemente, Darnell et al. (2007) abordaram o problema de predição de *hot spots* dentre os aminoácidos da região de interface através de uma perspectiva de análise discriminante. Eles compilaram um conjunto de dados de alta qualidade e não redundante, com aminoácidos da região de interface contendo tanto informações estruturais quanto informações experimentais de *alanine scanning*, a partir das quais os *hot spots* podem ser identificados e rotulados como tal. O melhor preditor que eles obtiveram utiliza propriedades estruturais, físico-químicas e energéticas, combinadas através de uma regra simples do tipo *OR*. Para este classificador, eles relataram um desempenho de 55%, mensurado pela medida F, correspondendo a cobertura de 72% e precisão de 44%. Utilizando o mesmo conjunto de dados e uma estratégia diferente de combinação de classificadores, Higa e Tozzi (2008a) relataram um desempenho marginalmente superior, correspondendo a um valor de medida F de 56.5%.

Todos os preditores de *hot spots* apresentados acima pressupõem que a estrutura do complexo formado pela proteína e sua parceira de interação seja conhecida. Portanto, estes preditores não

podem ser utilizados em complemento à utilização de preditores de regiões de interface, uma vez que esses são utilizados exatamente quando a estrutura do complexo não é conhecida.

Assim, este capítulo refere-se a ao desenvolvimento de um preditor de *hot spots* que possa ser utilizado em conjunto com o preditor de regiões de interface apresentado no capítulo 4, desta forma, provendo um refinamento na caracterização dos aminoácidos pertencentes à região de interface.

5.2 Estratégia de predição proposta

A estratégia proposta para o desenvolvimento do preditor consiste na utilização de um classificador para identificação dos *hot spots*. A curva ROC (vide seção 3.4 do capítulo 3) associada ao preditor também é determinada, de forma a indicar os diferentes pontos de operação em que ele pode ser utilizado.

Para formar o vetor de características, utiliza-se um conjunto de medidas de propriedades físico-químicas e estruturais, apresentado no capítulo 2 e empregadas na análise estrutural de proteínas. Desta forma, o vetor de características compreende:

x_1, x_2**:** os dois índices (Hagerty et al., 1999), derivados da base de dados Aaindex (Kidera et al., 1985), utilizados para representar os vinte tipos padrões de aminoácidos.

$x_3, \ldots, x_{22}$**:** o perfil evolucionário.

x_{23}**:** o grau de conservação de aminoácidos estimado pela pressão evolucionária.

$x_{24}, \ldots, x_{34}$**:** a área acessível a solvente (SAS), a superfície molecular (MS), área acessível a solvente relativa (rSAS), a energia de solvatação calculada considerando quatro conjuntos diferentes de parâmetros de solvatação atômica (ASP) (Eisenberg e McLachlan, 1986; Wesson e Eisenberg, 1992; Fernández-Recio et al., 2004) e as correspondentes energias de solvatação ponderadas pelo valor de SAS do aminoácido.

$x_{35}, \ldots, x_{41}$**:** as curvaturas principais (k_1 e k_2), as curvaturas média e gaussiana (K e H), o índice de forma (S), a *curvedness* (R) e o índice de planaridade.

x_{42}, x_{43}**:** os ângulos diedrais, ϕ e ψ.

5.3 Metodologia de desenvolvimento

5.3.1 Conjunto de dados

Para o desenvolvimento do preditor de *hot spots* é necessário um conjunto de dados formado por aminoácidos pertencentes à região de interface de proteínas com estruturas conhecidas e que se qualificam como "*hot spots*" ou "*não hot spots*". O repositório, que disponibiliza este tipo de informação, mais utilizado em estudos, apresentados na literatura, envolvendo *hot spots* é o AseDB (Bogan e Thorn, 1998). Contudo, para formar um conjunto de dados adequado ao desenvolvimento do preditor, os dados desta base de dados devem passar por um processo de filtragem a fim de obter um conjunto com dados de estruturas com alta resolução e não redundantes.

Seguindo a mesma estratégia utilizada no capítulo 4, optou-se por utilizar um conjunto de dados previamente compilado por outros autores, especificamente o compilado por Darnell et al. (2007). Desta forma, além de economizar o tempo necessário para compilar um conjunto de dados próprio, também é possível realizar uma comparação direta dos resultados obtidos neste trabalho com aqueles relatados por Darnell et al. (2007). Uma vez que a quantidade de dados sobre regiões de interfaces, caracterizadas tanto estruturalmente quanto do ponto de vista de sua contribuição energética para a interação proteína-proteína, é bastante limitada, este conjunto de dados constitui um dos mais representativos para analisar *hot spots*. Ele é composto por aminoácidos pertencentes à região de interface oriundos de duas fontes: a base de dados AseDB (Bogan e Thorn, 1998), acrescido de um conjunto de dados publicado por Kortemme e Baker (2002), com as correspondentes energias livre de interação ($\Delta\Delta$G) resultantes de experimentos de mutação para alanina. O critério utilizado para definir um aminoácido como pertencente à região de interface é o de que a distância entre um de seus átomos e um átomo da proteína com que ela interage seja menor ou igual a 4Å. Darnell et al. (2007) limitaram os dados às estruturas, cuja resolução é superior a 3Å, e para remover redundâncias mantiveram apenas estruturas para as quais a identidade em seqüência para qualquer outra seqüência do conjunto de dados fosse inferior a 35%.

Além disso, também foram removidos do conjunto de dados original aqueles aminoácidos para os quais não foi possível calcular a propriedade de grau de conservação. Um total de 15 aminoácidos foi removido, tal que o conjunto de dados efetivamente utilizado contém 233 aminoácidos, dos quais 24% correspondem a *hot spots*. Cada aminoácido foi rotulado como "*hot spot*" se o correspondente $\Delta\Delta$G relatado é maior ou igual a 2.0 kcal/mol. Em caso contrário, ele é rotulado como "não *hot spot*".

5.3.2 Cálculo das propriedades físico-químicas e estruturais

As propriedades físico-químicas e estruturais que formam o vetor de características utilizado pelo preditor de *hot spots* foram apresentadas no capítulo 2. Para o seu cálculo foi utilizada a mesma estratégia descrita na seção 4.3.2 do capítulo 4, referente ao desenvolvimento do preditor de regiões de interface.

Em adição ao cálculo das propriedades utilizadas no desenvolvimento do preditor de regiões de interface, também foram calculados:

- o perfil evolucionário, obtido através de um programa Python desenvolvido para este fim. Para isso, é utilizado o mesmo MSA utilizado para calcular o grau de conservação de aminoácidos.
- os ângulos diedrais, ϕ e ψ, correspondentes a cada aminoácido na superfície da proteína, calculados com a utilização do software Stride (Frishman e Argos, 1995).

Além disso, as seguintes propriedades, utilizadas no desenvolvimento do preditor de regiões de interface, não foram consideradas no desenvolvimento do preditor de *hot spots*: potencial eletrostático, HSE, RD, CN e as medidas na vizinhança do aminoácido. Dado o tamanho reduzido do conjunto de dados utilizado, estas modificações visam limitar a dimensão do vetor de características. Pois, de acordo com o fenômeno de *peaking* (Jain et al., 2000), o desempenho de um classificador é limitado pela interrelação entre a sua complexidade, o tamanho do conjunto de dados e o número de medidas utilizadas.

5.3.3 Critério de avaliação de desempenho

O critério utilizado para avaliação de desempenho do preditor de *hot spots* é a área sob a curva ROC (*A*rea *U*nder ROC *C*urve - AUC) (vide seção 3.4 do capítulo 3), que sumariza seu desempenho considerando os diferentes pontos de operação definidos pela curva. Esse foi o critério utilizado na escolha do método de classificação utilizado no desenvolvimento do preditor.

Para comparar o desempenho do preditor desenvolvido com aquele relatado por Darnell et al. (2007), os critérios utilizados são a cobertura, a precisão e a medida F (vide seção 3.4 do capítulo 3). Duas situações são analisadas:

- na primeira, comparam-se os valores de cobertura e medida F no ponto de operação da curva ROC correspondente ao nível de precisão relatado por Darnell et al. (2007).
- na segunda, comparam-se os valores de cobertura, precisão e medida F no ponto de operação da curva ROC em que o preditor desenvolvido apresenta o valor máximo para a medida F.

5.3.4 Desenvolvimento do preditor

Para escolha do método de classificação utilizado pelo preditor de *hot spots*, foram testados os classificadores bayesiano linear, bayesiano quadrático, dos k-vizinhos mais próximos (k-NN), para k = 1 e 3, de Parzen e SVM com função *kernel* de base radial e saída calibrada para representar a probabilidade *a posteriori* (vide seção 3.2.2 do capítulo 3). O critério de desempenho utilizado foi o AUC.

Dentre os métodos de classificação avaliados, o classificador SVM foi o que apresentou o melhor desempenho, tendo sido adotado no restante do processo de desenvolvimento do preditor. Embora, a questão de seleção do método de classificação mereça maior investigação, optou-se por adotar o classificador SVM para o restante do desenvolvimento do preditor e postergar uma investigação mais profunda a respeito da seleção do método de classificação para uma oportunidade futura.

Para selecionar os parâmetros da função *kernel*, γ, e do parâmetro de regularização, C, foi utilizado um procedimento de busca em *grid* com validação cruzada utilizando o conjunto de dados de treinamento. Ao final, os seguintes parâmetros foram utilizados: C = 0.03125 e γ = 0.0078125.

Para o treinamento do classificador SVM, as propriedades que compõe o vetor de características foram mapeadas para o intervalo [-1, 1]. Estes mesmos mapeamentos foram utilizados para os dados do conjunto de teste.

Além disso, uma vez que o método de classificação utilizado pelo preditor, SVM , embute mecanismos para prevenir a degradação de desempenho associado ao aumento da dimensionalidade do vetor de características, nenhuma tentativa de redução de dimensionalidade, utilizando as técnicas apresentadas na seção 3.3 do capítulo 3, foi realizada.

Para implementar o preditor foi utilizado o software LibSVM (Chang e Lin, 2001), que implementa o classificador SVM e o método de Platt (Platt, 2000) para calibração da saída do classificador SVM como probabilidade *a posteriori*.

5.4 Experimento

Para determinar a curva ROC e o valor de AUC para o preditor de *hot spots* (vide seção 5.3.3) foi utilizado um processo de valiação cruzada *5-fold* estratificado com cem repetições. Tanto a curva ROC quanto o valor de AUC relatados correspondem ao valor médio obtido destas repetições.

A título de ilustração, o preditor desenvolvido foi utilizado para identificar *hot spots* em duas proteínas não incluídas no conjunto de dados utilizado em seu desenvolvimento. Em ambos os casos, o preditor é treinado utilizando todo o conjunto de dados descrito na seção 5.3.1 com os parâmetros de regularização e *kernel* ajustados da mesma forma que no procedimento de validação cruzada.

5.5 Resultados

A curva ROC correspondente ao comportamento médio do preditor é apresentado na Figura 5.1(a). O correspondente valor de AUC é de 0.8386 (+/-0.0380) e representa a probabilidade de o classificador produzir uma ordenação com uma amostra positiva (*hot spots*) posicionada à frente de uma negativa (não *hot spots*), ambas escolhidas de forma aleatória.

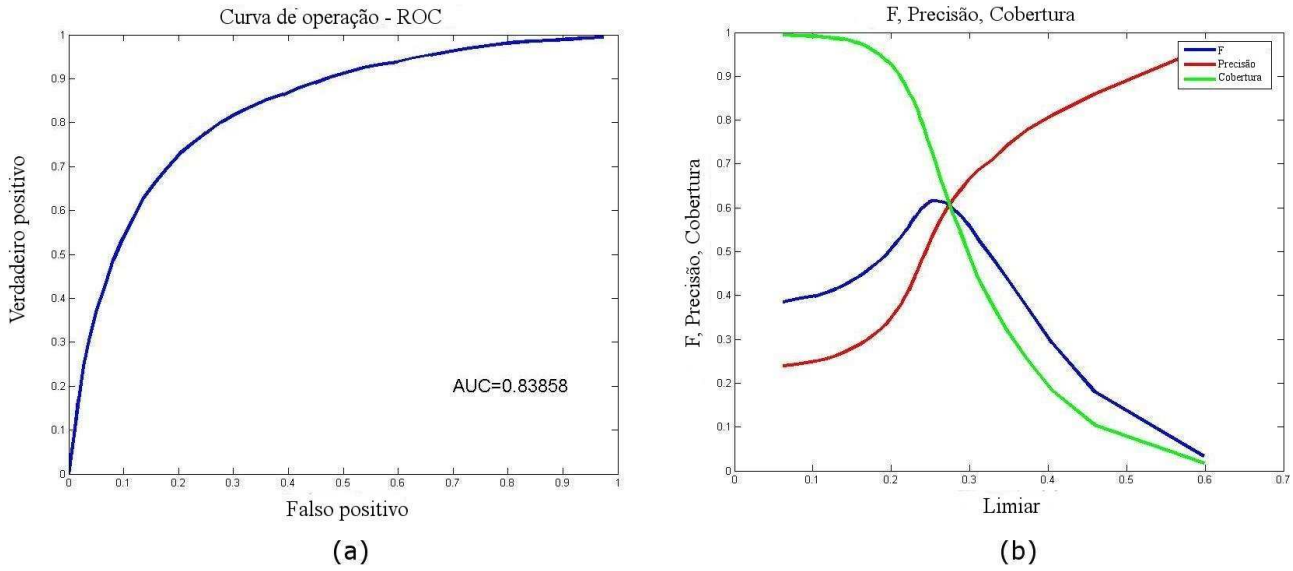

Figura 5.1: (a) Curva ROC. (b) Curvas F/Precisão/Cobertura x Limiar. Adaptado de (Higa e Tozzi, 2009).

Em seu trabalho, Darnell et al. (2007) utilizaram uma árvore de decisão como classificador e relataram uma precisão de 44% e uma cobertura de 72%, correspondendo a um valor de medida F de 55%. Neste nível de precisão, o preditor desenvolvido apresenta uma cobertura de 83.8% (+/- 5.1), correspondendo a uma medida F de 57.9% (+/-3.7). A Figura 5.1(b) mostra como a precisão, a cobertura e a medida F variam em função do limiar utilizado para definir a curva ROC, tal que o usuário pode utilizar estas medidas para escolher o ponto de operação mais apropriado para a sua aplicação. Por exemplo, escolhendo o ponto de operação que resulta no valor máximo de F (limiar $\approx$ 0.2427), o classificador apresenta uma medida F de 60.4% (+/-3.9), correspondendo a uma cobertura de 78.1% (+/-5.1) e uma precisão de 49.5% (+/- 4.2). Utilizando um teste-t (Kachigan, 1986) com nível de significância de 1%, mostra-se que este resultado é superior ao relatado por Darnell et al. (2007).

Para ilustrar a aplicação do preditor, dois exemplos não incluídos no conjunto de dados de Darnell et al. (2007) são apresentados. O primeiro exemplo é o domínio de tetramerização da proteína repressora de tumor p53. Essa proteína de 393 aminoácidos atua como um fator de transcrição, de-

sempenhando um papel chave na proteção de organismos contra o câncer (el Deiry et al., 1992). O domínio de tetramerização da p53 está localizado na porção terminal COOH e compreende os aminoácidos 325 a 356.

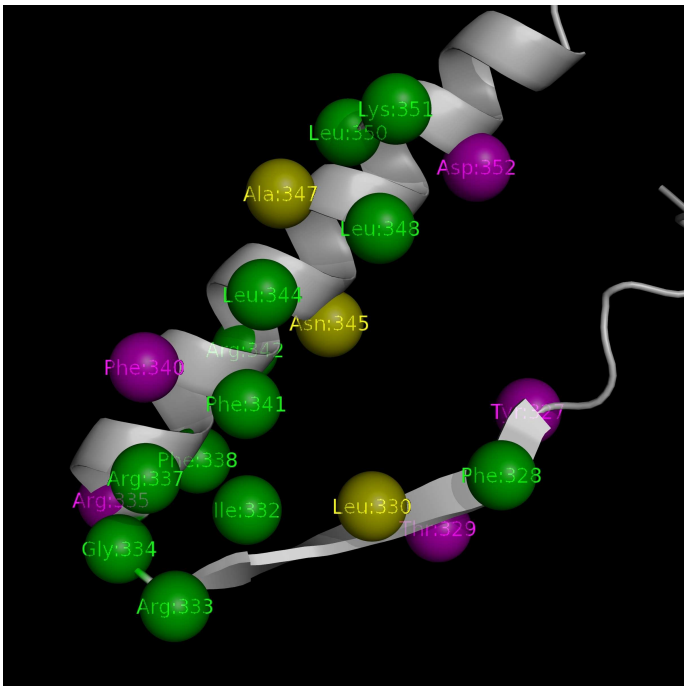

Figura 5.2: Um monômero do domínio de tetramerização da proteína repressora de tumor p53 (3sak:A). Positivos verdadeiros são indicados em verde, falsos positivos são indicados em roxo e falsos negativos são indicados em amarelo. Adaptado de (Higa e Tozzi, 2009).

Em um extensivo estudo de mutagênese sítio-dirigido, Kato et al. (2003) construíram 2.314 mutantes representando todas as possíveis substituições de aminoácidos causadas por uma mutação pontual. Ao avaliar a atividade dos mutantes, eles observaram um conjunto de 15 aminoácidos do domínio de tetramerização sensíveis à inativação do aminoácido por substituição: Phe:328, Leu:330, Ile:332, Arg:333, Gly:334, Arg:337, Phe:338, Phe:341, Arg:342, Leu:344, Asn:345, Ala:347, Leu:348, Leu:350 e Lys:351. Considerando os 32 aminoácidos do domínio de tetramerização, o método proposto identificou 12 dos 15 aminoácidos sensíveis à inativação (indicados em verde na Figura 5.2), bem como 5 falsos positivos (indicados em roxo), 3 falsos negativos (indicados em amarelo) e 12 negativos verdadeiros. Este resultado representa uma medida F de 75%, correspondendo a uma cobertura de 80% e uma precisão de 70.6%.

O segundo exemplo é a proteína denominada *bone morphogenetic protein-2* (BMP-2), um membro da família de proteínas *transforming growth factor-*β (TGF-β) com importante papel na formação e regeneração de ossos em vertebrados adultos (Reddi, 1998). Ela sinaliza ao se ligar a dois tipos diferentes de serina/treonina receptores kinase, classificados como tipo I e tipo II. Kirsch et al. (2000)

analisaram as interações de mutantes da proteína BMP-2 com receptores dos tipos I e II e observaram duas regiões de interface diferentes, cada uma correspondendo a um tipo diferente de receptor. Uma das regiões, a mais forte, compreende aminoácidos de ambos os monômeros (Val:26, Asp:30, Trp:31, Lys:101, Tyr:103 de um monômero e Ile:62, Leu:66, Asn:68, Ser:69, Phe:49, Pro:50, Ala:52, His:54 do outro), enquanto a outra inclui apenas aminoácidos de um monômero (Ala:34, His:39, Ser:88, Leu:90 e Leu:100).

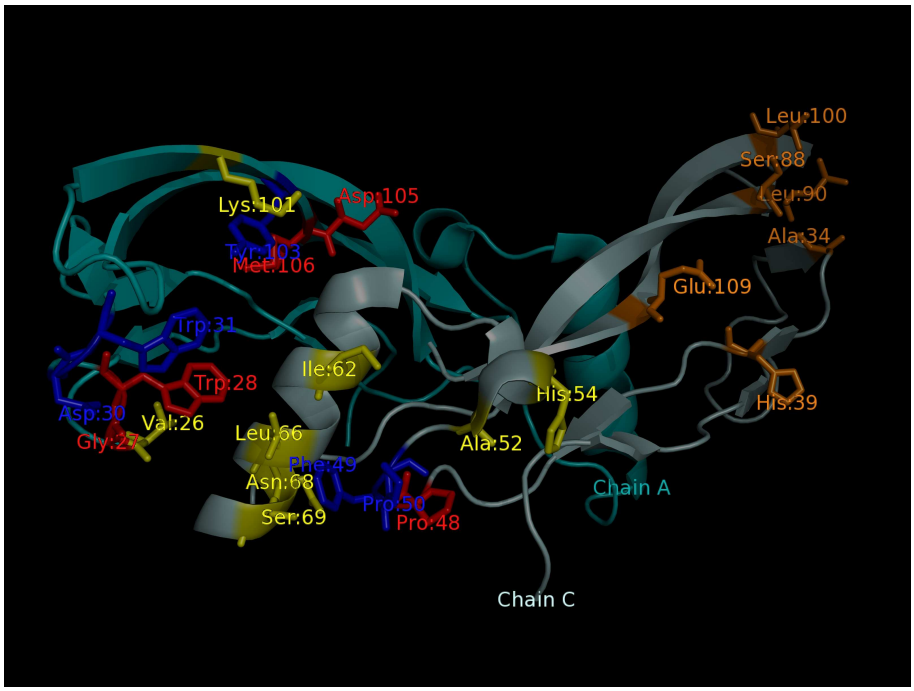

Figura 5.3: Homodímero correspondente à molécula da proteína *bone morphogenetic protein-2* (BMP-2) (1es7:A e C). Para a região de interface mais forte (à esquerda), os positivos verdadeiros são indicados em azul, os falsos positivos são indicados em vermelho e os falsos negativos são indicados em amarelo. Aminoácidos na região de interface mais fraca (à direita) são indicados em laranja. Adaptado de (Higa e Tozzi, 2009).

Neste exemplo, o preditor é utilizado para avaliar todos os aminoácidos da superfície de um monômero. A análise resultou em 29 aminoácidos preditos como *hot spots* dentre os 101 aminoácidos na superfície da proteína. Em seguida, as predições positivas foram filtradas da seguinte forma: uma predição positiva é mantida somente se a sua vizinhança (dois aminoácidos à esquerda e dois aminoácidos à direita) contém ao menos dois aminoácidos, cujas predições também são positivas. Este é um tipo de pós-processamento comumente utilizado por preditores de regiões de interface (Yan et al., 2004; Reš et al., 2005). Ao final, restaram 13 predições positivas, ilustradas na Figura 5.3. Dessas, 5 correspondem a aminoácidos da região de interface (positivos verdadeiros) e estão indicados em azul; 14 são falsos negativos e estão indicados em amarelo; 5 são falsos positivos e estão indicados

em vermelho; e 83 são negativos verdadeiros. Este resultado implica em uma medida F de 34.5%, correspondendo a uma precisão de 50% e uma cobertura de 26.3%. Se apenas a região de interface mais forte é considerada, obtém-se uma medida F de 43.5%, correspondendo a uma precisão de 50% e uma cobertura de 38.5%. Embora os níveis de cobertura obtidos sejam relativamente baixos, eles representam um resultado típico de preditores de regiões de interface (vide capítulo 4) e a um nível de precisão de 50% ele é considerado satisfatório para localizar a região de interface (Bradford e Westhead, 2005).

5.6 Discussão

Na seção 5.5, o desempenho do preditor de *hot spots* desenvolvido foi estimado e comparado com o relatado por Darnell et al. (2007), que utiliza o mesmo conjunto de dados e critério de avaliação. Verificou-se que o preditor desenvolvido foi o que apresentou o melhor desempenho.

Contudo, a diferença mais importante entre o preditor desenvolvido e os outros preditores de *hot spots* apresentados na literatura (vide seção 5.1), incluindo o relatado por Darnell et al. (2007), é o fato de que o preditor desenvolvido não pressupõe que a estrutura do complexo formado pela proteína de interesse e sua parceira de interação seja conhecido. Para calcular as propriedades físico-químicas e estruturais consideradas na formação do vetor de carcterísticas que descreve um aminoácido de interface, é utilizado apenas o conhecimento sobre a estrutura da proteína a que o aminoácido pertence. Isto abre a possibilidade de utilizar o preditor de *hot spots*, apresentado neste capítulo, em conjunto com o preditor de regiões de interface, apresentado no capítulo 4, uma característica única deste preditor.

O resultado da utilização conjunta desses dois preditores é a capacidade de fornecer uma informação mais detalhada que aquela fornecida apenas pela utilização dos preditores de região de interface proteína-proteína descritos na literatura.

5.7 Considerações finais

Neste capítulo foi apresentado um preditor de *hot spots* dentre os aminoácidos da região de interface consistindo de um classificador SVM com saída calibrada para representar a probabilidade *a posteriori*, de acordo com o método proposto por Platt (2000) (vide capítulo 3). Considerando o ponto de operação na curva ROC correspondente ao valor máximo da medida F, o classificador desenvolvido apresenta uma medida F de 60.4%, correspondendo a uma cobertura de 78.1% e uma precisão de 49.5%. Esses resultados indicam que o desempenho do preditor é superior ao do preditor apresentado por Darnell et al. (2007), que utiliza o mesmo conjunto de dados e critério de avaliação

de desempenho.

Além disso, o preditor proposto, diferente de outros preditores de *hot spots* descritos na literatura, utiliza um vetor de características formado por propriedades físico-químicas e estruturais, cujo cálculo não depende do conhecimento da estrutura do complexo formado pela proteína de interesse e sua parceira de interação. Isto, então, permite a sua utilização em conjunto com o preditor de regiões de interface desenvolvido no capítulo 4. Assim, os aminoácidos, identificados como pertencentes à região de interface, podem ser qualificados, de acordo com sua importância energética, provendo uma informação mais detalhada que a, usualmente, provida pelos preditores de regiões de interface descritos na literatura.

Capítulo 6

Conclusão e trabalhos futuros

6.1 Principais contribuições

Neste trabalho, foi desenvolvido um preditor de regiões de interface proteína-proteína, que utiliza um vetor de características construído a partir de medidas de propriedades físico-químicas e estruturais. Comparado com os preditores, descritos na literatura, que utilizam o mesmo tipo de medida, o preditor desenvolvido apresenta como principal contribuição a utilização do aminoácido como unidade básica de classificação em contraposição à utilização do conceito de *patches* (região contínua de forma aproximadamente circular, formada por um conjunto de aminoácidos amostrados sobre a superfície da proteína). Considerando o mesmo conjunto de dados e critério de avaliação, seu desempenho foi superior ao relatado por Bradford e Westhead (2005) e similar ao relatado pelos mesmos autores em um trabalho posterior (Bradford et al., 2006).

Adicionalmente, foi desenvolvido um preditor de *hot spots* dentre os aminoácidos da região de interface para identificação daqueles que apresentam maior contribuição energética para o processo de interação proteína-proteína (*hot spots*). Diferente de outros preditores de *hot spots* (Kortemme e Baker, 2002; Moreira et al., 2006; Ban et al., 2006; Guney et al., 2007; Darnell et al., 2007), o preditor desenvolvido não depende do conhecimento da estrutura do complexo formado pela proteína ao interagir com sua parceira, possibilitando sua utilização de forma complementar ao preditor de regiões de interface. O preditor de *hot spots* desenvolvido foi comparado com o relatado por Darnell et al. (2007) e, para o mesmo conjunto de dados e critério de avaliação, apresentou desempenho superior.

Além disso, a informação que resulta da utilização conjunta dos dois classificadores desenvolvidos é mais detalhada que aquela fornecida apenas pela utilização dos preditores de região de interface proteína-proteína descritos na literatura.

6.2 Trabalhos futuros

Diferentes linhas de investigação podem ser abordadas em estudos futuros para aprimoramento do trabalho desenvolvido, dentre elas destacam-se:

- A extensão do vetor de características com medidas extraídas do perfil evolucionário da proteína e avaliação de sua contribuição para a capacidade de discriminação do preditor de regiões de interface;
- A utilização de um modelo de classificação contextual bem fundamentado, como, por exemplo, os campos aleatórios markovianos (MRF), para a predição dos aminoácidos da interface com dependência de contexto. Isto permitiria a eliminação do procedimento empírico utilizado no segundo estágio do preditor de regiões de interface desenvolvido;
- A comparação do preditor de regiões de interface desenvolvido com um número maior de preditores descritos na literatura;
- A predição direta de *hot spots* a partir de aminoácidos da superfície da proteína. O conjunto de dados mais abrangente que permite o desenvolvimento deste preditor é formado pela união do conjunto de dados compilado por Darnell et al. (2007) com o compilado por Bradford e Westhead (2005), removendo-se eventuais redundâncias. Neste conjunto de dados, apenas uma parte dos dados é rotulada, implicando na necessidade de utilização de técnicas de classificação semi-supervisionadas (Chapelle et al., 2006) para o desenvolvimento do preditor; e
- Avaliação da capacidade de detecção de *hot pots* do preditor desenvolvido em relação à utilização de modelos teóricos para a estimação da energia de ligação proteína-proteína (Kortemme e Baker, 2002; Moreira et al., 2006).

Finalmente, deve-se observar que novas medidas físico-químicas e estruturais podem vir a ser propostas como resultado do contínuo estudo das características estruturais das proteínas. Desta forma, é recomendada a avaliação do potencial de novas medidas para melhoria da discriminação entre os aminoácidos da região de interface e os demais aminoácidos da superfície da proteína e, em caso positivo, sua inclusão no vetor de características utilizado pelos preditores desenvolvidos.

Referências Bibliográficas

Ahmed, S., Gromiha, M., Fawareh, H. e Sarai, A. (2004). Asaview: Database and tool for solvent accessibility representation in proteins, *BMC Bioinformatics* **5**: 51.

Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K. e Walter, P. (1998). *Essential Cell Biology - An Introduction to the Molecular Biology of the Cel*, Garland Publishing Inc, New York, NY, USA.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. e Watson, J. D. (1994). *Molecular Biology of the Cell*, Garland Publishing, New York, NY, USA. Third Edition.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. e Lipman, D. J. (1997). Gapped blast and psi-blast: A new generation of protein database search programs, *Nucleic Acids Research* **25**: 3389–3402.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New Jersey, USA. Third Edition.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. e Yeh, L.-S. L. (2004). Uniprot: the universal protein knowledgebase, *Nucleic Acids Research* **32**: D115–D119.

Bahandur, R. P., Chakrabarti, P., Rodier, F. e Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins, *Proteins: Structure, Function and Genetics* **53**: 708–719.

Baker, N. A. e McCammon, J. A. (2003). Eletrostatic interactions, *in* P. E. Bourne e H. Weissig (eds), *Structural Bioinformatics*, Wiley-Liss Inc, chapter 21, pp. 427–440.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J. e McCammon, J. A. (2001). Eletrostatics of nanosystems: Application to microtubes and ribossome, *Proceedings of the National Academy of Science USA* **98**: 10037–10041.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. e Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics* **16**(5): 412–424.

Ban, Y. A., Edelsbrunner, H. e Rudolph, J. (2006). Interface surfaces for protein-protein complexes, *Journal of the ACM* **5**(3): 361–378.

Bazaraa, M. S., Sherali, H. D. e Shetty, C. M. (1993). *Nonlinear Programming - Theory and Algorithms*, John Wiley and Sons, New York, NY, USA.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shyndyalov, I. e Bourne, P. E. (2000). The protein data bank, *Nucleic Acid Research* **28**(1): 235–242.

Besl, P. J. e Jain, R. C. (1988). Segmentation through variable-order surface fitting, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(2): 167–192.

Bishop, C. M. (1997). *Neural Networks for Pattern Recognition*, Oxford University Press, New York, NY, USA.

Blum, A. L. e Langley, P. (1997). Selection of relevant features and examples in machine learning, *Artificial Intelligence* **97**: 245–271.

Bogan, A. A. e Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces, *Journal of Molecular Biology* **280**: 1–9.

Bordner, A. e Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces, *Proteins: Structure, Function and Bioinformatics* **60**: 353–366.

Bradford, J. R., Needham, C. J., Bulpitt, A. J. e Westhead, D. R. (2006). Insights into protein-protein interfaces using a bayesian network prediction method, *Journal of Molecular Biology* **362**: 365–386.

Bradford, J. R. e Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using support vector machines approach, *Bioinformatics* **21**(8): 1487–1494.

Branden, C. e Tooze, J. (1999). *Introduction to Protein Structure*, Garland Publishing, New York, NY, USA. Second Edition.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

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

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* **2**: 121–167.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. e Huang, E. S. (2004). Are protein-protein interface more conserved in sequence than the rest of protein surface?, *Protein Science* **13**: 190–202.

Chakrabarti, P. e Janin, J. (2002). Dissecting protein-protein recognition sites, *Proteins: Structure, Function and Genetics* **47**: 334–343.

Chakravarty, S. e Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability, *Structure* **7**: 723–732.

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

Chang, C. C. e Lin, C. J. (2001). *LibSVM: a Library for Support Vector Machines*, Department of Computer Science, National Taiwan University. Acessado em 30/7/2008, em `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Chapelle, O., Schölkopf, B. e Zien, A. (2006). *Semi-Supervised Learning*, MIT Press, Cambridge, MA, USA.

Chen, H. e Zhou, H.-X. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against nmr data, *Proteins: Structure, Function and Bioinformatics* **61**: 21–35.

Chothia, C. e Janin, J. (1975). Principles of protein-protein recognition, *Nature* **275**: 705–708.

Chung, J.-L., Wang, W. e Bourne, P. E. (2006). Exploiting sequence and structure homologs to identify protein-protein binding sites, *Proteins: Structure, Function and Bioinformatics* **62**: 630–640.

Clackson, T. e Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface, *Science* **267**: 383–386.

Connolly, B. L. (1983). Analytical molecular surface calculation, *Journal of Applied Crystallography* **16**: 548–558.

Cristianini, N. e Shawe-Taylor, J. (2000). *Support Vector Machines - And other kernel-based learning methods*, Cambridge University Press, Cambridge, MA, USA.

Darnell, S. J., Page, D. e Mitchell, J. C. (2007). An automated decision-tree approach to predicting protein interaction hot spots, *Proteins: Structure, Function and Bioinformatics* **68**: 813–823.

De, S., Krishnadev, O., Srinivasan, N. e Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different, *BMC Structural Biology* **5**(15): 1–16.

DeLano, W., Ultsch, M. H., de Vos, A. M. e Wells, J. A. (2000). Convergent solutions to binding at a protein-protein interface, *Sience* **287**: 1279–1283.

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. e Baker, N. A. (2004). Pdb2pqr: An automated pipeline for the setup, execution, and analysis of poisson-boltzmann electrostatics calculations, *Nucleic Acids Research* **32**: W665–W667.

Duda, R. O., Hart, P. E. e Stork, D. G. (2001). *Pattern Classification*, John Wiley & Sons, Inc, New York, NY, USA. Second Edition.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D. e Tax, D. M. J. (2004). *PRTools4, a Matlab toolbox for pattern recognition*, Delft University of Technology, Delft. Acessado em 30/7/2008, em `http://www.prtools.org/`.

Durbin, R., Eddy, S. R., Krogh, A. e Mitchison, G. (1998). *Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.

Eisenberg, D. e McLachlan, A. D. (1986). Solvation energy inprotein folding and binding, *Nature* **319**(16): 199–203.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. e Scharf, M. (1995). The double cubic lattice method: efficient approaches to numerical integration of surface ara and volume and to dot surface contouring of molecular assemblies, *Journal of Computational Chemistry* **16**(3): 273–284.

el Deiry, W. S., Kern, S. E., Pientanpol, J. A., Kinzler, K. W. e Vogelstein, B. (1992). Definition of a consensus binding site for p53, *Nature Genetics* **1**: 45:49.

Everitt, B. S., Landau, S. e Leese, M. (2001). *Cluster Analysis*, Oxford University Press, New York, NY, USA. Fourth Edition.

Fariselli, P., Pazos, F., Valencia, A. e Casadio, R. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks, *European Journal of Biochemistry* **269**: 1356–1361.

Fauchère, J. e Pliska, V. (1983). Hydrophobic parameters p of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides, *European Journal of Medicinal Chemistry* **18**(4): 369–475.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters* **27**: 861–874.

Fernández-Recio, J., Totrov, M. e Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes, *Journal of Molecular Biology* **335**: 843–865.

Freund, Y. e Schapire, R. E. (1999). A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence,* **14**(5): 771–780.

Frishman, D. e Argos, P. (1995). Knowledge-based protein secondary structure assignment, *Proteins: Structure, Function and Genetics* **23**: 566–579.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press, New York, NY, USA. Second Edition.

Garey, M. P. e Johnson, D. S. (1979). *Computers and intractability - A guide to the theory of NP-completeness*, W. H. Freeman and Company, New York, NY, USA.

Gilson, M. K. (2002). Introduction to continuum eletrostatics with molecular applications, *in* D. A. Beard (ed.), *Computational and Theorical Biophysics*, Biophysiscs textbooks online, Biophysics Society. Acessado em 30/7/2008, em `http://www.biophysics.org/education/gilson.pdf`.

Greer, J. e Bush, B. L. (1978). Macromolecular shape ans surface maps by solvent exclusion, *Proceedings of the National Academy of Science, USA* **75**: 303–307.

Guney, E., Tuncbag, N., Keskin, O. e Gursoy, A. (2007). Hotsprint: Database of computational hot spots in protein interfaces, *Nucleic Acids Research* **36(Database issue)**: D662–666.

Guyon, I. e Elisseff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**: 1157–1182.

Hagerty, C. G., Muchnik, I. e Kulikowski, C. (1999). Two indexes can approximate four hundred and two amino acid properties, *Proceedings of the 1999 IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics*, Cambridge, MA, pp. 365–369.

Halperin, I., Ma, B., Wolfson, H. e Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions, *Proteins: Structure, Function and Genetics* **47**: 409–443.

Hamelryck, T. (2005). An amino acid has two sides: a new 2d measure provides a different view of solvent exposure, *Proteins* **59**: 38–48.

Hamelryck, T. e Manderick, B. (2003). Pdb file parser and structure implemented in python, *Bioinformatics* **19**: 2308–2310.

Hanley, J. A. e McNeil, B. J. (1982). The meaning and use of the area under a roc operating characteristic (roc) curve, *Radiology* **143**: 29–36.

Harold, E. R. e Means, W. S. (2001). *XML in a nutshell: a desktop quick reference*, O'Reilly, Sebastopol, CA, USA.

Haykin, S. (1999). *Neural Networks - A Comprehensive Foundation*, Prentice Hall, New Jersey, USA. Second Edition.

Higa, R. H. e Tozzi, C. L. (2008a). Prediction of protein-protein binding hot spots: A combination of classifiers approach, *Brazilian Symposium on Bioinformatics 2008 (Lecture Notes in Computer Science 5167)*, pp. 165–168.

Higa, R. H. e Tozzi, C. L. (2008b). A simple and efficient method for predicting protein-protein interaction sites, *Genetics and Molecular Research* **7**(3): 898–909. Versão expandida de trabalho apresentado na forma de resumo apresentado no 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology, X-Meeting 2007.

Higa, R. H. e Tozzi, C. L. (2009). Prediction of binding hot spot residues by using structural and evolutionary parameters, *Genetics and Molecular Biology* **32**(3): 626–633. Versão expandida de trabalho apresentado em forma de resumo no 4th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, X-Meeting 2008.

Hu, Z., Ma, B., Wolfson, H. e Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces, *Proteins: Structure, Function and Genetics* **39**: 331–342.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. e Hinton, J. E. (1991). Adaptative mixture of local experts, *Neural Computation* **3**: 79–87.

Jain, A. K., Duin, R. P. W. e Mao, J. (2000). Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1): 4–37.

Jain, A. K., Murty, M. N. e Flynn, P. J. (1999). Data clustering: A review, *ACM Computing Surveys* **31**(3): 264– 323.

Jain, A. e Zongker, D. (1997). Feature selection: Evaluation, application and small sample performace, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(2): 153–158.

Jolliffe, I. T. (2002). *Principal Component Analysis*, Springer Verlag, New York, NY, USA. Second Edition.

Jones, S. e Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches, *Journal of Molecular Biology* **272**: 121–132.

Jones, S. e Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis, *Journal of Molecular Biology* **272**: 133–143.

Kachigan, S. K. (1986). *Statistical Analysis - An Interdisciplinary Introduction to Univariate and & Multivariate Methods*, Radius Press, New York, NY, USA.

Kato, S., Han, S.-Y., Liu, W., Otsuka, K., Shibata, H. e Kanamaru, R. (2003). Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis, *Proceedings of the National Academy of Science USA* **100**(14): 8424–8429.

Kidera, A., Konishi, Y., Ooi, T. e Sheraga, H. A. (1985). Relation between sequence similarity and structural similarity in proteins, *Journal of Protein Chemistry* **4**: 265–297.

Kirsch, T., Nickel, J. e Sebald, W. (2000). Bmp-2 antagonists emerge from alterations in the low-affinity binding epitope for receptor bmpr-ii, *The EMBO Journal* **19**(13): 3314–3324.

Koenderink, J. J. (1990). *Solid Shape*, MIT Press, Cambridge, MA, USA.

Kohavi, R. e John, G. H. (1997). Wrappers for feature subset selection, *Artificial Intelligence* **97**: 272–234.

Koike, A. e Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines, *Protein Engineering, Design and Selection* **17**(2): 165–173.

Kortemme, T. e Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes, *Proceedings of the National Academy of Science USA* **99**(2): 14116–14121.

Kyte, J. e Doollittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology* **55**(3): 105–132.

Lee, B. e Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility, *Journal of Molecular Biology* **55**: 379–400.

Li, M.-H., Lin, L., Wang, X.-L. e Liu, T. (2007). Protein-protein interaction site prediction based on conditional random fields, *Bioinformatics* **23**(5): 597–604.

Li, X., Keskin, O., Ma, B., Nussinov, R. e Liang, J. (2004). Protein-protein interactions: Hot spots and structurally conserved residues often located in complemented pockets that pre-organized in the unbound states: Implications for docking, *Journal of Molecular Biology* **344**: 781–795.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. e Subramaniam, S. (1998). Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape, *Proteins: Structure, Function and Genetics* **33**: 1–17.

Lim, D., Park, H. U., de Castro, L., Kang, S. G., Lee, H. S. e Jensen, S. (2001). Cristal structure and kinetic analysis of betalactamase inhibitor protein-ii in complex with tem-1 beta-lactamase, *Nature Structural Biology* **8**: 848–852.

Lima, C. A. M. (2004). *Comitê de máquinas: Uma abordagem unificada empregando máquinas de vetores-suporte*, Tese de doutorado, Universidade Estadual de Campinas.

Lo Conte, L., Chothia, C. e Janin, J. (1999). The atomic structure of protein-protein recognition sites, *Journal of Molecular Biology* **285**: 2177–2198.

Lutz, M. (1996). *Programming Python*, O' Reilly, Sebastopol, CA, USA.

Ma, B., Elkayam, T., Wolfson, H. e Nussinov, R. (2003). Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proceedings of the National Academy of Science USA* **100**(10): 5772–5777.

Mardia, K. V., Kent, J. T. e Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, New York, NY, USA.

Mathworks (2004). *Matlab 7*, Mathworks Inc.

Mayrose, I., Mitchell, A. e Pupko, T. (2005). Site-specific evolutionary rate inference: taking uncertainty into account, *Journal of Molecular Evolution* **60**(3): 345–353.

McIvor, A. M. e Valkenburg, R. J. (1997). A comparison of local surface geometry estimation methods, *Machine Vision and Applications* **10**: 17–26.

Meyer, M., Desbrun, M., Schöder, P. e Barr, A. H. (2002). Discrete differential-geometry operators for triangulated 2-manifolds, *in* H. C. Hege e K. Polthier (eds), *Visualisation and Mathematics III*, Springer Verlag, pp. 35–57.

Moreira, I. S., Fernandes, P. A. e Ramos, M. J. (2006). Computational alanine scanning mutagenesis - an improved methodological approach, *Journal of Computational Chemistry* **28**: 644–654.

Moreira, I. S., Fernandes, P. A. e Ramos, M. J. (2007). Hot spots - an review of the protein-protein interface determinant amino-acid residues, *Proteins: Structure, Function and Bioinformatics* **68**: 803–812.

Nelson, D. L. e Cox, M. M. (2000). *Lehninger Principles of Biochemistry*, Worth Publishers, New York, NY, USA. Third Edition.

Neuvirth, H., Raz, R. e Schreiber, G. (2004). Promate: A structure based prediction program to identify the location of protein-protein binding sites, *Journal of Molecular Biology* **338**: 181–199.

Nguyen, C., Gardner, K. J. e Cios, K. J. (2007). A hidden markov model for predicting protein interfaces, *Journal of Bioinformatics and Computational Biology* **5**(3): 739–753.

Ofran, Y. e Rost, B. (2003a). Analysing six types of protein-protein interfaces, *Journal of Molecular Biology* **325**: 377–387.

Ofran, Y. e Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information, *FEBS Letters* **544**: 236–239.

Ofran, Y. e Rost, B. (2006). Isis: Interaction sites identified from sequence, *Bioinformatics* **23 ECCB 2006**: e13–e16.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelehoods methods, *in* B. S. D. S. A. Smola, P. Bartlett (ed.), *Advances in Large Margin Classifiers*, MIT Press.

Pressley, A. (2001). *Elementary Differential Geometry*, Springer Undergratuate Mathematics Series, Springer, London, UK.

Pudil, P., Novovicõvá, J. e Kittler, J. (1994). Floating search methods in feature selection, *Pattern Recognition Letters* **15**: 1119–1125.

Pupko, R., Bell, R. E., Mayrose, I., Glaser, F. e Ben-Tal, N. (2002). Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bionformatics* **18**(Supl. 1): S71–S77.

Radzika, A. e Wolfden, R. (1988). Comparing the polarities of amino acids: side chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution, *Biochemistry* **27**: 1664–1670.

Reddi, A. H. (1998). Role of morphogenetic proteins in skeletal tissue engineering and regeneration, *Nature Biotechnology* **16**: 247–252.

Reš, I., Mihalek, I. e Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures, *Bioinformatics* **21**(10): 2496–2501.

Rhodes, G. (1993). *Crystallography Made Crystal Clear - A Guide for Users of Macromolecular Models*, Academic Press, Inc, London, UK.

Richards, R. M. (1977). Areas, volumes, packing and protein structures, *Annual Review in Biophysics and Bioengineering* **6**: 151–176.

Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins - analytical equations for overlapping spheres and implications for the hydrophobic effects, *Journal of Molecular Biology* **178**: 63–89.

Rost, B. (1999). Twilight zones of protein sequence alignments, *Protein Engineering* **12**(2): 85–94.

Sander, C. e Schneider, R. (1991). Database of homology-derived of protein structure and structural meaning of sequence alignment, *Proteins: Structure, Function, and Genetics* **9**(1): 56–68.

Sanner, M. F., Olson, A. J. e Spehner, J. C. (1996). Reduced surface: An efficient way to compute molecular surface, *Biopolymers* **38**: 305–320.

Scheeff, E. D. e Fink, J. L. (2003). Fundamentals of protein structure, *in* P. E. Bourne e H. Weissig (eds), *Structural Bioinformatics*, Wiley-Liss Inc, chapter 2, pp. 15–39.

Schlkopf, B. e Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press.

Seber, G. A. F. (1984). *Multivariate Observations*, John Wiley and Sons, New York, NY, USA.

Sevcik, J., Urbanikova, L., Dauter, Z. e Wilson, K. S. (1998). Recognition of rnase sa by the inhibitor barstar: structure of the complex at 1.7 a resolution, *Acta Crystallography* **Sect D 54**: 954–963.

Shrake, A. e Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms - lysozyme and insulin, *Journal of Molecular Biology* **79**: 351–371.

Sutton, C. e McCallum, A. (2006). An introduction to conditional random fields for relational learning, *in* L. Getoor e B. Taskar (eds), *Introduction to Statistical Relational Learning*, MIT Press.

Tabernero, L., Bochar, D. A., Rodwell, V. W. e Stauffacher, C. V. (1999). Substrate-induced closure of the flap domain in the ternary complex structures provides insights into the mechanism of catalysis by 3-hydroxy-3-methylglutaryl-coa reductase, *Proceedings of the National Academy of Science USA* **96**: 7167–7171.

Thompson, J. D., Higgins, D. G. e Gibson, T. J. (1994). Clustal w: Improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research* **22**: 4673–4680.

Tsodikov, O. V., Jr., M. T. R. e Sergeev, Y. V. (2002). Novel computer program for fast exact calculation of accessbile and molecular areas and average surface curvature, *Journal of Computational Chemistry* **23**: 600–609.

Valdar, W. S. J. e Thornton, J. M. (2001). Protein-protein interfaces: Analysis of amino acid conservation in homodimers, *Proteins: Structure, Function and Genetics* **42**: 108–124.

van der Heijden, F., Duin, R. P. W., de Ridder, D. e Tax, D. M. J. (2004). *Classification, Parameter Estimation and State Estimation. and Engineering Approach Using Matlab*, John Wiley and Sons, Chichester, UK.

Vapnik, V. N. (1997). *Statistical Learning Theory*, John Wiley and Sons, New York, NY, USA.

Wang, B., Chen, P., Huang, D.-S., jing Li, J., Lok, T.-M. e Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate, *FEBS Letters* **580**: 380–384.

Webb, A. (2002). *Statistical Pattern Recognition*, John Wiley and Sons, Chichester, UK. Second Edition.

Wesson, L. e Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution, *Protein Science* (1): 227–235.

Westbrook, J. D. e Fitzgerald, P. M. D. (2003). The pdb format, mmcif, and other data formats, *in* P. E. Bourne e H. Weissig (eds), *Structural Bioinformatics*, Wiley-Liss Inc, chapter 8, pp. 161–180.

Wider, G. (2000). Structure determination of biological macromoleulces in solution using nmr spectroscopy, *BioTechniques* **29**: 1278–1294.

Yan, C., Dobbs, D. e Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues, *Bioinformatics* **20**(Suppl 1): i371–i378.

Zhou, H. X. e Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list, *Proteins: Structure, Function, and Genetics* pp. 336–343.

Apêndice A

Algoritmo de agrupamento *k-means*

O problema de agrupamento consiste em dividir um conjunto de padrões, representados por um vetor de características, em um conjunto de grupos, tal que padrões em um mesmo grupo sejam similares entre si e dissimilares quando comparados com padrões de outros grupos (Jain et al., 1999).

Apesar de existir uma diversidade enorme de métodos de agrupamento descritos na literatura (Everitt et al., 2001), nesta seção é apresentado apenas o algoritmo *k-Means* (Everitt et al., 2001), que é um dos métodos de agrupamento mais conhecidos e utilizados. Ele considera um número pré-fixado de grupos e particiona o conjunto de padrões de forma a minimizar o erro quadrático dos padrões com relação aos centróides de seus respectivos agrupamentos, $J = \sum_{i=1}^{k} \sum_{j=1}^{n_k} \|\mathbf{x}_{i,j} - \mathbf{c}_i\|$, onde k é o número de grupos considerado, n_k é o número de padrões no grupo k, $\mathbf{x}_{i,j}$ é o padrão j do grupo i e c_i é o centro do grupo i.

Algoritmo *k-means*

1. **Inicialização**: Seja n o número de padrões, k o número de grupos e c_1, c_2, ..., c_k os k centros de grupos, escolhidos de forma aleatória;

2. **Faça**:

   - Atribua cada padrão x_i, $i = 1, \ldots, n$, para o grupo l, tal que $l = arg\ \min_j \|\mathbf{x}_i - \mathbf{c}_j\|$, $1 \leq j \leq k$;
   - o Recalcule os k centros de grupos, c_1, c_2, ..., c_k, utilizando a corrente atribuição dos padrões aos k grupos;

3. **até que** o critério de convergência seja satisfeito (ex: nenhuma reatribuição de padrões para novos grupos).

O objetivo dos métodos de agrupamento, incluindo o método *k-means*, é descobrir a estrutura de classes embutida nos dados. A maioria dos métodos pode apresentar um comportamento diferente dependendo da natureza dos dados e da suposições iniciais consideradas, o número de classes no caso do método *k-means*. Dessa forma, os agrupamentos obtidos precisam ser avaliados com relação à sua validade. Medidas para avaliar a validade de agrupamentos geralmente são índices que refletem a magnitude relativa da similaridade intra e inter grupos, existindo uma diversidade de propostas na literatura (...) Contudo, esta seção restringe-se a apresentar a medida de avaliação da validade de agrupamentos denominada a *silhouette value*. A *silhouette value* para cada padrão é uma medida da similaridade do ponto com relação ao seu próprio grupo comparado com padrões em outros grupos, variando de -1 a 1. A *silhouette value*, $S(i)$, para o padrão i é definida como:

$$S(i) = \frac{\min_j(b(i,j),2) - a(i)}{\max_j(a(i), \min_j(b(i,j),2))}, \tag{A.1}$$

onde $a(i)$ é a distância média do padrão i para os outros padrões do seu grupo e $b(i,j)$ é a distância média do padrão i para padrões em outros grupos j.

O valor médio da *silhouette value* fornece uma medida da validade do agrupamento obtido e pode ser utilizado para se determinar o número ótimo de grupos para execução do algoritmo *k-means*. Quanto maior este valor, melhor o agrupamento obtido.

Apêndice B

Lista de aminoácidos

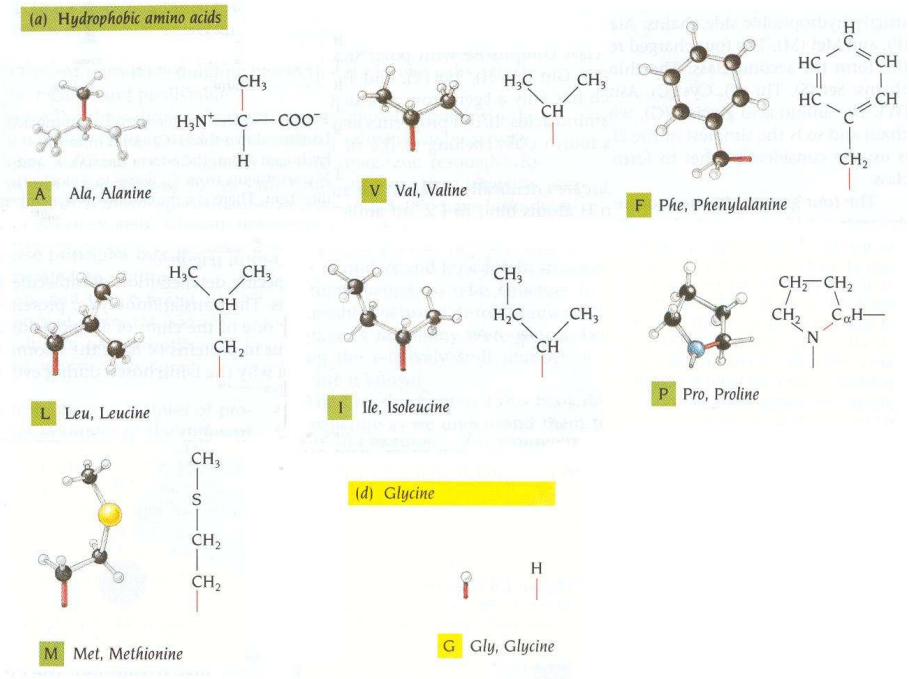

Figura B.1: Aminoácidos hidrofóbicos. Adaptado de Branden e Tooze (1999).

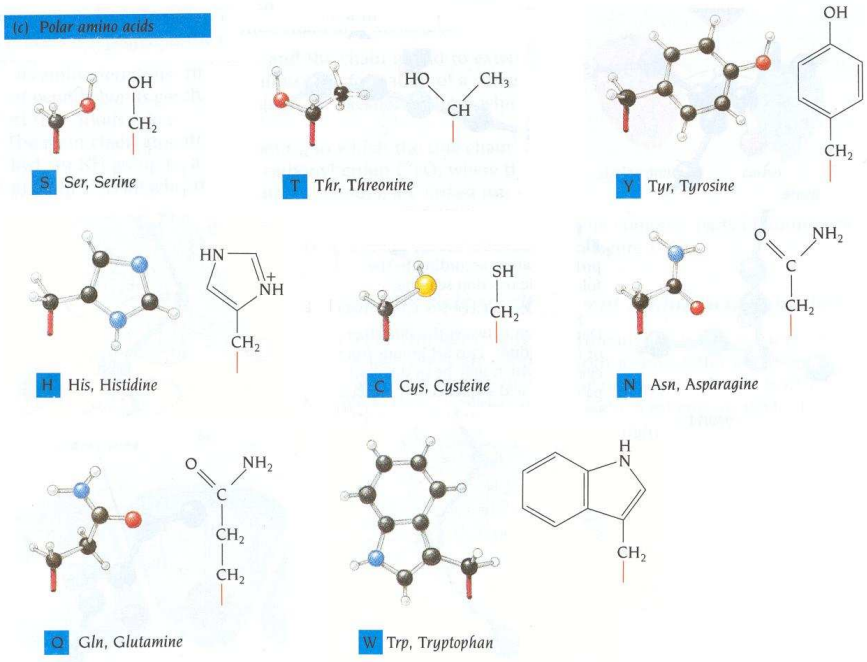

Figura B.2: Aminoácidos polares. Adaptado de Branden e Tooze (1999).

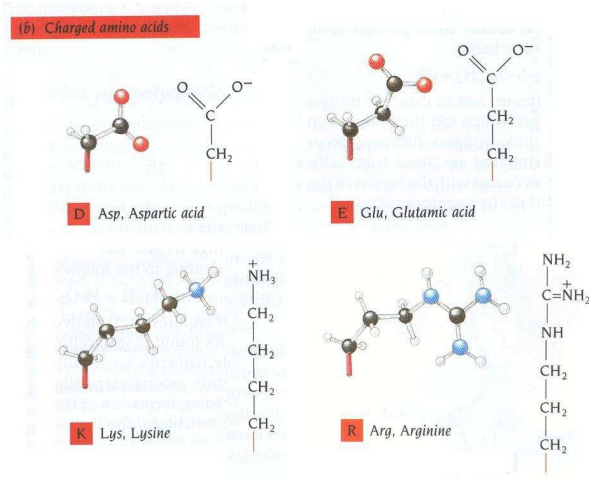

Figura B.3: Aminoácidos carregados. Adaptado de Branden e Tooze (1999).

Apêndice C

Conjuntos de dados

Tabela C.1: Estruturas extraídas a partir do conjunto compilado por Bradford e Westhead (2005) - Parte 1. As chaves {} indicam cadeias não incluídas no conjunto de dados e que foram utilizadas apenas para rotular a região de interface.

PDB ID	Cadeia 1	Cadeia 2	PDB ID	Cadeia 1	Cadeia 2
Enzima-Inibidores					
1a4y	A	B	1ava	A	{C}
1avw	A	B	1ay7	A	B
1bvn	T	{P}	1clv	I	{A}
1cse	I	{E}	1dpj	A	B
1dtd	A	B	1eai	C	{A}
1f34	B	{A}	1fss	A	B
1gla	F	{G}	1kxq	H	{A}
1mct	I	A	1smp	A	I
1tab	I	{E}	1tgs	I	{Z}
1udi	E	I	1viw	B	{A}
2ptc	I	{E}	2sic	E	I
4cpa	I	{_}	4sgb	E	I
7cei	A	B			
Transientes não enzima-inibidores					
1agr	E	{A}	1atn	D	{A}
1b6c	A	B	1bkd	S	{R}
1buh	B	{A}	1d2z	B	{A}
1dow	A	B	1eay	A	C
1euv	A	B	1f3v	B	{A}
1f5q	A	B	1he1	A	{C}
1hx1	B	A	1i2m	B	{A}
1i8l	C	A	1kac	B	{A}
1pdk	A	B	1qav	A	{B}
1tx4	A	B	1xdt	R	T
3ygs	C	P			
Permanentes heterogêneos					
1ahj	A	B	1aht	H	L
1b34	A	B	1bun	A	{B}
1dce	A	B	1dj7	A	{B}
1efv	A	B	1g4y	B	R
1gux	A	B	1h2a	L	S
1luc	B	{A}	1pnk	A	B
1req	A	B	1tco	A	B
2aai	A	B			

Tabela C.2: Estruturas extraídas a partir do conjunto compilado por Bradford e Westhead (2005) - Parte 2. As chaves {} indicam cadeias não incluídas no conjunto de dados e que foram utilizadas apenas para rotular a região de interface.

PDB ID	Cadeia 1	Cadeia 2	PDB ID	Cadeia 1	Cadeia 2
Permanentes homogêneos					
1a0f	B	{A}	1a4i	A	{B}
1a4u	A	{B}	1afr	F	{A}
1afw	A	{B}	1aj8	A	{B}
1ajs	A	{B}	1aom	A	{B}
1aq6	A	{B}	1at3	A	{B}
1az3	B	{A}	1b3a	B	{A}
1b5e	A	{B}	1b7b	A	{C}
1b8a	A	{B}	1b8j	A	{B}
1b9m	B	{A}	1bbh	A	{B}
1bft	A	{B}	1bjn	B	{A}
1bo1	A	{B}	1brm	A	{B}
1bw0	B	{A}	1byf	A	{B}
1byk	B	{A}	1c7n	A	{B}
1cli	B	{A}	1cmb	B	{A}
1cnz	A	{B}	1coz	A	{B}
1cp2	A	{B}	1dor	A	{B}
1e0b	A	{B}	1ete	A	{B}
1f5m	A	{B}	1f6y	A	{B}
1f8r	A	{C}	1gpe	B	{A}
1hgx	B	{A}	1hjr	C	{A}
1hss	A	{B}	1hul	A	{B}

Tabela C.3: Estruturas extraídas a partir do conjunto compilado por Bradford e Westhead (2005) - Parte 3. As chaves {} indicam cadeias não incluídas no conjunto de dados e que foram utilizadas apenas para rotular a região de interface.

PDB ID	Cadeia 1	Cadeia 2	PDB ID	Cadeia 1	Cadeia 2
Permanentes homogêneos (cont.)					
1isa	B	{A}	1jkm	A	{B}
1kpe	A	{B}	1mka	A	{B}
1msp	A	{B}	1nse	A	{B}
1nsy	A	{B}	1one	A	{B}
1pp2	L	{R}	1pvu	A	{B}
1qae	B	{A}	1qax	A	{B}
1qbi	A	{B}	1qfe	A	{B}
1qfh	B	{A}	1qi9	B	{A}
1qor	B	{A}	1qqj	A	{B}
1qu7	B	{A}	1scf	B	{A}
1smt	A	{B}	1sox	A	{B}
1spu	A	{B}	1trk	B	{A}
1vfr	B	{A}	1vhi	A	{B}
1vlt	A	{B}	1vok	A	{B}
1vsg	B	{A}	1wgj	A	{B}
1xik	A	{B}	1xso	A	{B}
1ypi	A	{B}	1yve	J	{I}
2ae2	A	{B}	2arc	A	{B}
2gsa	B	{A}	2hdh	A	{B}
2hhm	B	{A}	2nac	B	{A}
2pfl	B	{A}	2utg	B	{A}
3tmk	B	{A}	4mdh	A	{B}
5hvp	B	{A}			

Tabela C.4: Estruturas extraidas do conjunto compilado por Darnell et al. (2007). As chaves {} indicam cadeias não incluídas no conjunto de dados. Elas são utilizadas apenas para rotular a região de interface.

PDB ID	Cadeias 1	Cadeias 2
1a4y	A	B
1ahw	C	{AB}
1brs	A	D
1bsr	A	B
1bxi	A	{B}
1cbw	{BC}	D
1dan	{H}L	TU
1dvf	AB	CD
1dx5	M	{I}
1gc1	{G}	C
1nmb	HL	{N}
1vfb	AB	{C}
3hfm	HL	Y
3hhr	A	B{C}

Tabela C.5: Aminoácidos e respectivos ΔΔGs extraidos do conjunto de dados compilado por Darnell et al. (2007) - Parte 1.

AA	ΔΔG	AA	ΔΔG	AA	ΔΔG	AA	ΔΔG
1a4y:A							
W261	0.10	W263	1.20	S289	0.00	W318	1.50
K320	-0.30	E344	0.20	W375	1.00	E401	0.90
Y434	3.30	D435	3.50	Y437	0.80	I459	0.70
1a4y:B							
R5	2.30	H8	0.90	Q12	0.30	R31	0.20
R32	0.90	N68	0.20	H84	0.20	W89	0.20
E108	-0.30	H114	0.65				
1ahw:C							
Y156	4.00	T167	0.00	T170	1.00	V198	-0.30
1brs:A							
K27	5.40	N58	3.10	R59	5.20	E60	-0.20
R83	5.40	R87	5.50	H102	6.10		
1brs:D							
Y29	3.40	D35	4.50	D39	7.70	T42	1.80
E76	1.30						
1bsr:A							
C31	0.93	C32	0.75				
1bsr:B							
C31	0.93	C32	0.75				
1bxi:A							
C23	0.92	N24	0.14	T27	0.73	S28	0.17
S29	0.96	E30	1.41	L33	3.42	V34	2.58
V37	1.66	T38	0.90	E41	2.08	S48	0.01
S50	2.19	D51	5.92	I53	0.85	Y54	4.83
Y55	4.63						
1cbw:D							
T11	0.20	K15	2.00	R17	0.50	R39	0.20
1dan:L							
L39	0.00	K62	0.00	Q64	0.80	I69	1.90
F71	1.20	L73	0.00	E77	0.00	R79	1.20
Q88	0.00	V92	0.00	N93	0.00	E94	0.00
H115	0.00						
1dan:T							
T17	0.10	K20	2.60	I22	0.70	E24	0.70
Q37	0.55	K41	0.35	S42	-0.10	D44	0.70
W45	1.60	S47	0.05	K48	0.40	F50	0.40
D58	2.18	D61	0.24	F76	1.20		
1dan:U							
Y94	1.00	Q110	1.40	E128	0.10	R131	0.00
T132	0.00	L133	0.00	R135	0.55	F140	1.50
S163	0.00	T203	0.10	V207	-0.20	E208	0.00

Tabela C.6: Aminoácidos e respectivos ΔΔGs extraidos do conjunto de dados compilado por Darnell et al. (2007) - Parte 2.

AA	ΔΔG	AA	ΔΔG	AA	ΔΔG	AA	ΔΔG
1dvf:A							
H30	1.70	Y32	2.00	Y49	1.70	Y50	0.70
W92	0.30						
1dvf:B							
T30	0.90	Y32	1.80	W52	4.20	D54	4.30
N56	1.20	D58	1.60	E98	4.20	R99	1.90
D100	2.80	Y101	4.00				
1dvf:C							
Y49	1.90						
1dvf:D							
K30	1.00	H33	1.90	I97	2.70	Y98	4.70
Q100	1.60						
1dx5:M							
F34	2.60	Q38	1.40	R67	3.40	T74	0.80
R75	0.70	Y76	3.00	K81	1.00	I82	2.60
M84	0.30	K110	0.00				
1gc1:C							
Q25	0.03	H27	0.28	K29	0.59	N32	0.18
Q33	0.10	K35	0.32	Q40	-0.41	S42	0.00
L44	1.04	T45	-0.15	N52	0.70	R59	1.16
S60	-0.09	D63	-0.32	Q64	0.44		
1nmb:H							
D56	2.80	Y99	1.50	Y100A	0.50		
1nmb:L							
Y32	1.70	T93	0.30	L94	0.90		
1vfb:A							
Y32	1.30	Y49	0.80				
1vfb:B							
W52	1.23	D58	-0.20	E98	1.10	Y101	4.00
3hfm:H							
S31	0.20	D32	2.00	Y33	6.00	Y50	7.50
Y53	3.29	Y58	1.70				
3hfm:L							
N31	5.25	N32	5.20	Y50	4.60	Q53	1.00
Y96	2.80						

Tabela C.7: Aminoácidos e respectivos $\Delta\Delta$Gs extraidos do conjunto de dados compilado por Darnell et al. (2007) - Parte 3.

AA	$\Delta\Delta$G	AA	$\Delta\Delta$G	AA	$\Delta\Delta$G	AA	$\Delta\Delta$G
3hfm:Y							
H15	-0.44	Y20	5.00	R21	1.00	W63	0.31
R73	-0.20	L75	1.25	T89	0.00	N93	0.60
K96	7.00	K97	6.00	I98	0.00	S100	0.26
D101	1.50						
3hhr:A							
I4	0.41	R8	0.20	L9	-0.04	N12	0.10
L15	0.15	R16	0.24	H18	-0.50	H21	0.20
Q22	-0.2	F25	-0.40	Y42	0.20	L45	1.20
Q46	0.10	S62	0.20	N63	0.30	R64	1.60
Q68	0.60	Y164	0.30	R167	0.30	K168	-0.20
D171	0.80	K172	2.00	E174	-0.90	T175	2.00
R178	2.40	I179	0.80	C182	1.01		
3hhr:B							
R43	2.20	E44	1.80	W76	0.60	T77	-0.25
S102	-0.20	I103	1.80	W104	4.50	I105	2.00
C108	0.00	E120	-0.20	K121	0.10	C122	0.00
D126	1.00	E127	1.00	D164	1.60	I165	2.20
Q166	0.00	K167	0.00	W169	4.50	R217	0.20
N218	0.30						