**Universidade Federal de Alagoas**
**Mestrado em Modelagem Computacional**
**de Conhecimento**


Dissertação de Mestrado

# Predição da Função das Proteínas Sem Alinhamentos Usando Máquinas de Vetor de Suporte

Ulisses Martins Dias
ulimard@yahoo.com.br

Orientadores:

Roberta Vilhena Vieira Lopes
Eliana Silva de Almeida

Maceió, Março de 2007

Ulisses Martins Dias

# Predição da Função das Proteínas Sem Alinhamentos Usando Máquinas de Vetor de Suporte

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Modelagem Computacional de Conhecimento do Departamento de Tecnologia da Informação da Universidade Federal de Alagoas.

Orientadores:

Roberta Vilhena Vieira Lopes

Eliana Silva de Almeida

Maceió, Março de 2007

**Catalogação na fonte**
**Universidade Federal de Alagoas**
**Biblioteca Central**
**Divisão de Tratamento Técnico**
**Bibliotecária Responsável: Helena Cristina Pimentel do Vale**

D541p Dias, Ulisses Martins.
Predição da função das proteínas sem alinhamentos usando máquinas de vetor de suporte / Ulisses Martins Dias. – Maceió, 2007.
90 f. : il.

Orientadora: Roberta Vilhena Vieira Lopes.
Co-Orientadora: Eliana Silva de Almeida.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2007.

Bibliografia: f. 85-90.

1. Bioinformática. 2. Proteína. 3. Função. 4. Inteligência artificial. 5. Máquina de vetor suporte. 6. Gene Ontológico. I. Título.

CDU: 004.8

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional de Conhecimento pelo Programa Multidisciplinar de Pós-Graduação em Modelagem Computacional de Conhecimento, da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina:

Profa. Dra. Roberta Vilhena Vieira Lopes
UFAL – Instituto de Computação
Orientadora

Profa. Dra. Eliana Silva de Almeida
UFAL – Instituto de Computação
Co-orientadora

Prof. Dr. Luiz Marcos Garcia Gonçalves
UFRN – Departamento de Engenharia de Computação e Automação
Examinador

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Examinador

# Resumo

Este trabalho apresenta um novo modelo capaz de prever a função de proteínas utilizando máquinas de vetor de suporte, um método de aprendizagem de máquina treinado usando parâmetros estruturais calculados a partir da conformação espacial da própria proteína. O modelo difere do paradigma comum de predição por não ser necessário calcular similaridades por meio de alinhamentos entre a proteína que se deseja prever a função e as proteínas de função conhecida presentes nos bancos de dados públicos. Dessa forma, o modelo é capaz de associar função às proteínas que não possuem qualquer semelhança com proteínas conhecidas, podendo ser usado quando todos os outros métodos falham ou quando não se deseja utilizar o conceito de similaridade na predição da função. A justificativa de que o modelo é válido foi realizada analisando sua performance ao prever funções de proteínas desconhecidas, proteínas não usadas no treinamento, utilizando como estudo de caso um conjunto de proteínas de ligação.

# Abstract

This thesis presents a new model to protein function prediction using support vector machines, a machine learning approach trained using structural parameters calculated from protein tertiary structure. The model is different from the others paradigms because it is not necessary to search for similarities against the others known proteins in public databases by alignments. In this way, the model is able to associate functional relationships among proteins with no similarities and it could be used when all other methods fail or when the user don't want to use the concept of similarity in function predictions. The proof that the model is valid was accomplished analyzing its performance with unknown proteins, i.e proteins not used in the training set. The validation approach used a set of binding proteins.

# Agradecimentos

A Deus, por ter concedido inteligência suficiente para finalizar esse trabalho.

A todos os membros da banca. É uma grande honra poder contar com a contribuição de todos.

Aos professores Agamemnon e Roberta pela acolhida e pela oportunidade de conviver de tão perto com diversos debates intelectuais na beira da piscina.

Novamente à professora Roberta Lopes pela constante orientação, sempre indicando a direção a ser tomada nos momentos de maior dificuldade. Agradeço, principalmente, pela confiança, mais uma vez depositada, no meu trabalho de dissertação.

Aos meus pais pelo apoio incondicional e pelo exemplo de força e determinação.

A minha esposa Danielle, companheira de todas as horas, por tudo que ela tem feito por mim. Sem seu amor e cuidados sei que não teria sobrevivido.

Às minhas irmãs que, mesmo distantes fisicamente, sempre torceram pelo meu sucesso.

Aos meus amigos Ig e Rosemeire pelas boas influencias que de forma direta ou indireta contribuíram durante essa jornada acadêmica.

À FAPEAL pelo apoio financeiro.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

## 1.1 Predição da Função das Proteínas

A quantidade de métodos computacionais para a predição da função de proteínas aumentou consideravelmente durante os últimos cinco anos. A rapidez com que novos dados genômicos são produzidos diariamente e novas estruturas são obtidas permite entender melhor o relacionamento entre os dados estruturais e as funções das proteínas.

Entretanto, desvendar a estrutura de uma proteína não garante conhecer a função que ela desempenha na célula, já que proteínas com conformações espaciais semelhantes podem possuir funções diferentes se o sítio de ligação tiver divergido (Todd et al. 20002) e proteínas com as mesmas funções podem possuir conformações espaciais diferentes em vários aspectos (Kauvar & Vilar 1998, Lesk & Fordham 1996).

Além disso, para aumentar a complexidade do problema em questão, a palavra 'função' gera várias ambigüidades em biologia molecular, pois pode ser usada de várias maneiras em níveis diferentes de um organismo como, por exemplo, molecular, celular, tecidular e no organismo como um todo. Em geral, nenhum desses níveis ocorre isoladamente, funções de nível molecular trabalham juntas para produzir uma função de nível celular e assim por diante.

A complexidade do problema em questão aumenta quando é levado em conta que o conjunto de funções deverá ser formalizado para processamento automático via métodos computacionais. Logo, é necessário padronizar o vocabulário utilizado para descrever as características funcionais das proteínas.

Na literatura já existem duas maneiras consolidadas de descrever as possíveis funções desempenhadas pelas proteínas. A primeira delas foi criada por bioquímicos para caracterizar e padronizar a descrição das diversas re-

ações enzimáticas que ocorrem em nível celular criando uma comissão para classificação de enzimas - EC (Webb 1992, Tipton & Boyce 2000). A segunda maneira de se descrever as funções é utilizando a ontologia conhecida como *Gene Ontology* - GO (Ashburner et al. 2000). GO é um projeto que compila um vocabulário controlado e dinâmico dos termos relacionados a diferentes aspectos de genes e proteínas, sendo escrito em linguagem natural para os pesquisadores, mas com uma estrutura hierárquica que facilita o processamento automático.

## 1.2 Motivação

Após o término do seqüenciamento do genoma humano e de vários outros genomas um dos grandes desafios da bioinformática passou a ser decifrar a função de cada gene expresso e criar métodos capazes de fornecer uma descrição, mesmo não tão precisa, de novos genes a serem descobertos futuramente.

Grande parte dos métodos computacionais existentes atualmente são baseados no alinhamento de seqüências de genes ou proteínas, com o objetivo de inferir informações funcionais através de similaridades. Mais recentemente, o método foi extendido para o alinhamento de estruturas de proteínas.

Entretanto, os métodos baseados em similaridade falham quando, para uma dada proteína alvo, não é possível encontrar proteínas conhecidas similares seqüencialmente ou estruturalmente. Esse fato torna o método limitado às proteínas que são comuns na natureza. Nesse contexto, o presente trabalho se engloba no grupo de novos paradigmas confeccionados para suprir essa lacuna e efetuar predições mais precisas.

## 1.3 Contextualização

Os métodos computacionais existentes para predizer a função das proteínas podem ser divididos em três grupos: os que trabalham com a seqüência de aminoácidos, os que trabalham com a estrutura terciária, alguns especificamente dão maior atenção à superfície da proteína e os que trabalham com a seqüência e a estrutura ao mesmo tempo.

Os métodos baseados na seqüência de aminoácidos, também chamados de métodos por transferência de homologia, se propõem a obter a funcionalidade da proteína partindo-se apenas da seqüência de nucleotídeos. Nesse caso, a primeira e mais comum técnica de predição é efetuada a partir do alinhamento da proteína alvo com outras proteínas utilizando a ferramenta

PSI-BLAST (Altschul et al. 1990) a fim de encontrar proteínas com um alto grau de similaridade presentes no banco de dados e cuja função já foi definida experimentalmente.

De posse dessas proteínas, é possível aprender sobre alguns aspectos funcionais da proteína alvo apenas associando-a com as proteínas encontradas. A razão biológica para esse tipo de abordagem é que, se duas seqüências possuem alto grau de similaridade, então elas evoluíram de um ancestral comum e possuem funções similares, se não idênticas.

Essa afirmação é bastante limitada em sua validade, pois estudos comprovam que mesmo proteínas com seqüências muito semelhantes podem possuir funções completamente diferentes (Devos & Valencia 2000, Gerlt & Babbitt 2000), isso torna os métodos de predição baseados unicamente na seqüência pouco seguros.

Além disso, Friedberg (2006) identificou três aspectos que diminuem a eficácia desses métodos à medida que os bancos de dados de seqüência crescem. O primeiro aspecto é a observação de que mesmo com alto grau de similaridade a predição pode ser errônea.

Exemplificando com os processos enzimáticos das proteínas, Shah & Hunter (1997) foram os primeiros a concluir que, para utilizar apenas a seqüência, é necessário encontrar subregiões conservadas que sejam funcionalmente importantes como, por exemplo, regiões responsáveis pela formação de sítios de ligação. Recentemente, Rost (2002) concluiu que a função enzimática de uma proteína pode não ser conservada mesmo com alto grau de similaridade.

O segundo aspecto responsável pela diminuição da eficácia é o fato de que muitas das novas seqüências coletadas nos laboratórios são diferentes das seqüências baseadas nos bancos de dados, ou seja, é impossível encontrar uma outra seqüência que compartilhe uma similaridade capaz de sugerir determinada função. Isso faz com que esse método não possa ser utilizado em um grande número de casos. Uma possível solução para esse problema seria aumentar a sensibilidade da busca utilizando, ao invés de alinhamentos globais, um alinhamento com o perfil das famílias de proteínas.

O terceiro e último aspecto não é um problema particular dos métodos baseados em alinhamentos seqüenciais, mas também de qualquer método de anotação. Tal aspecto se refere a quantidade de proteínas anotadas com função errada, incompleta ou sem padronização. Quanto mais seqüências entram nos bancos de dados, mais são anotadas por transferência de homologia, o que faz com que os erros sejam propagados (Gilks et al. 2005, Friedberg 2006).

Os métodos baseados na estrutura da proteína permitem analisar o mecanismo bioquímico pelo qual as proteínas implementam a sua funcionalidade.

Eles buscam um alinhamento tridimensional da estrutura da proteína alvo com outras proteínas de função conhecida para associar informações sobre sua funcionalidade. Essa metodologia possui motivação biológica semelhante a dos métodos baseados em similaridades seqüenciais. Entretanto, essa abordagem é favorecida pelo fato de a estrutura espacial das proteínas ser evolutivamente bem mais preservada que a sua seqüência. Assim, muitas proteínas com seqüências bem diferentes podem possuir a mesma conformação tridimensional (Brenner et al. 1996) e compartilhar funções similares.

Entretanto, é possível encontrar semelhanças funcionais entre estruturas bem diferentes. Nesse caso, a utilização de alinhamento estrutural da proteína alvo com proteínas conhecidas presentes em um banco de dados é de pouca valia, necessitando de outros métodos capazes de obter informações da estrutura protéica.

Além disso, estruturas similares podem ter funções diferentes se, por meio de mutações durante o processo evolucionário, o sítio de ligação divergiu. Como conseqüência, métodos puramente baseados em comparações estruturais globais geram previsões erradas ou sem acurácia, já que são poucos os resíduos responsáveis pela especificidade da ligação (Seção 2.4.1). Outro fator que dificulta a utilização dessa metodologia é que nem sempre é fácil encontrar um alinhamento estrutural satisfatório. Isso faz com que seja impossível para muitas proteínas utilizar esse tipo de abordagem.

Apesar de todas as críticas a essa abordagem, muitos sistemas a utilizam em conjunto com outras abordagens para melhorar a acurácia do sistema (Pal & Eisenberg 2005, Yao et al. 2003, Bowie et al. 1991, Holm & Sander 1998).

Nos casos em que um bom alinhamento estrutural não foi encontrado com a proteína alvo, ainda é possível buscar informações funcionais analisando padrões estruturais conservados, com o objetivo de encontrar marcadores únicos associados à função. A razão biológica em se utilizar padrões estruturais é a mesma dos perfis de seqüência.

Infelizmente, esse método, assim como todos os outros, perdem acurácia na identificação de funções mais específicas para cada proteína. Resultados melhores vêm sendo obtidos com a utilização de uma combinaçãos de vários métodos diferentes, em que a relevância de cada método é obtida através de redes bayesianas (Pal & Eisenberg 2005), mas mesmo assim, a predição ainda está longe de poder substituir os métodos experimentais.

Os métodos baseados na superfície das proteínas consideram a importância dos sítios de ligação para que a ação biológica de uma proteína surtisse efeito. Como os sítios de ligação são formados por cavidades presentes na superfície das moléculas conclui-se que a função de uma proteína é influen-

ciada pelas propriedades físicas, químicas e geométricas da superfície (Norel et al. 1994)

Além disso, proteínas executam suas tarefas celulares interagindo com outras moléculas e os resíduos de interação estão dispersos em diversas regiões da cadeia primária, o que dificulta o trabalho de prever a função de uma proteína usando apenas a seqüência ou alinhamentos globais de estrutura terciária.

Assim, utilizar as propriedades da superfície da proteína alvo apresenta vantagens sobre os métodos que utilizam a conformação espacial como um todo. Já existem casos descritos na literatura que relatam proteínas com seqüência e estruturas não relacionadas possuindo a mesma distribuição espacial de resíduos funcionalmente importantes (Kauvar & Vilar 1998, Kobayashi & Go 1997).

Nesse contexto, alguns trabalhos de grande importância podem ser encontrados como, por exemplo, o trabalho de Schmitt et al. (2002) que utiliza um conjunto de descritores gerados a partir de uma série de cavidades pré-computadas para armazenar os padrões conservados nas superfícies. Tais descritores são gerados com a idéia de que a detecção de um padrão presente em duas cavidades corresponde ao problema de encontrar um subgrafo completo comum aos dois conjuntos de descritores (Schmitt et al. 2002). Nesse caso, a solução encontrada para o problema é a utilização de algorimtos de *clique-detection* (Bron & Kerbosch 1973).

Apesar de elegante, o algoritmo é computacionalmente oneroso e não pode ser usado em uma análise *all-against-all* (Ferre et al. 2005). Um outro trabalho que utiliza a superfície das proteínas é Binkowski et al. (2003), que alinha a superfície da proteína alvo com um representativo conjunto de padrões residuais que formam o arcabouço das regiões de cavidades. Entretanto, o método não consegue detectar padrões similares na superfície que estejam fora de ordem na cadeia primária.

Recentemente, Ferre et al. (2005) relatou um método que consegue identificar padrões conservados independente da ordem em que aparecem na seqüência de aminoácidos, que é uma das principais vantagens sobre métodos anteriores. O método, além disso, possui como característica o fato de ser bastante rápido.

Yao et al. (2003) apresentou um sistema capaz de obter uma boa performance utilizando uma constatação simples, mas normalmente ignorada, a de que todas as proteínas existentes são resultado de um processo de divergências evolucionárias. Dessa forma, criou-se a abordagem de rastreamento evolucionário (*evolutionary trace method*).

Nessa abordagem, a importância evolucionário dos resíduos em uma família de proteínas é classificada correlacionando suas variações com divergências evolucionárias. Dessa forma, esse método é capaz de identificar com maior precisão vários sítios funcionais que podem ser usados posteriormente em trabalhos de *docking* de proteínas (Aloy et al. 2001).

A técnica de Yao et al. (2003) se baseia no fato de que, em uma grande variedade de famílias de proteínas, os rastros evolucionariamente mais importantes formam agrupamentos (*clusters*) que são significantemente diferentes do que se resíduos fossem selecionados de modo aleatório (Madabushi et al. 2002).

A característica central da análise de rastreamento evolucionário é uma classificação hierárquica das características funcionais, aproximadas através de árvores evolucionárias. Como essas árvores são intrínsecas a qualquer família de proteínas, conclui-se que a abordagem pode ser aplicável a qualquer estrutura protéica com suficientes seqüências homólogas para criar o histórico natural de variações evolucionárias e seleções de *clusters*.

Recentemente, métodos híbridos que englobam alguns dos vários métodos citados acima vêm sendo utilizados. Nesse contexto, encontra-se o trabalho de Pal & Eisenberg (2005), que criou um sistema híbrido em que características extraídas da seqüência e da estrutura são usadas para melhorar a previsão funcional. Para isso, há um banco de dados que armazena as proteínas anotadas em conjunto com cada aspecto que pode ser utilizado. A partir desse banco é possível gerar uma predição analisando separadamente cada aspecto disponível. Redes bayesianas geram o relatório final de acordo com a estrutura hierárquica da *gene ontology*.

Entre os novos paradigmas, é possível citar ainda os trabalhos que utilizam técnicas de aprendizagem de máquina em algum contexto como, por exemplo, Eisner (2005), que utiliza no contexto de seqüência de proteínas discriminantes como palavras chaves extraídas do Swiss-Prot, um número fixo de características extraídas do PFAM (Bateman et al. 2000) e alinhamentos através da ferramenta Blast (Bedell et al. 2003). Informações detalhadas sobre tais discriminantes podem ser encontradas no trabalho de León & Markel (2003).

O tabalho de Eisner (2005) serviu de inspiração para o modelo aqui apresentado, inclusive o método de aprendizagem de máquina utilizado, máquinas de vetor de suporte, foi utilizado em Eisner (2005) e se mostraram bastante eficientes no contexto das informações obtidas com as seqüências das proteínas. Outra semelhança entre ambos os trabalhos pode ser vistas nos métodos de análise estatística que serão apresentados no capítulo 5, pois o presente trabalho utilizou o modelo validação estatística apresentado por Eisner (2005).

A fundamental diferença entre Eisner (2005) e o presente trabalho é o fato de aqui se utilizar apenas informações relacionadas a conformação espacial das proteínas. Nesse novo contexto surgem várias complicações, pois a estrutura das proteínas contém uma gama de informação impossível de ser utilizada completamente, o que obriga a uma série de restrições com o intúito de tornar o modelo viável computacionalmente e, ao mesmo tempo, eficaz.

## 1.4 Objetivos e Contribuições

Propor uma extensão do trabalho de Eisner (2005) para utilizar a estrutura da proteína ao invés da seqüência de aminoácidos utilizando a GO para classificar as funções preditas, tal que:

- Seja capaz de relacionar proteínas funcionalmente semelhantes independente da similaridade seqüencial ou estrutural.
- Represente a proteína, independente de seu tamanho, através de um conjunto fixo de parâmetros físicos, químicos e geométricos.
- Verifique se uma dada função molecular pode ser da alçada de uma proteína, permitindo utilizar em laboratório apenas os testes específicos que confirmem a predição e reduzindo o tempo e os custos que seriam necessários em testes laboratoriais mais abrangentes.
- Obtenha, para uma dada proteína, uma lista de funções moleculares possivelmente desempenhadas.

## 1.5 Descrição da Dissertação

Neste trabalho será apresentado o desenvolvimento de uma máquina de vetor de suporte para predição da função de uma proteína a partir de sua estrutura. Neste sistema, a função predita irá se restringir apenas aos aspectos da função molecular da proteína, que define a atividade bioquímica de uma proteína na célula.

O *kernel* utilizado pela máquina de vetor de suporte foi a função de base radial. O treinamento da máquina ocorre mediante uma busca exaustiva por quais parâmetros do *kernel* melhor se adaptam aos dados de entrada. Encontrados esse parâmetro, o cálculo do maior hiperplano de separação é um problema determinístico.

Os dados de entrada foram obtidos a partir da base de dados STING_DB associada às funções presentes no projeto de anotação da GO através do PDB_Id, que funcionou como chave para a ligação das bases de dados.

É importante ressaltar que, devido a limitação dos objetivos deste trabalho, ficaram fora do estudo alguns detalhes importantes para a consolidação do modelo proposto. Por exemplo, o conjunto de funções analisadas é limitado a um subgrupo das proteínas de ligação e, mesmo nesse subgrupo, funções muito específicas não foram contempladas.

Outras limitações desta dissertação dizem respeito a escolha do modelo de aprendizagem de máquina utilizado, das técnicas de pré-processamento realizadas e do modo como a proteína é codificada computacionalmente. Estudos mais detalhadas acerca dessas três etapas ainda precisam ser feitos para verificar possíveis aperfeiçoamentos do modelo. Por exemplo, checar se existem outros classificadores mais indicados que as máquinas de vetor de suporte ou outras maneiras de se representar as proteínas que favoreçam o processo de inferência de sua função.

## 1.6 Organização da Dissertação

Para cumprir os objetivos definidos na Seção 1.4 deste capítulo, organizou-se este trabalho em seis capítulos, incluindo esta introdução.

O capítulo 2 apresenta as noções biológicas necessárias para se entender o trabalho e essenciais para compreender sua importância.

O capítulo 3 está relacionado ao desenvolvimento de ferramentas utilizando inteligência artificial na área da bioinformática, com especial ênfase para as técnicas de redes neurais.

O capítulo 4 descreve o modelo computacional criado na presente dissertação que cumpre os objetivos descritos na seção 1.4.

O capítulo 5 mostra os resultados da análise estatística do modelo utilizando como estudo de caso algumas proteínas de ligação.

O capítulo 6 apresenta as conclusões da dissertação, suas limitações e ressalta as perspectivas de trabalhos futuros.

# Capítulo 2

# Noções Bioquímicas

## 2.1 As Ligações Interatômicas

### 2.1.1 Os Átomos

Os átomos presentes na natureza possuem um núcleo de carga elétrica positiva rodeado, a uma certa distância, por uma nuvem de elétrons carregados negativamente. O núcleo é formado por duas classes de partículas subatômicas: prótons, carregados positivamente, e nêutrons, que não possuem carga elétrica significativa.

A característica que diferencia um átomo de outro é a quantidade de prótons presente no núcleo, o número atômico ($Z$). Um átomo de carbono, elemento mais comum em moléculas orgânicas, possui seis prótons em seu núcleo, enquanto que um átomo de hidrogênio, o elemento mais leve da natureza, possui apenas um (Tabela 2.1). A carga elétrica carregada pelos prótons é exatamente igual em módulo e oposta à carga elétrica de um elétron. Assim, para que um átomo se mantenha eletricamente neutro o número de elétrons necessários é igual a $Z$.

Tabela 2.1: Elementos comuns em moléculas orgânicas

| Elemento | Número Atômico ($Z$) |
|---|---|
| Hidrogênio | 1 |
| Carbono | 6 |
| Nitrogênio | 7 |
| Oxigênio | 8 |

Os elétrons se dispõem ao redor do núcleo atômico de acordo com o diagrama energético (Figura 2.1). No diagrama é possível notar a presença de vários níveis e subníveis energéticos. Os níveis energéticos, identificados pelo

Número Quântico Principal ($n$) que é um inteiro variando de 1 a 7, correspondem às sete camadas (K, L, M, N, O, P, Q) do modelo de Rutherford-Bohr.

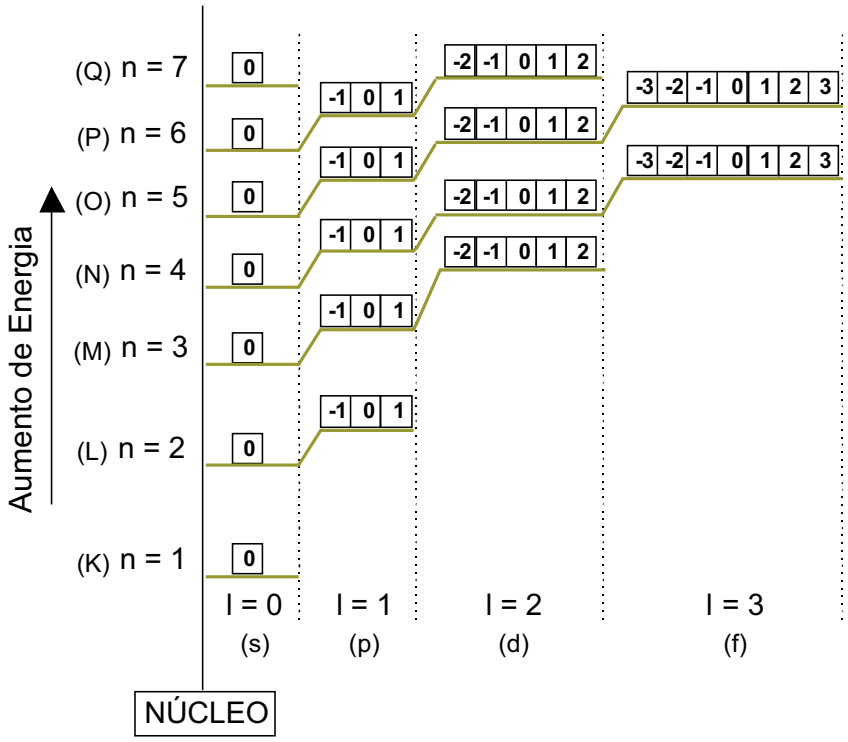


Figura 2.1: Diagrama Energético

Os subníveis energéticos são identificados pelo Número Quântico Secundário ou Azimutal ($l$), que é um inteiro variando de 0 a 3, mas habitualmente representado pelas letras 's', 'p', 'd' e 'f' respectivamente. No diagrama percebe-se que cada subnível comporta um número variável de orbitais (representados pelos quadradinhos). Os subníveis 's', 'p', 'd' e 'f' contém sucessivamente 1, 3, 5 e 7 orbitais, que são identificados pelo Número Quântico Magnético ($m$). Num dado subnível, o orbital central tem o número quântico magnético igual a zero, os orbitais da direita são numerados em ordem crescente e os da esquerda são numerados em ordem decrescente como exemplifica a figura 2.1.

Como cada orbital é capaz de receber no máximo 2 elétrons, pode-se calcular o número máximo de elétrons por camadas. Dessa forma, como a primeira camada só possui um orbital no subnível 's' e não possui outros subníveis ela apresenta capacidade para apenas 2 elétrons. A segunda camada possui os subníveis 's' e 'p', que possuem respectivamente 1 e 3 orbitais, o que resulta em 8 elétrons possíveis. A tabela 2.2 mostra o número máximo de elétrons das camadas energéticas.

A importância de conhecer a distribuição eletrônica reside no fato de os elétrons serem, entre todas as partículas subatômicas, os mais importantes

Tabela 2.2: Número máximo de elétrons por camada

| $n$ | Camada | Elétrons |
|---|---|---|
| 1 | K | 2 |
| 2 | L | 8 |
| 3 | M | 18 |
| 4 | N | 32 |
| 5 | O | 32 |
| 6 | P | 18 |
| 7 | Q | 2 |

para a formação das ligações interatômicas, já que os prótons e os nêutrons se agrupam fortemente uns aos outros no núcleo e mudam de átomo apenas em condições extremas como, por exemplo, no interior do sol. Como em tais condições não existe matéria orgânica, em tecidos vivos apenas os elétrons sofrem rearranjos o que determina as leis químicas pelas quais os átomos se combinam.

Duas regras regem a maneira como os elétrons se organizam:

1. O arranjo de elétrons em um átomo é mais estável quando todos os elétrons preenchem orbitais de menor energia. O que significa que o diagrama da figura 2.1 será preenchido de baixo para cima de acordo com o número máximo de elétrons.

2. Um átomo cuja camada mais externa está inteiramente preenchida com elétrons é especialmente estável e não reativo como, por exemplo, os gases nobres (Figura 2.2). Por outro lado, os átomos encontrados em tecidos vivos possuem a última camada energética incompleta, o que gera uma forte tendência em doar, aceitar ou compartilhar elétrons.

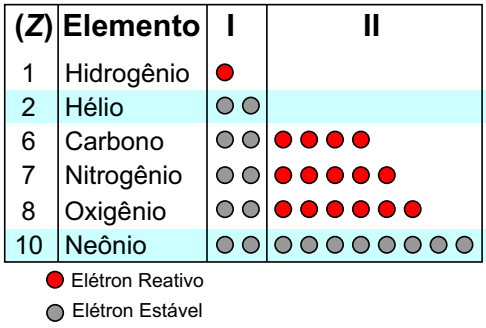


Figura 2.2: Camadas Energéticas de Átomos Comuns

A figura 2.2 mostra que o carbono precisa de quatro elétrons para completar a sua última camada, enquanto o oxigênio precisa apenas de dois. Nesse

caso, o átomo de carbono é capaz de se ligar a até quatro outros átomos, uma capacidade de ligação superior ao oxigênio. Essa capacidade de um átomo ligar-se a outros chama-se valência. Pela figura 2.2 o hidrogênio tem uma valência (monovalente), o oxigênio tem duas (bivalente), o nitrogênio três (trivalente) e o carbono possui quatro (tetravalente).

A seção 2.1.2 detalha as ligações iônicas, formadas quando um átomo doa os elétrons da sua última camada. Quando dois átomos compartilham elétrons têm-se as ligações covalentes, que são muito mais comuns em matéria orgânica. As ligações covalentes são explicadas na seção 2.1.3 e outros tipos de interação não covalentes são mencionadas na seção 2.1.4.

### 2.1.2 Ligações Iônicas

As ligações iônicas ocorrem geralmente quando o átomo que doa elétrons possui apenas um ou dois na última camada (camada de valência). Tais átomos atingem mais facilmente a configuração estável perdendo esses dois elétrons do que compartilhando.

Um exemplo de reação iônica é a que ocorre entre o metal alcalino sódio (Na) e o halogênio cloro (Cl). O sódio possui apenas um elétron em sua camada de valência e sua penúltima camada está completa, caso ele consiga doar esse único elétron atingirá a configuração estável. O caso do cloro é um pouco diferente, ele possui sete elétrons em sua camada de valência, precisando de apenas um para completá-la.

Quando um átomo de sódio encontra um átomo de cloro, um elétron abandona o sódio e preenche a última camada do cloro, deixando ambos com sua configuração estável. O resultado desse evento é a formação do sal de cozinha (NaCl), como mostra a figura 2.3.

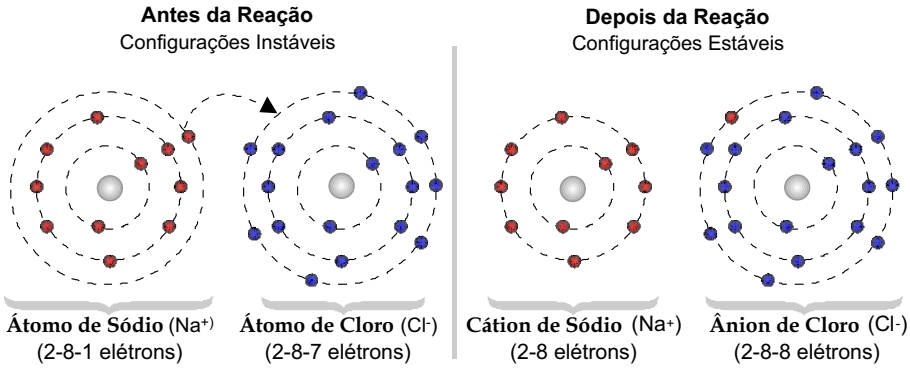


Figura 2.3: Ligação Iônica

Os átomos que participam de uma ligação iônica são chamados de íons

e podem ser classificados em cátions, os que doam elétrons, e ânions, os que recebem elétrons. Com a perda de elétrons o cátion possui carga elétrica positiva enquanto que o ânion possui carga elétrica negativa. Devido às cargas opostas, há uma atração entre os cátions e ânions que constitui a ligação iônica.

### 2.1.3 Ligações Covalentes

Ligações covalentes ocorrem quando dois átomos compartilham o mesmo elétron para preencher sua camada de valência, nesse caso não há transferência que induza o ganho ou perda de elétron. Quando um agrupamento de átomos é formado via ligações covalentes caracterizam o que é chamado de uma molécula.

A molécula mais simples da natureza é aquela constituída por dois átomos de hidrogênio (H) que se unem para formar a molécula de hidrogênio ($H_2$). Pela figura 2.4 percebe-se que cada átomo de hidrogênio possui um único elétron na camada K, mas essa camada precisa de dois elétrons para ficar completa. Assim, os átomos de hidrogênio compartilham entre si os elétrons que possuem a fim de preencherem a camada de valência.

Os dois elétrons compartilhados formam uma nuvem muito densa de carga negativa entre os dois núcleos de carga positiva, ajudando-os a se manterem unidos apesar da natural repulsão que um núcleo possui sobre o outro. As forças atrativas e repulsivas entram em equilíbrio quando os núcleos estão separados por uma distância característica chamada tamanho da ligação.

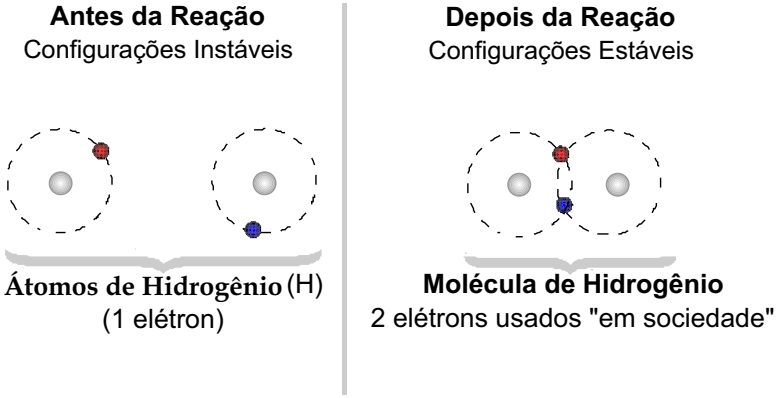


Figura 2.4: Ligação Covalente

Uma propriedade de suma importância é a força da ligação, definida como a quantidade de energia requerida para que a ligação seja quebrada. A tabela 2.3 mostra a força e o tamanho médio das principais forças interatômicas. É

possível dizer que as interações não covalentes são muito mais fracas que as covalentes, tal diferença, entretanto, se torna muito mais marcante em meio aquoso.

| Tipo de Ligação | Força ($KJmol^{-1}$) | Distância Média ($nm$) |
|---|---|---|
| Covalente | $200 - 800$ | $0,10$ |
| Ponte de Hidrogênio | 10-40 | $0,17$ |
| van der Waals | $1$ | $0,26$ |

Tabela 2.3: Forças Interatômicas Principais

#### 2.1.3.1 Tipos de Ligações Covalentes

Enquanto o hidrogênio forma apenas uma ligação covalente, outros átomos como oxigênio, nitrogênio e carbono podem formar várias (Figura 2.2). Nesses casos, é possível que cada átomo compartilhe dois de seus elétrons, tal evento é chamado de uma dupla ligação.

As duplas ligações são muito mais fortes que o compartilhamento de apenas um elétron por átomo, a ligação simples. Além disso, uma dupla ligação produz um arranjo de átomos mais rígido e menos flexível, pois impede que os átomos girem livremente como acontece nas ligações simples.

O conjunto de todas as ligações feitas por um átomo é fator determinante de sua orientação no espaço, refletindo a orientação das órbitas dos elétrons compartilhados. Dessa forma, ligações covalentes são caracterizadas por ângulos de ligação específicos. Por exemplo, se um átomo de carbono efetua quatro ligações simples, a precisa orientação das ligações forma a base para a geometria das moléculas orgânicas, como as proteínas.

Quando os átomos formadores das ligações são diferentes, atraem os elétrons compartilhados para si com intensidades diferentes. Comparados com o carbono, por exemplo, os átomos de oxigênio e o nitrogênio exercem uma forte atração sobre os elétrons, enquanto que um hidrogênio atrai os elétrons fracamente.

A maneira diferente com que os elétrons são atraídos cria pólos com carga negativa de um lado e positiva de outro, o que chamamos ligações polares, cuja importância é essencial para formar agrupamentos de moléculas, pois criam dipólos permanentes que permitem que as moléculas interajam através de forças elétricas.

Além das ligações duplas e simples existem outros tipos de ligações covalentes: as ligações triplas e as ligações dativas. As ligações triplas são formadas quando cada átomo da ligação compartilha três elétrons, essa ligação

pode acontecer com átomos trivalentes e tetravalentes, como o nitrogênio e o carbono respectivamente.

As ligações dativas envolvem um tipo diferente de compartilhamento de elétrons, pois o par de elétron compartilhado pertence a apenas um dos átomos que permite o compartilhamento sem receber nada em troca. As ligações covalentes triplas e dativas são incomuns nas proteínas.

### 2.1.4 Interações Não Covalentes

Além das ligações iônicas, existem mais tipos de interações não covalentes que desempenham um papel fundamental na formação da estrutura de macromoléculas. Tais interações são mais fracas que as covalentes, mas em conjunto tornam-se uma força efetiva capaz de 'dobrar' a conformação espacial da molécula, ou ainda, agrupar duas moléculas diferentes.

#### 2.1.4.1 Pontes de Hidrogênio

As pontes de hidrogênio são um tipo de ligação não covalente de papel primordial dentro das células. Essas ligações representam uma forma especial de interação polar no qual um hidrogênio eletropositivo é parcialmente compartilhado por dois átomos fortemente eletronegativos.

Nesse tipo de ligação, o hidrogênio pode ser visto como um próton parcialmente dissociado de um átomo doador, o que permite que ele seja compartilhado com um segundo átomo aceptor. As pontes de hidrogênio são mais fortes quando os núcleos dos três átomos envolvidos formam uma reta.

As pontes de hidrogênio serão mencionadas novamente na seção 2.2.1 devido a sua importância na organização dos solutos em meio aquoso e a sua importância no surgimento do efeito hidrofóbico (seção 2.2.2).

#### 2.1.4.2 Atrações de Van der Waals

As atrações de Van der Waals são ligações temporárias que ocorrem entre átomos apolares. A nuvem de elétrons ao redor de qualquer átomo apolar tende a sofre flutuações aleatórias, em que os elétrons se concentram em uma parte em detrimento de outras. Nesses casos, um dipolo é parcialmente formado.

Os dipolos parciais induzem em outras moléculas o surgimento de novos dipolos, o que gera uma leve atração entre os átomos. Entretanto, como muitos átomos estão em contato, acaba-se criando uma rede de forças fracas que, em conjunto, tornam-se bastante significantes.

## 2.2 A Água

A água é a substância mais importante para a vida. A busca por ambientes fora da terra onde possa haver vida normalmente recai em uma busca por planetas onde a presença de água seja possível. Nesse contexto, a distância entre o planeta e a sua estrela deve ser tal que a temperatura no planeta não atinja valores muito maiores que o ponto de ebulição da água.

A vida na terra começou nos oceanos e as condições do ambiente selou as propriedades químicas dos atuais seres vivos. Dessa forma, a vida em nosso planeta depende das propriedades físico-químicas da água.

No homem, a água corpórea total varia de $55$ a $65\%$ da massa, sendo essa percentagem menor para indivíduos obesos; estes volumes são aproximadamente $10\%$ menores para as mulheres (Murray et al. 1994). Dois terços desse volume de água está presente dentro das células (fluído intracelular) e muitas das reações intracelulares ocorrem em meio aquoso. O terço restante é fluído extracelular, concentrado principalmente no plasma.

As proteínas são macromoléculas inseridas em meio aquoso. Assim, a estrutura química da água causa um grande impacto no comportamento esperado e nas propriedades químicas, interferindo de maneira direta na formação da estrutura tridimensional e como conseqüência na função desempenhada dentro da célula.

### 2.2.1 Uma Molécula Polar de Estrutura Tetraédrica

A molécula de água é angular e possui uma pequena fração de carga elétrica negativa no átomo de oxigênio central, além de carga positiva nos hidrogênios terminais de modo a formar uma distribuição irregular de cargas elétricas internamente, como mostra a figura 2.5. Essa distribuição irregular é o que caracteriza uma molécula dipolo.

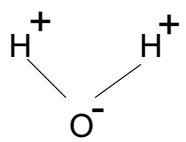


Figura 2.5: Estrutura dipolar da molécula da água

A diferença de cargas nas ligações $H^+ - O^-$ resulta do fato de o átomo de oxigênio possuir maior eletroafinidade, com tendência a atrair para si os elétrons da ligação covalente com o hidrogênio (veja seção 2.1.3). Como a di-

ferença de eletroafinidade entre o hidrogênio e o oxigênio é um fator intrínseco aos próprios átomos, a ocorrência do dipolo é permanente e independente de quaisquer fatores externos.

O fato de a molécula de água ser polar permite a ela orientar outras moléculas polares e permanecer associada a elas via um mecanismo de atração eletromagnética. Em solução aquosa ocorre que essas outras moléculas polares são, na verdade, também moléculas de água, tornando a solução aquosa uma matriz bem ordenada e fortemente coesa de moléculas de água.

Entretanto, existem muitas moléculas angulares na natureza que não possuem as mesmas características da água. Isso ocorre porque o fato de ser angular não é suficiente para entender a complexidade envolvida na interação da água com os compostos em solução.

Um outro fator que não pode ser ignorado é a molécula de água ser um tetraedro irregular com o centro ocupado pelo átomo de oxigênio (Figura 2.6). Os quatro vértices desse tetraedro são ocupados por dois átomos de hidrogênio, que formam entre si um ângulo de 105º (um tetraedro regular teria um ângulo de 109,5º) e pelos dois elétrons não compartilhados.

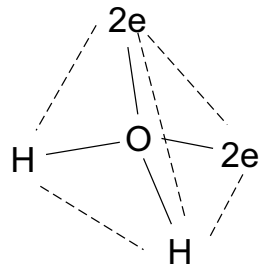


Figura 2.6: Coordenadas tetraédricas da molécula de água

A atração que ocorre entre o hidrogênio de carga positiva de uma molécula e um par de elétrons não compartilhados de outra é denominada ponte de hidrogênio e, embora sejam individualmente fracas (Tabela 2.3), em conjunto se tornam significante.

A estrutura tetraédrica da água permite formar no máximo quatro pontes de hidrogênio distintas, duas envolvendo os seus hidrogênios e duas envolvendo os pares de elétrons não compartilhados. Dessa maneira, uma solução aquosa é uma matriz ininterrupta de pontes de hidrogênio.

### 2.2.2 Moléculas em Solução Aquosa

A seção anterior explicou o comportamento da água limpa. Entretanto, é preciso ampliar o escopo desse conhecimento para abordar a maneira como a água interage com os solutos que estão imersos. Essa interação ocorrerá de modo diferente dependendo da polaridade do soluto.

Antes de mais nada é preciso entender que qualquer soluto imerso na água gera o efeito de quebrar a matriz ininterrupta de pontes de hidrogênio e, como conseqüência, faz com que a água abandone sua formação mais estável. (Israelachvili 1991).

Caso o soluto seja polar as moléculas de água poderão utilizá-lo como parceiros de pontes de hidrogênio e, com isso, restituir a matriz que fora rompida e alcançar novamente uma conformação estável.

Entretanto, caso o soluto seja apolar não haverá maneira de se repor as ligações de hidrogênio perdidas criando novas ligações com o soluto, o que ocorre nesse caso é que algumas moléculas se reorientarão para desviar do soluto a fim de ligar-se com outras que também estão na mesma situação na superfície do soluto.

Caso o soluto apolar seja pequeno será possível para água se reorientar a fim de empacotá-lo sem ter que perder nenhuma ponte de hidrogênio, isso ocorre graças a habilidade que as moléculas de coordenadas tetraédricas possuem de ligar-se ao redor de praticamente qualquer molécula inerte (Israelachvili 1991). A configuração final alcançada lembra muito uma espécie gaiola ao redor do soluto e é mostrada na figura 2.7.

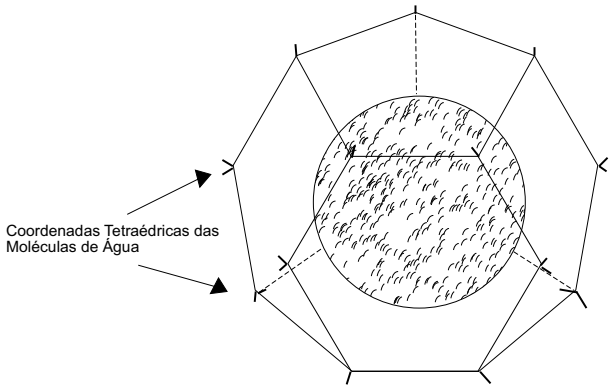


Figura 2.7: Soluto apolar em meio aquoso

Essa reorientação obriga as moléculas de água a assumirem uma conformação menos flexível, fazendo com que as moléculas próximas à superfície do soluto percam a liberdade de girar livremente e fragilizando, por conseguinte, as pontes de hidrogênio formadas. Além disso, essa nova organiza-

ção é entropicamente desfavorável e ocorre com perda de energia. (Gibas & Jambeck 2001, Israelachvili 1991).

Macromoléculas como as proteínas são formadas tanto por subgrupos polares quanto apolares. Nesse caso, a água precisará procurar os subgrupos polares que estão presentes e trazê-los para a superfície, a fim de torná-los parceiros de pontes de hidrogênio. Assim, os grupos apolares, que não podem formar ligações, são automaticamente direcionados para o centro da macromolécula.

O resultado é que uma proteína em meio aquoso se desdobra de modo a deixar em sua superfície os grupos polares, que, por isso mesmo, são chamados de hidrofílicos, enquanto o centro é composto por grupos apolares ou hidrofóbicos. Esse efeito, chamado de força hidrofóbica, tem papel importante para manter as moléculas unidas na água, além de ser central para o enovelamento das moléculas das proteínas.

## 2.3 Estrutura das Proteínas

A seção anterior explicou os aspectos gerais que determinam a conformação de macromoléculas em meio aquoso, bem como as forças que atuam nesse processo. Entretanto, detalhes mais específicos sobre a estrutura das proteínas em si ainda não foram fornecidos e são importantes para o entendimento deste trabalho. A presente seção objetiva preencher essa lacuna.

As proteínas são longas moléculas formadas por aminoácidos. Apesar de aproximadamente 300 aminoácidos diferentes ocorrerem na natureza, apenas 20 deles estão presentes nas proteínas de todos os seres vivos.

Os aminoácidos possuem uma estrutura fixa formada por um carbono central, carbono $\alpha$, ao qual são ligados quatro grupamentos: um átomo de hidrogênio, um grupamento carboxila, um grupamento amino e um grupamento **R**, que difere a cada aminoácido (Figura 2.8).

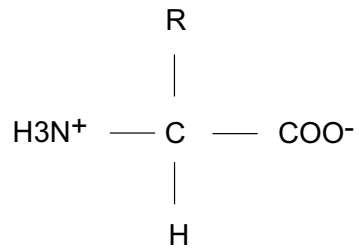


Figura 2.8: Estrutura de um aminoácido

Percebe-se por essa estrutura que os aminoácidos possuem no mínimo dois grupos funcionais que são o amínico e o carboxílico. Entretanto, é o grupamento **R** o responsável pelas diferentes propriedades químicas dos aminoácidos (Brown 1999).

Os grupamentos **R** variam consideravelmente em complexidade: na glicina, o grupamento se resume a apenas um átomo de hidrogênio enquanto o triptofano possui duas cadeias cíclicas aromáticos (Figura 2.9).

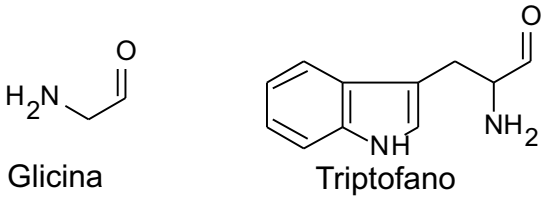


Figura 2.9: Os aminoácidos variam em complexidade

Os aminoácidos que constituem uma proteína são unidos por ligações peptídicas formadas pela condensação entre o grupamento carboxila de um aminoácido e o grupamento amino de outro. Tal reação química ocorre com a liberação de uma molécula de água. A figura 2.10 mostra o esquema de como ocorre a reação.

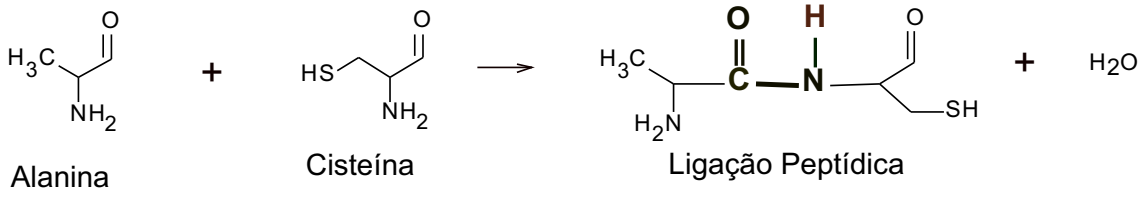


Figura 2.10: Reação peptídica

A ligação peptídica entre dois aminoácidos resulta em uma molécula com duas extremidades quimicamente distintas: uma tem um grupamento amino livre e é chamada N-terminal; na outra extremidade o grupamento livre é o carboxílico e é chamada C-terminal. Essas diferentes extremidades ocorrerão em todas as proteínas, independente de seu tamanho.

A seqüência repetitiva de átomos ao longo da cadeia polipeptídica é denominada de cadeia central. Ligados a essa cadeia estão a porção de aminoácidos não envolvidos na construção das ligações peptídicas, os grupamentos **R**, essa nova seqüencia denomina-se cadeia lateral.

Outra questão importante a ser abordada são os ângulos que podem ser assumidos pela ligação do carbono central, carbono $\alpha$, com os grupos amínico

e carboxílico, esses ângulos são respectivamente representados pelas letras gregas $\Phi$ e $\Psi$.

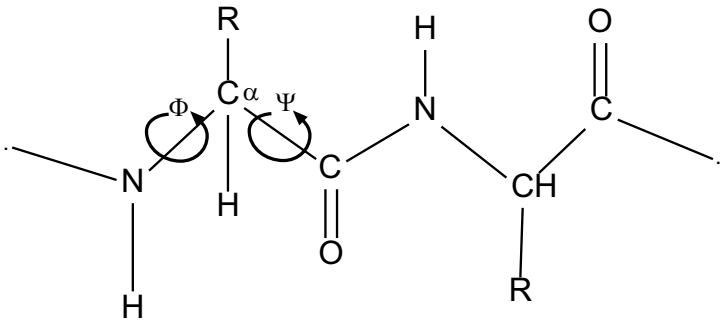


Figura 2.11: Ângulos dos aminoácidos em ligações peptdícas

Cada um desses ângulos possui liberdade de girar livremente (Figura 2.11). Entretanto, muitas combinações de ângulos não são possíveis para muitos aminoácidos devido a restrições espaciais do grupo **R** e de vizinhos na cadeia. A conformação espacial da proteína é determinada pelos ângulos $\Phi$ e $\Psi$, cuja distribuição para os aminoácidos em uma dada proteína pode ser desenhada em um gráfico chamado de gráfico de Ramachandram (Mount 2001).

### 2.3.1 Níveis Diferentes da Estrutura Protéica

Para facilitar o estudo da estrutura protéica quatro níveis são reconhecidos nas moléculas de proteínas: a estrutura primária, secundária, terciária e quaternária.

A estrutura primária é a seqüência em que ocorrem os aminoácidos nas moléculas de proteína, enquanto que a estrutura secundária possui maior complexidade com a identificação de elementos regulares localizados na conformação espacial da proteína. As duas estruturas regulares mais importantes são a hélice $\alpha$ e a folha $\beta$, ambas são estabilizadas pela ação das pontes de hidrogênio (Brown 1999).

A estrutura terciária é a conformação tridimensional propriamente dita, sendo formada pelo enovelamento dos componentes da estrutura secundária. Como já foi mencionado na seção 2.2, a tendência natural é a cadeia se empacotar de modo a 'esconder' os grupamentos **R** não polares (Israelachvili 1991).

A estrutura quaternária se refere ao modo pelo qual duas moléculas protéicas são orientadas a fim de formar uma multissubunidade, podendo envolver duas ou mais moléculas iguais ou ainda várias moléculas polipeptídicas diferentes.

As estruturas superiores (secundária, terciária e quaternária) são determinadas pela estrutura primária, sendo isto uma das bases pela qual a modelagem computacional de proteínas é possível. Esta teoria pode ser provada aquecendo uma molécula de modo a fazê-la perder suas estruturas superiores, quando a molécula for resfriada guarda a capacidade inata de voltar a assumir as estruturas anteriores (Anfinsen 1973).

### 2.3.2 Elementos da Estrutura Secundária

Esta seção detalha os aspectos relevantes das duas principais estruturas secundárias encontradas nas proteínas: hélice $\alpha$ e folha $\beta$.

#### 2.3.2.1 Hélices $\alpha$

As hélices $\alpha$ são as estruturas mais abundantes nas proteínas. Elas possuem $3,6$ aminoácidos por giro da hélice com uma ponte de hidrogênio formada de quatro em quatro resíduos, ou seja, cada grupo carboxílico de um aminoácido de uma dada posição $n$ na seqüência formará uma ponte de hidrogênio com o grupo amínico do aminoácido da posição $n+4$ (Figura 2.12). Em média, uma hélice $\alpha$ possui 10 aminoácidos, mas essa quantidade pode variar de 5 a 40.

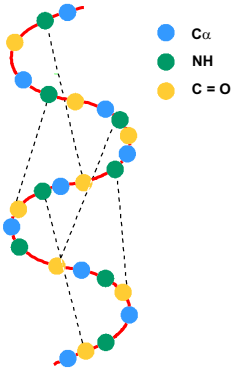


Figura 2.12: Hélice $\alpha$

A estrutura de uma hélice $\alpha$ é fixa, as hélices possuem giro para direita, mas pequenas seções de 3 a 5 aminoácidos com giro para esquerda podem ser encontradas. Em média, os ângulos $\Phi$ e $\Psi$ são aproximadamente 60 graus e 40 graus respectivamente (Mount 2001).

As regiões com grande probabilidade de possuir hélices $\alpha$ são aquelas ricas em alanina, ácido glutâmico, leucina e metiolina; além de serem pobres em glicina, tirosina e serina.

A localização mais comum de uma hélice $\alpha$ é na superfície das moléculas, onde podem fazer uma interface com o meio aquoso com o grupamento **R** sempre voltado para fora da hélice. Seguindo o padrão relatado na seção 2.2.2 para macromoléculas grandes em meio aquoso, a face da hélice voltada para o interior da proteína tende a possuir componentes hidrofóbicos e a face externa hidrofílicos (Mount 2001).

#### 2.3.2.2 Folhas $\beta$

As folhas $\beta$ (Figura 2.13) são estruturas menos comuns que as hélices $\alpha$ e sua predição via métodos computacionais é bem mais difícil, pois podem ser formadas por subseqüências muito distantes na cadeia.

As folhas $\beta$ são formadas por pontes de hidrogênio entre, em média, 5-10 aminoácidos consecutivos em uma porção da cadeia com outros 5-10 aminoácidos em outra parte, que pode ser logo abaixo ou em regiões muito distantes.

As folhas $\beta$ podem ser classificadas em paralelas, anti-paralelas e mistas. As paralelas são aquelas em que as seqüências que a formam estão na mesma direção, já nas anti-paralelas essas seqüências estão em direções opostas. As folhas mistas apresentam uma mistura de paralelas e anti-paralelas (Mount 2001).

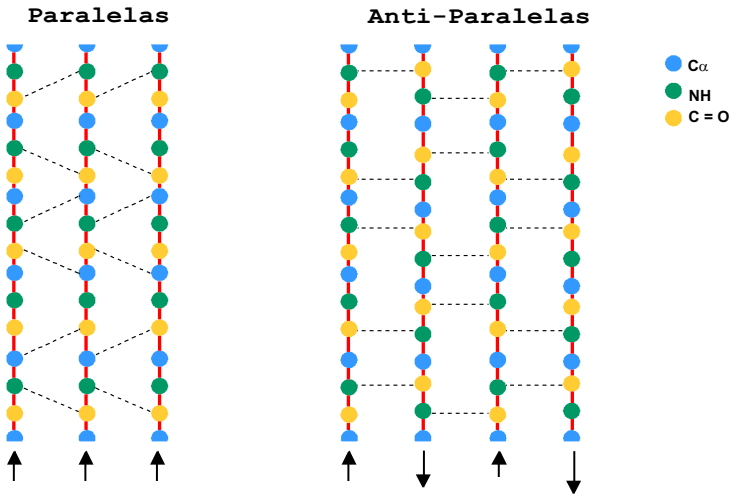


Figura 2.13: Folha $\beta$

O padrão de ligações de hidrogênio varia conforme a classificação da folha em paralela, anti-paralela e mista e, além disso, pode variar de uma folha anti-paralela a outra. A figura 2.13 exemplifica a configuração de folhas $\beta$ paralelas e anti-paralelas.

## 2.4 Função das Proteínas

Nas seções anteriores foi visto que cada tipo de proteínas possui uma seqüência precisa de aminoácidos que permite a formação de uma conformação espacial particular. Entretanto, ainda não foi mencionado que as proteínas não são moléculas rígidas e inflexíveis, elas podem possuir partes móveis precisamente construídas de modo que a ação mecânica trabalhe em conjunto com as propriedades químicas (Alberts et al. 2002).

A capacidade de agir em praticamente todos os ciclos metabólicos de qualquer organismo presente na natureza é, em parte, gerada pela interação entre as ações mecânicas e as propriedades químicas. O resultado mais visível dessa interação é a grande flexibilidade das proteínas em se acoplar a outras moléculas.

A habilidade de se acoplar a outras moléculas permite às proteínas agir como catalisadoras, receptoras de sinais, reguladoras de expressão gênica e muitas outras funções. A região da proteína onde ocorre o acoplamento é chamada de sítio de ligação e será detalhada na próxima seção. Por conseguinte, alguns exemplo ilustrativos da ação de algumas proteínas serão mostrados na seção 2.5.

### 2.4.1 Sítios de Ligação

A ação biológica de uma proteína depende de sua interação física com outras moléculas. Assim, as proteínas possuem a capacidade de ligação, que pode ser estável ou de curta duração. Em ambos os casos, a proteína se liga apenas a uma molécula ou a um grupo específico de poucas moléculas, essa propriedade de se ligar a apenas um pequeno grupo de moléculas é chamada de especificidade.

A substância que se acopla a uma proteína, independente de suas particularidades químicas ou físicas, é conhecida como um ligante para aquela proteína.

A especificidade depende da formação de uma série de ligações fracas, tais como pontes de hidrogênio, forças iônicas e atrações de van der Walls favorecidas pelas relações hidrofóbicas e hidrofílicas. Como individualmente cada uma dessas ligações é fraca, são necessárias muitas ligações em conjunto para que o acoplamento seja efetivo. Isso só é possível se o contorno da superfície da molécula ligante preencher corretamente as cavidades da superfície das proteínas, como uma chave no cadeado.

A cavidade na superfície da proteína onde o ligante é acoplado se chama

sítio de ligação e pode ser formada por aminoácidos distantes na cadeia primária, que se aproximam devido ao enovelamento da proteína na estrutura terciária, daí a importância da estrutura terciária na definição da função da proteína.

Apesar de os átomos do núcleo da molécula não entrarem em contato direto com o ligante, possuem importância crucial como esqueleto da molécula que dá à superfície os contornos e propriedades químicas necessários, pequenas mudanças de aminoácidos nesse esqueleto pode fazer com que a molécula mude completamente a sua conformação, acabando por destruir os sítios de ligação (Alberts et al. 2002).

Alguns sítios de ligação podem funcionar como alças, para que a célula leve a proteína do lugar onde é sintetizada para o lugar onde ela desempenhará a sua função. Exemplos desse tipo de funcionalidade serão vistos na seção 2.5.

### 2.4.2 Domínio Funcional da Proteína

Estudos da conformação, função e evolução de proteinas revelaram a importância de uma unidade de organização chamada domínio funcional, uma pequena estrutura que pode conter entre $40$ e $350$ aminoácidos e constituem a unidade modular pelas quais muitas proteínas longas são construídas.

Os diferentes domínios funcionais de uma proteína são geralmente associados com funções diferentes. A proteína Kinase, mostrada na figura 2.14, é uma proteína de sinalização especializada em mensagens que controlam o crescimento das células. Ela possui quatro domínios funcionais: os domínios $SH2$, mostrado na cor azul, e $SH3$, mostrado na cor verde, possuem função regulatória enquanto que os outros dois domínios, em amarelo, são responsáveis pela atividade catalítica.

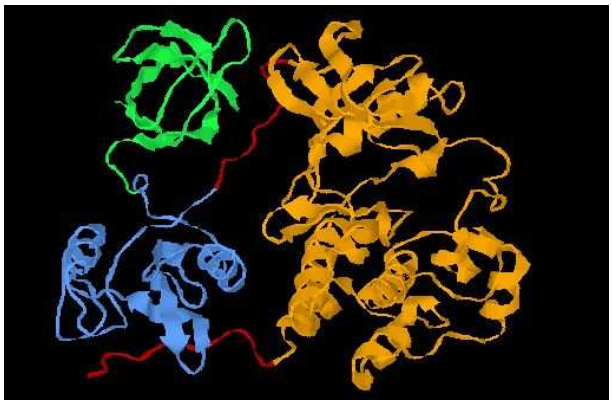

Figura 2.14: Proteína Kinase possui vários domínios funcionais

As proteínas menores possuem apenas um domínio, enquanto as maiores possuem cerca de uma dúzia deles. Os domínios são conectados uns aos outros por pequenas cadeias polipeptídicas. Como a figura 2.14 ilustra, o núcleo central do domínio pode ser construído a partir de $\alpha$ hélices, folhas $\beta$ ou combinações desses dois elementos.

### 2.4.3 Classificação em Famílias

Uma proteína com conformação espacial estável e propriedades úteis dentro da célula pode sofrer várias modificações durante o processo evolutivo de modo a poder executar novas funções. Eventos desse tipo ocorreram várias vezes no passado e muitas proteínas atualmente podem ser agrupadas em famílias onde cada membro possui características que lembram os outros.

Os vários membros de uma família podem possuir funções distintas, já que algumas mudanças de aminoácidos são selecionadas no curso da evolução por resultarem em atividade biológica significativa. Isso cria famílias cujos membros possuem atividade biológicas diferenciadas.

Entretanto, muitas mudanças de aminoácidos são completamente neutras, não gerando vantagens nem destruindo a estrutura funcional das proteínas. Isso resulta em famílias onde os membros possuem a mesma atividade biológica, apesar de certas diferenças em sua estrutura primária.

Finalmente, como as mutações são um processo aleatório, podem haver mudanças danosas, que alteram a conformação espacial a ponto de destruir os sítios de ligação presentes na superfície ou alterar as propriedades químicas necessárias para que a molécula crie uma interação estável com o ligante. Nesse caso, o processo de seleção natural tende a eliminar o ser vivo onde essa mutação ocorreu, impedindo as proteínas de seguirem para gerações futuras.

As mutações danosas geralmente são aquelas ocorridas no domínio funcional (Seção 2.4.2) da proteínas, mas as mais comuns são neutras, esse fato acaba por gerar um conjunto de proteínas diferentes, mas que possuem domínios semelhantes e, por conseguinte, funções parecidas. Pode-se pensar nos domínios como módulos especializados que fazem partes de várias proteínas diferentes.

Uma característica desses módulos especializados é a facilidade com que eles podem ser integrados em outras proteínas. As integrações ocorrem porque duplicações gênicas, nesse caso dos genes que expressaram o domínio funcional, são comuns na evolução dos genômas. Os módulos duplicados podem ser ligados em série para formar estruturas mais extensas com sítios funcionais distintos.

## 2.5 Exemplos da Atuação de Proteínas

As proteínas são moléculas muito flexíveis e executam funções muito diferentes em um organismo. Dessa forma, uma das maneiras de classificar as proteínas é de acordo com a função biológica em que atuam, como por exemplo: proteínas estruturais, catalíticas ou de transporte. O grupo de proteínas mais comum é o de proteínas catalíticas, cujos membros são conhecidos como enzimas e, por sua vez, são classificados pelo tipo de reação que catalizam (seção 2.5.3).

A tabela 2.4 ilustra as principais atividades protéicas em nível molecular e cita algumas proteínas que as desempenham.

| Função | Proteínas (Exemplos) |
|---|---|
| Estrutural | Colágeno, queratina |
| Contração | Actina, miosina |
| Enzimática | Hexocinase, RNA Polimerase |
| Transporte | Hemoglobina |
| Regulatória | Insulina |
| Proteção | Imunoglobulina |
| Armazenamento | Ovalbulina |

Tabela 2.4: Principais funções das proteínas

As subseções seguintes apresentam exemplos ilustrativos das funções mais comuns desempenhadas pelas proteínas, detalhando um pouco mais a tabela 2.4. Entretanto, é importante ressaltar que a lista não é exaustiva e muitas funções importantes não foram abordadas.

### 2.5.1 Proteínas Reguladoras

A cadeia de DNA de um organismo carrega uma imensa quantidade de informação. Alguns genes são tão importantes que a informação que eles carregam é necessária a todo momento, tais genes são chamados de genes de manutenção e incluem os genes codificadores de ribossomos, codificadores de algumas enzimas como a RNA polimerase e outros genes envolvidos em vias metabólicas básicas (Brown 1999).

Outros genes são necessários apenas em determinadas circunstâncias, sendo, dessa forma, necessários mecanismos que impeçam sua expressão nos momentos em que são indesejáveis, caso contrário muita energia seria desperdiçada em vão.

Dentre os vários mecanismos que a célula possui para impedir a expressão de certos genes, um dos mais interessante é a utilização de proteínas espe-

cializadas em parar a transcrição de determinado intervalo da cadeia de DNA, tais proteínas são chamadas de proteínas reguladoras.

As Proteínas reguladoras reconhecem as seqüências específicas pelo acoplamento do seu domínio funcional ao DNA. Esses domínios são relativamente pequenos, em geral menores que $100$ resíduos de aminoácidos (Branden & Tooze 1991) e possuem geralmente uma estrutura comum composta de duas hélices $\alpha$ ligadas por uma pequena região polipeptídica no formato de um giro. As duas hélices $\alpha$ possuem a mesma orientação e apresentam propriedades funcionais comuns em todas as proteínas reguladoras.

Na seção 4.1, uma representação hierárquica das funções das proteínas será apresentada e, para tornar o texto da dissertação mais intuitivo, usou-se como exemplo a função de regulação, mais especificamente a do uso da lactose na *E. Coli* (Brown 1999).

A regulação do uso da lactose na *E. Coli* é um exemplo de como as proteínas reguladoras agem. A lactose é um dissacarídeo composto de uma única glicose ligada a uma única galactose. A glicose é o sacarídeo necessário como fonte de energia na célula, para ser utilizada é preciso que a lactose seja transportada do meio extracelular para dentro da célula e depois quebrada a fim de isolar a glicose.

Três enzimas trabalham para essa função:

**permease (*lacY*)** Transporta a lactose do meio extracelular para dentro da célula.

**$\beta$-galactosidade (*lacZ*)** Quebra a lactose em glicose e galactose.

**$\beta$-galactosídeo transacetilase (*lacA*)** O papel exato não é conhecido por completo.

Quando não existe lactose no meio extracelular essas enzimas não são necessárias, havendo na célula apenas um pequeno número de moléculas de cada uma em quantidades bassais. Quando a lactose é encontrada, a síntese de proteínas é rapidamente induzida e níveis de até $5.000$ moléculas de cada enzima são alcançados. Como as proteínas mantém quantidades semelhantes conclui-se que são induzidas ao mesmo tempo e na mesma quantidade.

Esse fato indica que os genes dessas enzimas ocupam posições em série na cadeia de DNA (Figura 2.15) e são transcritos na mesma molécula de mRNA, controlada por um único promotor que antecede o gene *lacZ* e um único finalizador posterior ao gene *lacA* (Brown 1999).

Pouco antes na cadeia de DNA existe o gene *lacI*, que é expresso independentemente, pois possui sua própria região promotora e seu próprio finalizador. O produto do gene *lacI* é a proteína responsável por regular a expressão dos outros três genes. Assim, se *lacI* for inativado por alguma mutação a célula continuamente produzirá *lacZ*, *lacY* e *lacA* independente da existência da lactose no meio extracelular.

O produto gênico de *lacI*, portanto, é uma proteína reguladora chamada de repressor. Elá é capaz de se ligar à molécula de DNA em um lugar chamado operador, que fica próximo à região promotora (Figura 2.15) de modo que quando o repressor está acoplado o acesso ao promotor é bloqueado devido ao tamanho da proteína, desse modo a RNA polimerase não pode se ligar ao DNA e a transcrição dos três genes *lac* não ocorre (Figura 2.16).

Entretanto, isso apenas acontece se a lactose não existir no meio extracelular, pois caso contrário a proteína repressora se ligará a um isômero da lactose, a alolactose. Quando a célula encontra um novo suprimento de lactose, ela captura algumas moléculas e as converte em alolactose que se liga ao repressor, causando uma mudança na conformação desta última de tal modo que ela não é mais capaz de se ligar ao operador, permitindo que a RNA Polimerase o encontre e inicie a transcrição (Figura 2.17).

No caso acima mencionado, a alolactose age como um indutor, pois induz a transcrição quando existir um suprimento de lactose no meio extracelular. Caso as enzimas de metabolismo de lactose esgotem o suprimento disponível, o número de ligações repressor-indutor diminui e as moléculas repressoras livres começam a predominar. Estes repressores livres recuperam a sua conformação original e, assim, podem se ligar novamente ao operador e impedir a transcrição dos genes.

Como foi mencionado anteriormente, o sacarídeo realmente necessário para produzir energia para célula é a glicose. No caso de já existir uma suficiente quantidade de glicose no meio extracelular a célula não precisa quebrar a lactose para produzí-la. Nesse caso é necessário um outro mecanismo capaz de manter os genes *lac* desligados.

Esse novo mecanismo envolve uma segunda proteína reguladora, a proteína CAP, e uma segunda localidade de ligação, o sítio CAP. A glicose é uma

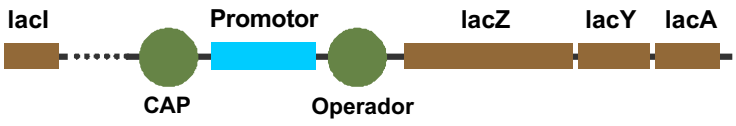


Figura 2.15: Genes envolvidos na utilização da lactose

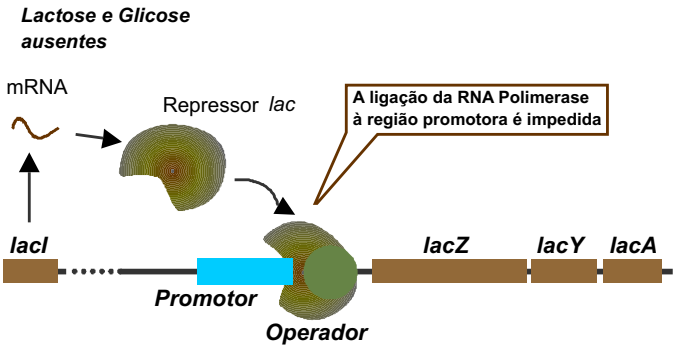


Figura 2.16: Lactose ausente no meio extracelular

inibidora de uma proteína chamada adenilato ciclase que catalisa a formação de um nucleotídeo chamado de cAMP (AMP cíclico) a partir de moléculas de ATP. Caso exista uma grande quantidade de cAMP na célula a proteína CAP forma com ele o complexo CAP-cAMP que se liga ao sítio CAP e estimula a ligação da RNA polimerase ao promotor. Dessa forma, através do controle da quantidade de cAMP na célula, a glicose indiretamente regula a transcrição dos genes *lac*.

O propósito do exemplo acima mencionado é servir como referência de capítulos futuros, pois ilustra vários aspectos funcionais de proteínas como a atuação como inibidores, catalizadores e reguladores.

### 2.5.2 Proteínas de Proteção

O sistema imunológico é um mecanismo de defesa contra parasitas externos como vírus e bactérias. Três propriedades são essenciais para que a operação de defesa seja satisfatória: reconhecimento específico das moléculas externas,

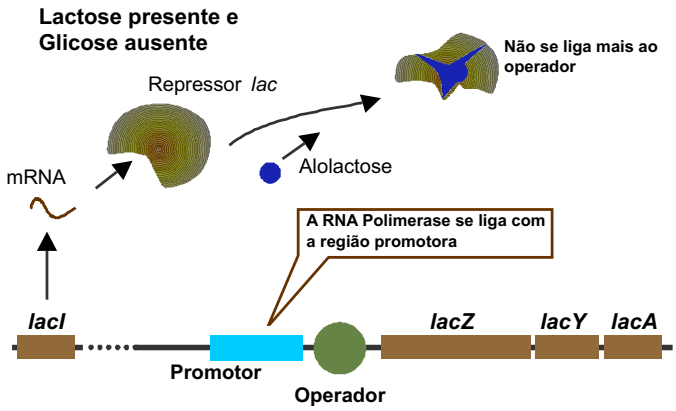


Figura 2.17: Lactose presente no meio extracelular

a habilidade de destruir o parasita e um mecanismo de memória que permita uma resposta mais rápida em uma segunda infecção pelo mesmo parasita.

Anticorpos, ou imunoglobulinas, são as proteínas produzidas em resposta às moléculas externas, chamadas antígenos. Cada anticorpo se acopla fortemente à molécula alvo particular inativando-a ou criando marcadores que levam a sua destruição por macrófagos ou outras células. Como existem bilhões de antígenos diferentes, os organismos dos seres vivos precisam ser capazes de criar bilhões de anticorpos (Branden & Tooze 1991).

Os anticorpos são moléculas em formato de Y com dois sítios de ligação idênticos complementares a uma pequena porção da superfície da molécula do antígeno (Alberts et al. 2002). A estrutura básica compreende duas cadeias leves idênticas (*light chain*) e duas cadeias pesadas (*heavy chain*), que são mantidas juntas por pontes dissulfídricas (Figura 2.18).

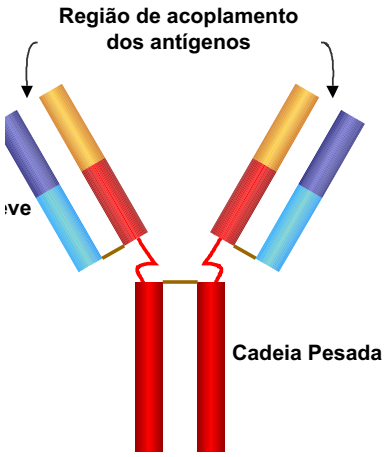


Figura 2.18: Anticorpos

Existem dois isótopos diferentes para as cadeias leves, entretanto nenhuma distinção funcional entre eles é conhecida. As cadeias pesadas possuem cinco diferentes isótopos que dividem as imunoglobulinas em classes funcionais diferentes, cada uma com propriedades diferentes na eliminação dos antígenos. (Branden & Tooze 1991).

### 2.5.3 As Enzimas

As enzimas são proteínas que possuem função catalizadora, ou seja, função de acelerar as reações bioquímicas que ocorrem em meio intracelular. A maioria dessas reações ocorreria muito lentamente se não fossem catalizadas por enzimas.

É importante frizar que o papel enzimático é um dos mais comuns exercidos pelas proteínas, pois, essencialmente, todas as reações bioquímicas são catalizadas por enzimas (Murray et al. 1994).

Cada enzima catalisa um pequeno número de reações, freqüentemente apenas uma, sendo, portanto, catalisadoras com especificidade de reação. A determinação da atividade das enzimas, dessa forma, está intimamente ligada a identificação do substrato e das vias metabólicas sobre a qual ela atua.

A simples função de catalizar uma dada reação pode parecer, a primeira vista, uma atividade banal e sem maiores implicações práticas no organismo. Entretanto, essa idéia distorcida é facilmente refutada quando se analisa a importância biomédica das enzimas, pois muitas doenças são provocadas devido às anormalidades da síntese de enzimas.

Quando as células são agredidas por inflamações ou por diminuição do suprimento de sangue, certas enzimas vazam para o plasma. A determinação da atividade destas enzimas tem se tornado parte integrante do processo de diagnóstico de um número de importantes doenças como infarto do miocárdio (Murray et al. 1994). Além disso, as enzimas também podem ser utilizadas em processos terapêuticos (Murray et al. 1994).

A especificidade das enzimas é sua propriedade mais significativa. As velocidades de processos podem, dessa forma, ser reguladas com precisão e eficiência. Todavia, a maioria das enzimas podem catalizar o mesmo tipo de reação com um número reduzido de substratos estruturalmente relacionados. Vale ressaltar que esta especificidade das enzimas quanto a reação e aos substratos nada mais é que uma conseqüência da especificidade do sítio de ligação mencionada na seção 2.4.1.

### 2.5.4 Outras Funções

Algumas funções que constam na tabela 2.4 não foram explicadas nas seções anteriores. Essa seção preenche essa lacuna.

**Proteínas Estruturais** São responsáveis por formar algumas partes da estrutura dos organimos como, por exemplo, o colágeno que está associado a tendões e ossos.

**Proteínas Contráteis** Permitem a movimentação dos organismos atuando em músculos como a miosina e actina ou em cílios e flagelos como a dineína.

**Proteínas de Transporte** Permitem o movimento direcionado de substâncias como moléculas e íons por todo o corpo. Os exemplos são: hemoglobulina (hemoglobina), que transporta oxigênio pela corrente sangüínea

dos vertebrados e a hemocianina, que realiza a mesma função em alguns invertebrados.

**Proteínas de Armazenamento** Também conhecida como atividade de reservatório de nutrientes, permite o armazenamento de substratos nutritivos para serem utilizados futuramente pelo organismo. Os exemplos são a ovoalbumina que armazena aminoácidos na clara do ovo e a alburitina, armazena ferro no fígado.

# Capítulo 3

# Aprendizagem de Máquina e Bioinformática

O presente capítulo visa apresentar os métodos de aprendizagem de máquina, em especial os métodos de redes neurais utilizados para criar o sistema deste trabalho.

O capítulo é dividido da seguinte maneira: na primeira parte, encontram-se noções introdutórias de aprendizagem de máquina (seção 3.1), na segunda parte, seção 3.2, são mostradas as principais técnicas de classificação usando aprendizagem de máquina e, por conseguinte, seção 3.3, a teoria de redes neurais é apresentada.

## 3.1 Conceitos Iniciais

Aprendizagem de máquina é um termo que engloba um conjunto de metodologias e abordagens com o objetivo de programar sistemas capazes de reconhecer padrões e comportamentos em dados que representam exemplos de acontecimentos do mundo real ou experiências passadas.

Segundo Haykin (2001), no contexto de redes neurais, o termo "aprendizagem" é definido da seguinte forma: "aprendizagem é um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida. O tipo de aprendizagem é determinado pela maneira pela qual a modificação dos parâmetros ocorre".

Em outras palavras, ao modelar um determinado problema, o termo "aprendizagem" se refere a executar um algoritmo que induz um modelo usando dados de treinamento ou experiências passadas. Dessa forma, os dois objetivos de qualquer projeto de aprendizagem são: induzir o modelo processando uma

grande quantidade de dados e realizar inferências a partir dele. Dentre esses objetivos, processar a grande quantidade de dados é o que exige mais tempo e esforço computacional.

Os métodos de aprendizagem de máquina fazem parte de um conjunto maior de ferramentas computacionais utilizadas para lidar com um grande volume de dados biológicos gerados diariamente. Tais ferramentas devem ser abrangentes a ponto de permitir que, partindo-se de uma mera descrição dos dados, seja obtido um modelo do conhecimento que possa ser validado por meio de técnicas estatísticas.

Existem vários casos na bioinformática onde a aprendizagem de máquina pode ser utilizada para extração de conhecimento. Larrañaga et al. (2006) criou uma categorização nos seguintes domínios:

**Genômica** Se refere à extração do conhecimento a partir de informações contidas nas seqüências de DNA como, por exemplo, extrair a localização e a estrutura de genes, identificar elementos regulatórios e regiões codificantes, prever a função do gene e a estrutura do RNA associado.

**Proteômica** Como a conformação espacial das proteínas desempenha um papel chave na funcionalidade, a maioria das aplicações computacionais se relacionam com isso, por exemplo, predição da estrutura e função das proteínas.

***Microarrays*** É um campo relacionado ao gerenciamento de dados experimentais, mais tipicamente, identificações de quais genes são expressos em determinadas circunstâncias como, por exemplos, vegetais em locais secos ou sem incidência direta de luz solar.

**Sistemas Biológicos** Domínio relacionado a modelos de processos que ocorrem na célula, especialmente redes de tradução de sinais e caminhos metabólicos.

**Evolução** Árvores filogenéticas, representações esquemáticas da evolução das espécies, podem ser construídas a partir de informações diversas, como informações fenotípicas, metabólicas e gênicas (comparação de genomas).

**Mineração de Textos** Com a crescente aplicação de técnicas computacionais houve um aumento das publicações científicas sobre o tema bioinformática. Tais publicações podem ser consideradas uma nova fonte de informações e as técnicas de mineração de texto (do inglês *text mining*) são utilizadas.

## 3.2 Técnicas de Classificação

Um problema de classificação consiste em, dado um conjunto de elementos divididos em classes e uma instância desse conjunto, atribuir uma classe a essa instância de acordo com as suas características que se assemelham com os demais membros da mesma classe.

Em outras palavras, o agrupamento de objetos em classes é feito pela escolha de características que descrevem tais objetos e são capazes de diferenciá-los de objetos de outras classes. Basicamente existem dois tipos de classificação:

**Classificação não supervisionada** Nesse tipo de classificação o interesse é agrupar consistentemente os dados de modo a permitir a descoberta de similaridades e diferenças entre os padrões (capturar uma organização inerente aos dados), bem como derivar conclusões úteis a partir deles.

**Classificação supervisionada** Nesse tipo de classificação há dados organizados por rótulos (classes) que são utilizados como exemplos de treinamento de um sistema que, depois de treinado, será capaz de fazer inferências quanto aos rótulos de novos dados.

No presente trabalho utilizou-se apenas a classificação supervisionada e, por isso, a mesma será explanada com um pouco mais de detalhes na seção 3.2.1.

### 3.2.1 Classificação Supervisionada

Quando as regras de classificação são induzidas de exemplos de treinamento, diz-se que a classificação ocorreu via um método de aprendizagem supervisionada ou, simplesmente, classificação supervisionada.

No campo da classificação de proteínas por suas funções, cada instância é composta de um vetor de características abstraídas da proteína e a classe de função a que esta proteína pertence. Pode-se então, entender a instância como sendo algo do tipo $(\overline{x}, y)$, onde $\overline{x}$ é o vetor de atributos e $y$ é um nome que identifica a classe.

Existem vários paradigmas de classificação, cada um com suas vantagens e desvantagens, facilidades e limitações, sendo impossível concluir que existe um melhor classificador para todos os possíveis conjuntos de treinamento. Alguns dos métodos de classificação que possuem ampla aceitação na bioinformática são: árvores de decisão (Breiman et al. 1993), redes bayesianas (Duda

& Hart 1973) e redes neurais artificiais (Haykin 2001), todos com aplicações nas diversas categorias citadas na seção 3.1.

As árvores de decisão são uma maneira intuitiva de classificar os padrões através de um modelo que se assemelha a uma seqüência de questões que devem ser respondidas, tais questões são bastante legíveis no formato de árvore, na qual a primeira questão corresponde ao nó raiz e as questões seguintes estão conectadas a ela através de arcos.

Os diferentes caminhos a partir de um nó correspondem a diferentes valores possíveis. Baseado na resposta segue-se o caminho apropriado até chegar a um nó folha (sem caminhos possíveis) que indica a previsão dada pelo sistema. É importante que os caminhos sejam mutuamente exclusivos e exaustivos, isto é, um e somente um caminho pode ser seguido e sempre haverá um caminho a ser seguido até chegar ao nó folha.

O problema de induzir uma árvore de classificação a partir de um conjunto de dados pode ser visto como um problema de organizar as variáveis na árvore, dando ênfase àquelas mais importantes no processo de decisão.

Apesar de intuitivas, as árvores de decisão não são adequadas para uma série de problemas, já que os nós das árvores não devem ser infinitos. Esse fato é um limitante que impede a utilização de árvores de decisão no presente trabalho, já que a camada de entrada é composta de uma superfície de decisão difícil de ser discretizada (Larrañaga et al. 2006).

As redes bayesianas são um tipo de classificador que utiliza o teorema de bayes, uma formulação rígida de regras probabilísticas, como base para inferir a classe mais provável de um dado exemplo. Dentre os vários tipos de modelos bayesianos, o mais utilizado na bioinformática é o mais simples, o Naive Bayes (Larrañaga et al. 2006).

O modelo Naive Bayes é construído supondo-se independência condicional do vetor de características dada a classe rotulada. A dificuldade desse modelo é que essa suposição é violada em numerosos casos, sendo difícil obter um modelo em que a suposição seja matematicamente comprovada (Larrañaga et al. 2006). Entretanto, apesar disso, o paradigma funciona bem em muitas situações.

As redes neurais artificiais são um paradigma capaz de resolver vários problemas dada uma precisão estabelecida. O paradigma de redes neurais foi o utilizado para se criar o modelo apresentado nesta dissertação e, dessa forma, a seção 3.3 as apresenta com mais detalhes.

## 3.3 Noções Básicas de Redes Neurais Artificiais

As redes neurais artificiais, usualmente conhecidas apenas como redes neurais, são modelos matemáticos motivados pelo reconhecimento de que o cérebro humano processa informações de uma forma não-linear e paralela, organizando os constituintes estruturais, neurônios, de forma a realizar processamentos rapidamente (Haykin 2001).

O cérebro humano (ou o de outros animais) possui habilidade de desenvolver suas próprias regras através de experiências que vão sendo acumuladas com o passar do tempo. Em redes neurais artificiais, um modelo pré-estabelecido é capaz de reconhecer uma série de regras a partir de exemplos (experiências) empregando entidades de processamento simples, denominadas neurônio.

De uma forma mais específica, as redes neurais artificiais podem ser vistas como grafos conexos, conforme a figura 3.1, onde os nós são os neurônios e os arcos são as forças de conexão conhecidas como pesos sinápticos. É importante notar na figura 3.1, a organização dos neurônios em camadas, que podem ser a camada de entrada, as camadas ocultas e a camada de saída.

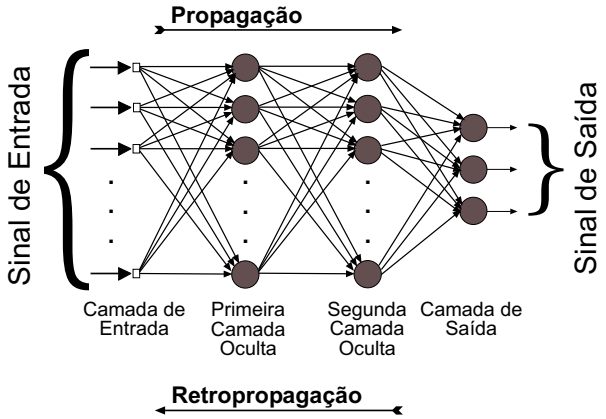


Figura 3.1: Redes Neurais

O modo como essa estrutura consegue induzir regras a partir de experiência é o ponto mais interessante de redes neurais. Basicamente, o aprendizado das regras ocorre mediante a simples alteração dos pesos sinápticos que ligam os neurônios e, desse modo, gerando uma saída diferente que, espera-se, esteja mais de acordo com as experiências fornecidas.

### 3.3.1 Os Neurônios

Os neurônios, unidade de processamento da informação, são de fundamental importância para que a rede neural cumpra seu objetivo, pois formam a base para o projeto de redes neurais artificiais. A figura 3.2, adaptada de Haykin (2001), mostra o modelo de um neurônio. Por essa figura é possível perceber que três elementos estão envolvidos:

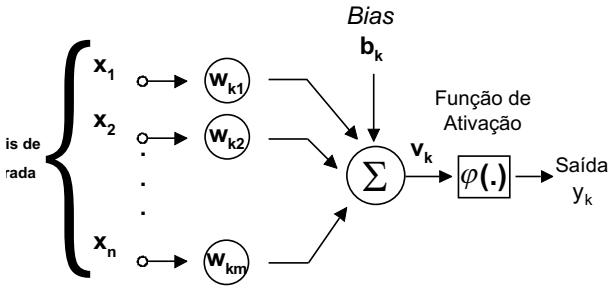


Figura 3.2: Modelo de um Neurônio

**Conjunto de sinapses (W)** São as ligações das entradas com o neurônio. Cada ligação possui um valor (peso), que representa a sua força. Assim

**Somatório ($\sum$)** Trata-se de um somador para os sinais de entrada $X_j$, ponderados pelos pesos sinápticos $w_{kj}$ que ativam o neurônio: $\sum_{j=1}^{m} w_{kj}X_j$, onde o primeiro índice $k$ se refere ao neurônio em questão e o segundo índice $j$ se refere à entrada à qual o peso se refere. Em outras palavras, esse elemento é responsável por integrar os sinais que chegam ao neurônio.

Na figura 3.2 é possível visualizar um *bias*, representado por $b_k$, que tem o efeito de aumentar ou diminuir o valor do somatório, gerando o que se denomina valor de ativação: $v_k = \sum_{j=1}^{m} w_{kj}X_j + b_k$

**Função de Ativação** Para restringir a amplitude de saída de um neurônio é preciso aplicar ao valor de ativação uma função que restrinja o intervalo permitido. Usualmente essa função restringe para um intervalo unitário fechado $[0, 1]$ ou $[-1, 1]$. Tal função será melhor detalhada na seção 3.3.2.

### 3.3.2 Tipos de Função de Ativação

A função de ativação, $\varphi(v)$, calcula a saída de um neurônio em função do valor de ativação $v$. Os tipos mais simples de função de ativação são:

**Função de limiar:** A saída de um neurônio assume o valor 1, se o valor de ativação é não negativo e 0 caso contrário, descrevendo a propriedade *tudo-ou-nada*. A figura 3.3.a ilustra o comportamento da função de limiar.

$$\varphi(v) = \begin{cases} 1 \text{ se } v \geq 0 \\ 0 \text{ se } v < 0 \end{cases}$$

**Função sigmóide:** Esta função, ao contrário da função limiar, pode assumir todos os valores entre 0 e 1. A representação mais utilizada para esta função é a função logística, definida por:

$$\varphi(v) = \frac{1}{1 + \exp(-\alpha v)}$$

Onde $\alpha$ é o parâmetro de inclinação da função sigmóide, quando $\alpha \to \infty$, esta função se comporta como a função limiar, como pode ser visto na representação da figura 3.3.b

**Tangente hiperbólica:** As funções de ativação definidas anteriormente possuem o intervalo $[0, 1]$. Entretanto, algumas vezes é desejável que a função de ativação exista no intervalo $[-1, 1]$. A função tangente hiperbólica possui a forma correspondente de uma função sigmóide definida no intervalo -$[1, 1]$ (figura 3.3.c)

$$\varphi(v) = \tanh(v)$$

### 3.3.3 Modelo do Perceptron

O perceptron é um modelo proposto por Rosemblatt (Rosenblatt 1962) composto por um único neurônio, como mostrado na figura 3.2, com função limiar de ativação e aprendizado supervisionado. O objetivo do perceptron é classificar corretamente os dados de entrada $x_1, x_2, \cdots, x_n$ em duas classes $C_1$ e $C_2$. A regra de decisão para a classificação é atribuir à entrada a classe $C_1$ se a saída do perceptron for 1 e a classe $C_2$ se for 0.

Aparentemente, o modelo do perceptron atenderia as necessidades de prever se uma dada função pode ou não ser exercida por uma proteína. Por exemplo, dada uma proteína $P_1$, poderia ser criado um classificador usando o modelo do perceptron com um neurônio para retornar 1 se $P_1$ exerce uma dada função molecular e 0 caso contrário.

O modelo do perceptron, entretanto, é simples demais e seu poder de classificação é bastante limitado. Na verdade, ele só consegue lidar com a classifi-

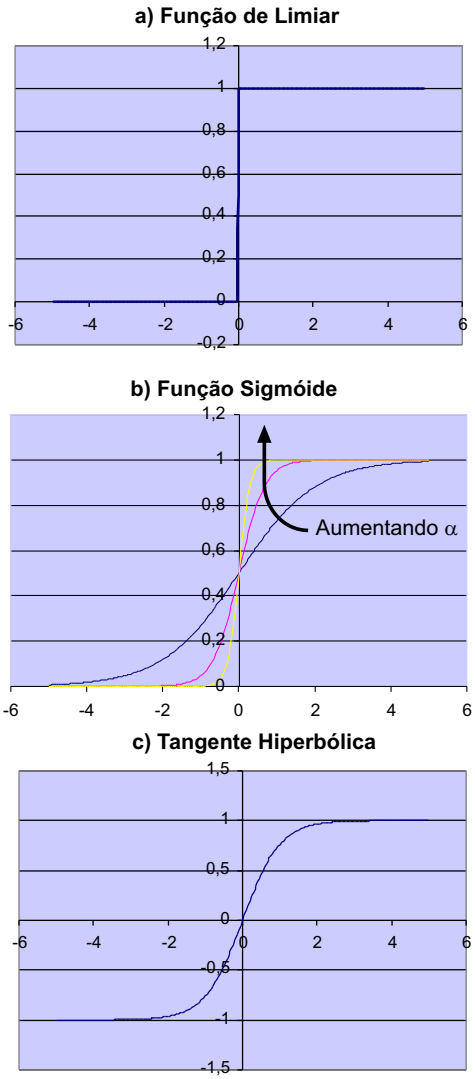


Figura 3.3: Funções de Ativação

cação de conjuntos linearmente separáveis, onde a fronteira de decisão entre uma classe e outra toma a forma de uma reta.

Assim, tomando o exemplo mais simples de um perceptron de duas dimensões e tomando o ponto $(x, y)$, se este ponto se encontra acima da linha de fronteira é atribuído à classe $C_1$, mas se estiver abaixo é atribuído à classe $C_2$, como mostra a figura 3.4.

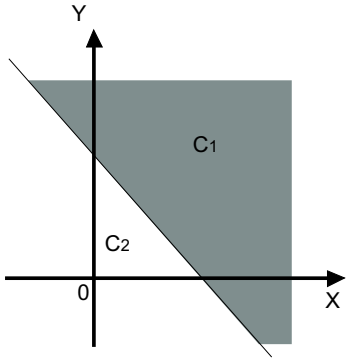


Figura 3.4: Fronteira de Decisão de um Perceptron Simples

Entretanto, o problema de predição da função de proteínas é muito mais complexo do que um espaço de decisão linearmente separável. Dessa forma, utilizar um único perceptron não seria apenas inadequado, mas também incorreto. Uma boa abordagem seria utilizar um conjunto maior de perceptrons organizados de modo a permitir um espaço de decisão mais complexo (seção 3.3.4).

#### 3.3.3.1 Aprendizagem no Modelo Perceptron

Em uma rede neural, o erro pode ser entendido simplesmente como a diferença entre a saída real gerada pela rede e a saída desejada. considere-se, agora, o caso do modelo perceptron, que possui apenas um neurônio $k$, que é o único constituinte da camada de saída de uma rede neural. Tal neurônio é ativado por um vetor de sinais de entrada $x(n)$ (veja figura 3.2), onde $n$ representa um contador de tempo de um processo iterativo envolvido nos ajustes de pesos sinápticos do neurônio $k$. A saída de $k$ é representada por $y_k(n)$ e é comparada a uma resposta desejada $d_k(n)$. Assim, define-se o sinal de erro $e_k(n)$ pela equação 3.1.

$$e_k(n) = d_k(n) - y_k(n) \tag{3.1}$$

Supondo que $w_{kj}(n)$ represente o valor do peso sináptico $w_{kj}$ do neurônio $k$ que recebeu como entrada o elemento $x_j(n)$ do vetor de sinal $\overrightarrow{X}(n)$ no instante

de tempo discreto $n$, é possível utilizar uma regra conhecida como regra delta, o ajuste a ser aplicado ao peso $w_{kj}(n)$ é definido pela equação 3.2

$$\Delta w_{kj}(n) = \eta e_k(n) x_j(n) \tag{3.2}$$

Onde o parâmetro $\eta$, conhecido como taxa de aprendizagem, é uma constante positiva que determina a taxa de aprendizado de um passo a outro no processo de aprendizagem.

De posse do valor de $\Delta w_{kj}(n)$ é possível ajustar o peso sináptico em questão usando a equação 3.3 de modo a torná-lo mais próximo do resultado desejado.

$$w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n) \tag{3.3}$$

O modelo de redes neurais a ser explanado na próxima seção é uma extensão do modelo do perceptron de modo a permitir a utilização de vários neurônios. Dessa forma, apesar de o modelo perceptron não ser usado no presente trabalho, entender a maneira como o aprendizado ocorre é essencial para se obter uma noção intuitiva do aprendizado em outros modelos.

### 3.3.4 Perceptrons de Múltiplas Camadas

Quando a estrutura de uma rede neural possui vários neurônios organizados em camadas, conforme a figura 3.1, gera-se uma arquitetura de rede conhecida como perceptrons de múltiplas camadas (*multilayer perceptron*) e se cada neurônio de uma camada se conecta com todos os neurônios da camada seguinte diz-se que a rede é totalmente conectada.

A modelagem de uma rede neural do tipo perceptron de múltiplas camadas requer algumas escolhas que interferem sensivelmente no seu comportamento. Alguns dos aspectos mais importantes são descritos abaixo.

1. **Projeto da Camada de Entrada:** A escolha da camada de entrada, mais especificamente a escolha do número de nós, é uma importante tarefa a ser cumprida, pois causa impacto tanto na precisão do sistema, quanto no tempo necessário para se realizar o treinamento.

2. **Quantidade de Camadas Ocultas:** As camadas ocultas funcionam como detectores de características que serão representadas internamente como pesos sinápticos, aumentando o número de camadas ocultas de uma para duas faz com que a rede neural consiga aproximar um número superior de funções matemáticas e aprender tarefas mais complexas extraindo progressivamente as características mais significantes dos veto-

res de entrada. Entretanto, não existem métodos para se descobrir o número de camadas ideal, o que obriga o projetista a realizar vários testes com diferentes arquiteturas.

3. **Número de Neurônios por Camada:** Para se escolher o número de perceptrons em cada camada alguns fatores precisam ser levados em conta. O primeiro deles é que, ao escolher um número muito grande de unidades (maior que o necessário) pode-se fazer com que a rede memorize os dados de treinamento e tenha uma boa precisão apenas quando esses dados são fornecidos e se, por outro lado, dados novos forem fornecidos a rede não conseguirá reconhecê-los. Esse problema é comumente conhecido como *overfitting*.

   O segundo fator é que um número muito pequeno de unidades (inferior ao necessário) faz com que a rede demore demais para criar regras de classificação dos padrões, vindo, algumas vezes, a não conseguir esse objetivo. Como no caso da quantidade de camadas ocultas, não é tão simples escolher o número de neurônios e alguns experimentos devem ser feitos antes da escolha final, apesar de existirem na literatura algumas heurísticas que ajudam a solucionar esse problema como o procedimento de decaimento de pesos (Hinton 1989), o de eliminação de pesos (Weigende et al. 1991) e o suavizador aproximativo (Moody & Rögnvaldsson 1997).

4. **Definição da Função de Ativação:** O modelo de cada neurônio da rede inclui a escolha de função de ativação. Deve-se levar em conta que funções como a limiar não são recomendadas, pois diminuem muito a flexibilidade da rede. Além disso, é importante que a função escolhida possua certas características como, por exemplo, seja contínua como a sigmóide. Em vários casos, uma boa escolha é a função logística com a particularidade de $\alpha = 1$.

#### 3.3.4.1 Aprendizagem no Modelo Perceptrons de Múltiplas Camadas

A maneira mais conhecida de se treinar uma rede neural do tipo perceptrons de múltiplas camadas é usando o algoritmo de retropropagação (do inglês *backpropagation*) (Rumelhart et al. 1986), que é um algoritmo de aprendizagem supervisionada por correção de erro, do mesmo modo que o algoritmo visto na seção 3.3.3.1.

O algoritmo de retropropagação fornece uma solução eficiente para o treinamento dos pesos sinápticos, apesar de não oferecer sempre uma solução ótima para todos os problemas resolúveis. De uma maneira geral o algoritmo

segue uma heurística que será explicada logo a seguir. Entretanto, é preciso entender que esse algoritmo traz uma base formal que não será mostrada, mas, em todo caso, recomenda-se Haykin (2001) para uma leitura mais aprofundada.

Considere-se uma rede neural do tipo perceptrons de múltiplas camadas composta de vários neurônios. Nessa rede há um neurônio $k$ que é ativado pela saída $Y_j$ dos neurônios $j$ da camada à esquerda. Nesse caso, o valor de ativação do neurônio $k$ pode ser obtido por $v_k = \sum_{i=1}^{m} w_{kj} Y_j + b_k$, onde $m$ representa o número de neurônios da camada $j$. Tal ativação ocorre no sentido de propagação do sinal mostrada na figura 3.1.

Na iteração $n$, para calcular o erro $e_k$ na saída $y_k$ do neurônio $k$, utiliza-se a equação 3.4, que é análoga à vista na seção 3.3.3.1. Nesta equação, o termo $d_k$ representa o valor desejado.

$$e_k(n) = d_k(n) - y_k(n) \tag{3.4}$$

O cálculo do ajuste nos pesos sinápticos do neurônio $k$ difere um pouco do que foi visto no modelo perceptron, pois o ajuste depende da camada onde o neurônio está inserido. Assim, insere-se o termo $\delta_k(n)$ no cálculo da variação a ser aplicado no peso sináptico de entrada $w_{kj}$, como mostra a equação

$$\Delta w_{kj}(n) = \eta \delta_k(n) y_j(n) \tag{3.5}$$

O termo $\delta_k(n)$ varia conforme a camada no neurônio, se o mesmo estiver na camada de saída usa-se a equação 3.6.

$$\delta_k(n) = e_k(n) \varphi'_k(v_k(n)) \tag{3.6}$$

Onde $\varphi'_j(v_j(n))$ é a derivada da função de ativação na saída do neurônio $k$, daí a importância de a função de ativação possuir como propriedade a continuidade. Caso o neurônio $k$ não esteja na camada de saída utiliza-se a equação 3.7.

$$\delta_k(n) = \varphi'_k(v_k(n)) \sum_i \delta_i(n) w_{ik}(n) \tag{3.7}$$

Onde o segundo fator da equação, o somatório sobre $i$, requer conhecimento de todos os termos $\delta_i(n)$ dos neurônios $i$ que estão na camada imediatamente à direita do neurônio oculto $k$ (seguindo a representação da figura 3.1) e que a este estão conectados. $w_{ik}(n)$, outro termo que aparece nesse somatório, representa os pesos sinápticos associados com as conexões entre os

neurônios $i$ (à direita) e o neurônio $k$.

Pelo exposto acima, é possível entender o porquê do algoritmo ser conhecido como retropropagação, pois, para que o erro em determinado neurônio oculto seja calculado, é preciso que se conheça os erros nas camadas à direita. Assim, calcula-se primeiro os erros na camada saída e retropropaga-se esse erro para a última camada oculta, que por sua vez, repete o procedimento com a camada oculta anterior até chegar à primeira camada oculta.

O algoritmo de retropropagação do modo como foi mostrado não é o único que pode ser utilizado para treinamento, existem várias heurísticas capazes de melhorar o seu desempenho. Além disso, como a função de ativação interfere no algoritmo, alterá-la também pode trazer ganhos de desempenho.

### 3.3.5 Máquinas de Vetor de Suporte

As máquinas de vetor de suporte (SVM, do inglês *Support Vector Machine*) são modelos baseados no fato de que, em altas dimensões do espaço de características, todos os problemas se tornam linearmente separáveis (Haykin 2001).

Um exemplo simples que ilustra esse conceito é mostrado na figura 3.5. Com os dados seguindo uma representação bidimensional (figura 3.5.a) o problema é não linearmente separável, mas se os dados forem representados em um espaço tridimensional (figura 3.5.b) é possível encontrar um hiperplano de separação.

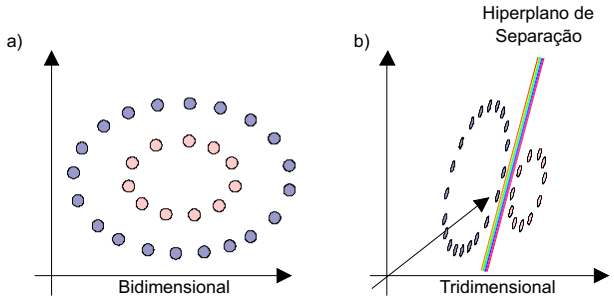


Figura 3.5: Projeção em um espaço de maior dimensão

Dessa forma, a primeira particularidade das SVM é que projetam o espaço real em um espaço de características de maior dimensão, na qual os mesmos sejam linearmente separáveis. Tal projeção é feita através da função $\phi$, que caracteriza o kernel da SVM, dado por: $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. Em geral, uma máquina de vetor de suporte requer a solução do seguinte problema de otimização.

$$\min_{W,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i \tag{3.8}$$

Sujeito às seguintes restrições:

$$\begin{cases} y_i(W^T \phi(X_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{3.9}$$

onde $C > 0$ é o parâmetro que penaliza o erro $\xi$ para cada instância de treinamento $i$.

Apesar de novos *kernels* serem continuamente propostos por pesquisadores, os mais consolidados são os quatro mostrados a seguir.

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polinomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$, com $\gamma > 0$
- Função de Base Radial: $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2)$,com $\gamma > 0$
- Sigmóide: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

As variáveis $\gamma$, $r$ e $d$ são parâmetros dos *kernels*, para uma informação mais aprofundada recomenda-se Haykin (2001). O *kernel* escolhido para ser utilizado nesse trabalho foi a função de base radial, sendo assim, a seção 3.3.5.1 foi destacada para explicar o porque da escolha.

Do que já foi exposto é fácil concluir que SVM são classificadores binários, ou seja, dividem o espaço de decisão em duas classes apenas. Assim, caso o domínio de um dado problema possua várias classes será necessário utilizar vários desses classificadores binários organizados de alguma maneira. Uma das maneiras mais comuns é utilizar vários classificadores par-a-par e organizá-los em uma árvore, conforme a figura 3.6.

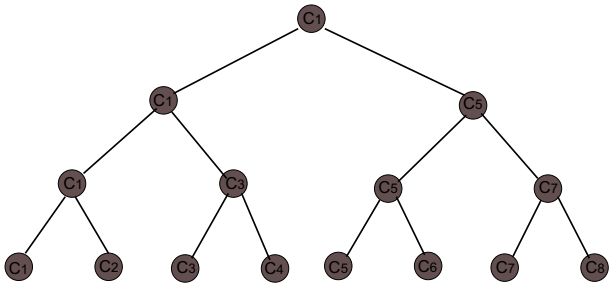


Figura 3.6: Projeção em um espaço de maior dimensão

Dessa forma, se existem $n$ classes diferentes $C_1, C_2, \ldots, C_n$, será preciso um classificador binário que decida entre $C_1$ e $C_2$, outro classificador que decida

entre $C_1$ e $C_3$ e assim sucessivamente, de modo que cada classe possua um classificador com todas as outras. A organização em árvores é feita para que cada classe possa competir com um subconjunto das outras apenas, e a decisão final do sistema será aquela classe que alcançar a raiz da árvore.

Uma outra maneira de se utilizar SVM com várias classes é criando vários classificadores locais que respondam sim ou não para uma determinada classe. Dessa forma, o número de classificadores que precisam ser criados é exatamente igual ao número de classes existentes. Para se obter uma decisão final usa-se na maioria dos casos alguma estratégia de recombinação (Mayoraz & Moreira 1996) como, por exemplo, estratégias baseadas em matrizes de código e hierárquicas (Lorena 2006).

#### 3.3.5.1 Função de Base Radial

O *kernel* do tipo função de base radial é um dos menos complicados, com ele é mais fácil obter um resultado relevante com uma modelagem relativamente simples. Tal simplicidade é, em grande parte, resultado da quantidade de parâmetros existentes (apenas dois), que é menor do que nos outros *kernels* normalmente usados.

Os dois parâmetros que existem ao se trabalhar com o *kernel* do tipo função de base radial são $C$, que não caracteriza diretamente o *kernel*, mas é um parâmetro que penaliza os erros, e $\gamma$. Ambos os parâmetros devem ser estimados conforme os dados de entrada. Assim, é preciso fazer uma pesquisa para se saber quais valores são mais adequados. Tal pesquisa é, em geral, computacionalmente onerosa, pois vários pares $(C, \gamma)$ devem ser tentados.

Como o objetivo é encontrar um bom par $C, \gamma$ que possa fazer inferências em dados desconhecidos, não se deve utilizar os dados de treinamento para verificar se o par $C, \gamma$ é adequado ou não. A maneira mais comum seria, então, separar os dados disponíveis em dados de teste e dados de treinamento e, dessa forma, a acurácia da classificação refletiria o comportamento do sistema ao serem apresentados dados novos.

Em geral, é preciso que haja um compromisso entre os dados de treino e os dados de teste, pois quanto maior o conjunto de dados de treino, melhor o classificador e quanto maior o conjunto de dados de testes, mais confiável será a estimativa do erro.

Uma versão melhorada do método de dividir os dados disponíveis em dados de treino e de validação é a técnica de validação cruzada (Stone 1974), que é o método mais utilizado na literatura, pois as estimativas obtidas são as que mais se aproximam da realidade. A figura 3.7 ilustra o procedimento utilizado

pelo método de validação cruzada, que funciona da seguinte maneira:

1. Inicialmente os dados disponíveis são divididos em $k$ subconjuntos disjuntos e de tamanhos aproximadamente iguais. O valor de $k$ pode variar livremente, mas $k = 5$ ou $k = 10$ são geralmente usados. A figura 3.7 utiliza $k = 10$.

2. Após isso se inicia um processo iterativo, onde em cada rodada um subconjunto é utilizado para teste enquanto os outros são utilizados para treinamento. Dessa forma, a acurácia obtida nos dados de validação é uma medida de quanto o par $(C, \gamma)$ é adequado.

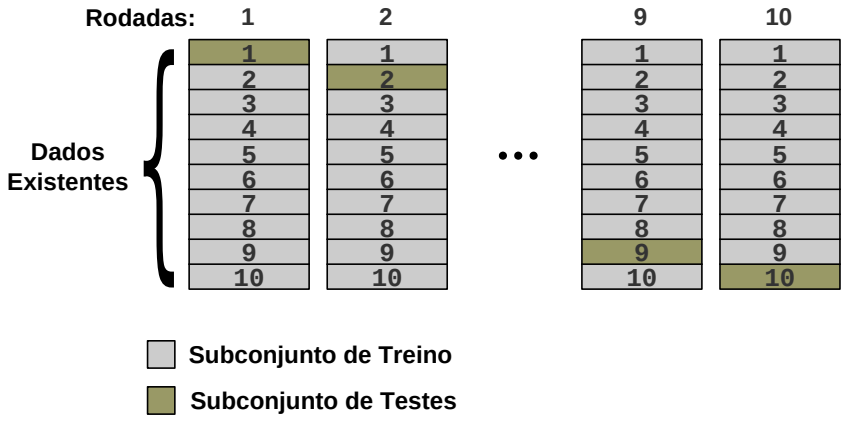


Figura 3.7: Método de Validação Cruzada

A vantagem de se utilizar o método de validação cruzado é que ele acusa quando a abordagem de aprendizagem de máquina apenas "decora"os dados de treinamento (*overfitting*), pois a técnica simula o processo de predição em dados desconhecidos.

A busca utilizada no presente trabalho foi feita utilizando seqüências exponencialmente crescentes de pares $(C, \gamma)$ com a acurácia sendo medida através do método de validação cruzada com $k = 5$. No presente trabalho, a busca foi feita com a variação ocorrendo da seguinte maneira:

- $C = 2^{-5}, 2^{-3}, ..., 2^{15}$
- $\gamma = 2^{-15}, 2^{-13}, 2^{3}$

Outras heurísticas para melhorar o desempenho da busca existem, mas a busca exaustiva é a que, naturalmente, obtém os resultados mais satisfatórios.

A ferramenta utilizada para realizar computacionalmente a escolha dos parâmetros $C$ e $\gamma$, bem como a geração do modelo final a ser validado no capítulo 5 foi a libSVM (Chang & Lin 2001), que é uma plataforma especializada na geração de modelos que usem máquinas de vetor de suporte.

#### 3.3.5.2 Aprendizagem em Máquinas de Vetor de Suporte

A aprendizagem por máquina de vetor de suporte constrói um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja a maior possível. Tal hiperplano é conhecido como hiperplano ótimo (figura 3.8).

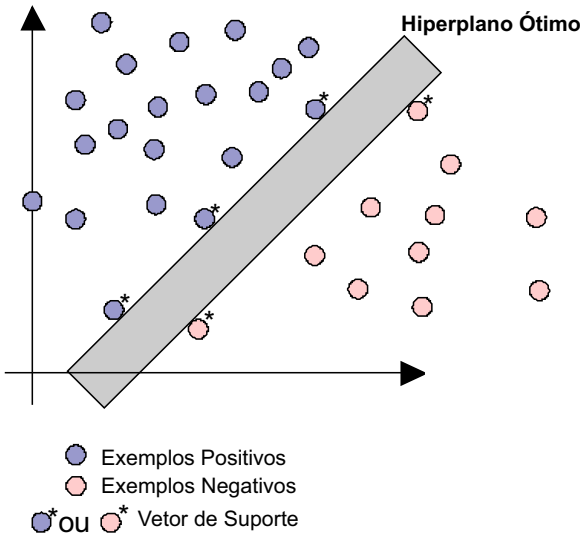


Figura 3.8: Redes Neurais

Os vetores de suporte consistem de um pequeno subconjunto dos dados de treinamento extraído pelo algoritmo. A característica principal é o fato de estarem na borda do hiperplano ótimo, sendo portanto os mais difíceis de classificar. Dessa forma, os vetores de suporte têm uma influência direta na localização ótima da superfície de decisão.

Define-se matematicamente um hiperplano da seguinte maneira: considere-se duas classes linearmente separáveis $(\overrightarrow{X}_i, y_i)$ onde $x_i \in \mathbb{R}^{\kappa}$ e $y_i \in \{-1, 1\}$ e dados um vetor de pesos $\overrightarrow{W}$ e um *bias* $b$. Um hiperplano é uma região no espaço da forma:

$$\overrightarrow{W}.\overrightarrow{X} + b = 0$$

Para discriminar entre uma classe e outra basta verificar que $\overrightarrow{W}.\overrightarrow{X} + b > 0$ representa uma das classes e $\overrightarrow{W}.\overrightarrow{X} + b < 0$ representa a outra. Para que o

hiperplano em questão seja ótimo, deve-se encontrar os parâmetros $W_0$ e $b_0$ que maximizem a margem de separação entre as classes.

Maximizar a margem equivale a minimizar a norma $||W||$ (Hearst et al. 1998). Dessa forma, o aprendizado em máquinas de vetor de suporte recai em um problema de otimização dado pela equação 3.8 sujeito a restrições dadas pela equação 3.9

# Capítulo 4

# Modelo Computacional

O presente capítulo pretende utilizar os conceitos abordados na parte referente às noções biológicas (capítulo 2) a fim de definir todos os aspectos protéicos utilizados na confecção do modelo. Além disso, o próprio modelo é definido utilizando a estrutura de uma máquina de vetor de suporte mencionada no capítulo 3.

Por motivo de clareza, dividiu-se o capítulo em cinco seções para que os conceitos fossem colocados de forma seqüencial: a seção 4.1 é um resumo da arquitetura da ontologia conhecida como *Gene Ontology* - GO, que foi utilizada na implementação do modelo apresentado neste trabalho e a seção 4.2 cita um conjunto de funções que podem ser desempenhadas pelas proteínas dentro da célula, esse conjunto é resultado de uma seleção realizada a partir da *gene ontology* e ilustra o escopo de funções que podem ser utilizadas com o presente modelo. Entretanto, para minimizar os custos computacionais, nem todas as funções cobertas pelo sistema foram utilizadas na análise estatística, apenas algumas foram extraídas para estudo de caso.

A seção 4.3, apresenta os atributos computados a partir da estrutura terciária das proteínas e utilizados no presente modelo. Tais atributos foram préselecionados por Borro et al. (2006) com a utilização de medidas estatísticas e métodos de mineração de dados aplicados ao banco de dados STING_DB, (http://www.cbi.cnptia.embrapa.br/StingRDB), o mesmo banco utilizado no presente trabalho.

A seção 4.4, define o projeto de entrada do modelo a partir dos parâmetros mencionados na seção 4.3. É importante ressaltar que a seção 4.4 foi separada da seção 4.3 por esta abordar apenas os aspectos biológicos relacionados a escolha dos parâmetros, enquanto que aquela aborda apenas os aspectos computacionais que tornam o modelo viável.

A última seção (seção 4.5) cria um agregado das outras, mostrando a orga-

nização propriamente dita do modelo.

## 4.1 Gene Ontology

As proteínas são moléculas muito flexíveis e executam funções muito diferentes em um organismo. A utilização de ontologias ajuda a lidar com a complexidade em questão, permitindo a utilização de métodos computacionais.

A escolha da ontologia requer muito cuidado, pois interfere no funcionamento do sistema como um todo, além de padronizar o vocabulário com que este se comunicará com o usuário final e a interface com que será acessível a outros sistemas.

Nesse contexto, é possível classificar as ontologias existentes em não hierárquicas como a GeneQuiz (Andrade et al. 1999) e hierárquicas como a *Gene Ontology* - GO. Apesar de as ontologias não hierárquicas serem capazes de descrever a função geral de uma proteína, uma abordagem hierárquica é útil para descrever termos mais específicos (Eisner et al. 2005).

Em ontologias como a GO, tanto o conhecimento geral quanto o específico sobre uma proteína está representado de uma maneira hierárquica sob a forma de uma estrutura que, devido a hierarquia, lembra uma árvore, mas que na verdade é um grafo direcionado acíclico melhor conhecido pela sigla DAG (Figura 4.1), do inglês *directed acyclic graph*.

Nesse DAG, os nós representam uma categoria (seção 4.1.1) e os arcos direcionados indicam que o nó destino é um caso específico do nó origem. Por essa semântica associada aos arcos é fácil perceber que nesse tipo de estrutura os termos gerais estão próximos da raiz e os termos específicos estão próximos das folhas. A figura 4.1 ilustra bem esse fato.

Por exemplo, se uma proteína é associada ao nó *metal ion binding* e no DAG existe um arco cuja origem é *ion binding* e o destino é *metal ion binding*, então essa proteína também está associada a *ion binding* e a todos os nós $n$ tais que existe um caminho de $n$ a *metal ion binding*. Essa regra é conhecida como *true path rule* e é bastante útil em metodologias que fazem uso de algum tipo de aprendizagem de máquina (Eisner et al. 2005). Sem essa regra a maioria das predições seriam inconsistentes.

É necessário, porém, notar que o fato de determinada proteína estar associada explicitamente a um nó $n$ não significa que esteja excluída a possibilidade de estar também associada aos nós $n_i$ tais que existe um arco $i$ cuja origem é $n$ e o destino é $n_i$, pois o que pode ter acontecido é que testes mais específicos ainda não foram realizados.

Além disso, é importante perceber que uma dada proteína pode possuir vários domínios funcionais, o que faz com que a mesma apresente diversas funções moleculares e seja associada explicitamente a mais de um nó do DAG, não havendo, por conseguinte, a necessidade de existir um caminho entre esses nós.

Assim, tomando como exemplo o caso estudado na seção 2.5.1 sobre a proteína expressa pelo gene *lacI*, que é uma proteína reguladora que se liga a uma região específica do DNA, conclui-se que a mesma está associada ao termo '*DNA binding*', enquanto que a proteína expressa pelo gene *lacY*, responsável por transportar a lactose do meio extracelular para o meio intracelular está associada ao termo '*lactose transporter activity*' (veja figura 4.2).

Um outro fato, além da existência de vários domínios funcionais, que pode fazer com que uma dada proteína seja associada a diversos nós da GO é a existência de termos correlacionados como, por exemplo, o termo '*lactose binding*' também pode ser associado à proteína expressa pelo gene *lacY*. Isso ocorre porque, para que uma proteína seja capaz de transportar a lactose é necessário que ela também seja capaz de se ligar a ela.

### 4.1.1 Categorias

Como pode ser visto nas figuras 4.1 e 4.2 , tanto o termo '*DNA biding*' quanto o termo '*lactose transporter activity*' herdam de '*Molecular Function*'. Isso ocorre porque a classificação utilizada pela GO é dividida em três categorias principais, que são:

**Função Molecular** Define-se como a atividade bioquímica de uma proteína. Essa categoria define apenas o que é realizado sem especificar onde ou quando o evento ocorre na realizade.

**Processo Biológico** Refere-se ao objetivo biológico ao qual o gene ou proteína contribui. Um processo é realizado por um conjunto de funções moleculares e geralmente envolve modificações químicas ou físicas como, por exemplo, manutenção e crescimento celular.

**Localização Celular** Refere-se à localização subcelular de uma proteína como, por exemplo, núcleo ou lisossomos. Apesar de esse item não constituir um aspecto funcional é de suma importância, pois a proteína não desempenha sua função no vácuo ou em solução salina.

Apesar de as três categorias serem aspectos importantes, o presente trabalho, por simplicidade, focou apenas o aspecto de função molecular das proteí-

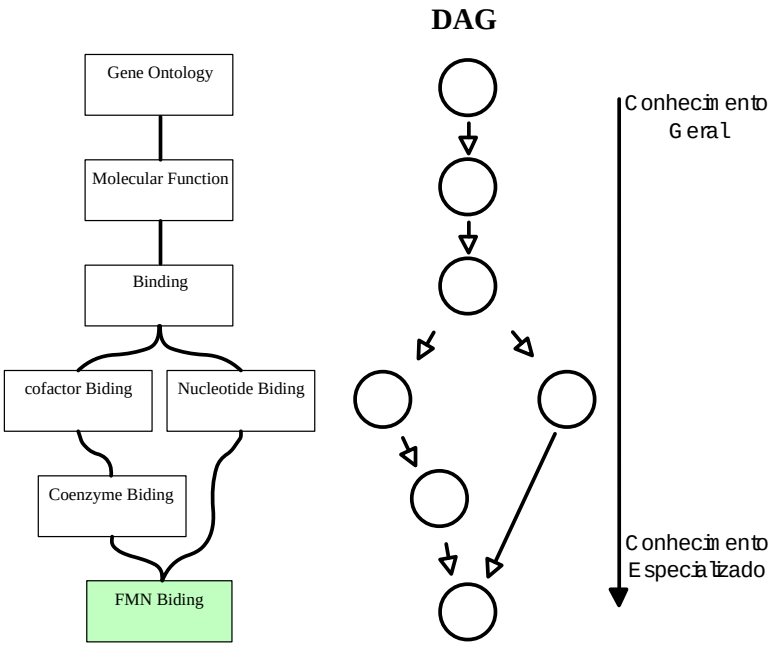


Figura 4.1: Arquitetura em DAG para *FMN Binding*

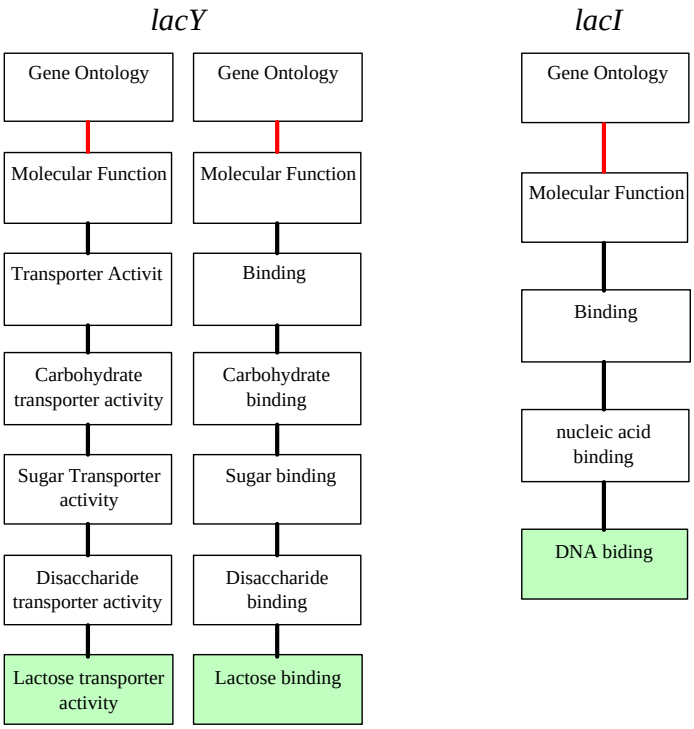


Figura 4.2: Hierarquia da *Gene Ontology*

nas.

## 4.2 Escopo do Modelo

O presente trabalho possui dois objetivos principais. O primeiro deles é verificar se determinada funcionalidade é da alçada de uma dada proteína. Nesse caso, o modelo deverá ser capaz de fornecer uma resposta do tipo "sim" ou "não" para essa pergunta.

O segundo objetivo é, dada a estrutura de uma proteína, o modelo deve sugerir um conjunto de prováveis funções que essa proteína exerce em nível molecular dentro da célula. Apesar dos dois objetivos serem parecidos, a abordagem exigida pelo segundo traz algumas complicações que precisam ser superadas.

De um modo geral, ambos os objetivos necessitam que o escopo de funções que podem ser trabalhadas pelo modelo seja delimitado. Como a GO é uma biblioteca dinâmica, que recebe atualizações mensalmente, o alcance do presente modelo será também dinâmico. Além disso, a metodologia exige um certo número de proteínas com estrutura tridimensional decifrada associadas para que haja uma boa predição. Em vários casos isso não ocorre, pois existem termos com apenas duas ou três proteínas associadas.

Assim, é necessária uma heurística para delimitar um subgrupo de funções que atenda a esses requisitos e, como regra, selecionou-se apenas as funções com mais de 50 conformações espaciais associadas a elas e que não ultrapassam o quarto nível de profundidade na categoria 'molecular function'. As subseções seguintes apenas ilustram a cobertura delimitada a partir destas duas regras.

### 4.2.1 Proteínas de Ligação

A habilidade de se acoplar a outras moléculas é uma das características principais das proteínas e é responsável, como visto na seção 2.4, por permitir que a proteína atue. A especificidade dos sítios de ligação estudada na seção 2.4.1 permite classificar uma proteína de acordo com o ligante. A figura 4.3 mostra algumas funções de ligação selecionadas de acordo com as regras já citadas. Nessa mesma figura é possível visualizar em um fundo cinza as funções que foram selecionadas como estudo de caso para a análise estatística realizada no capítulo 5.

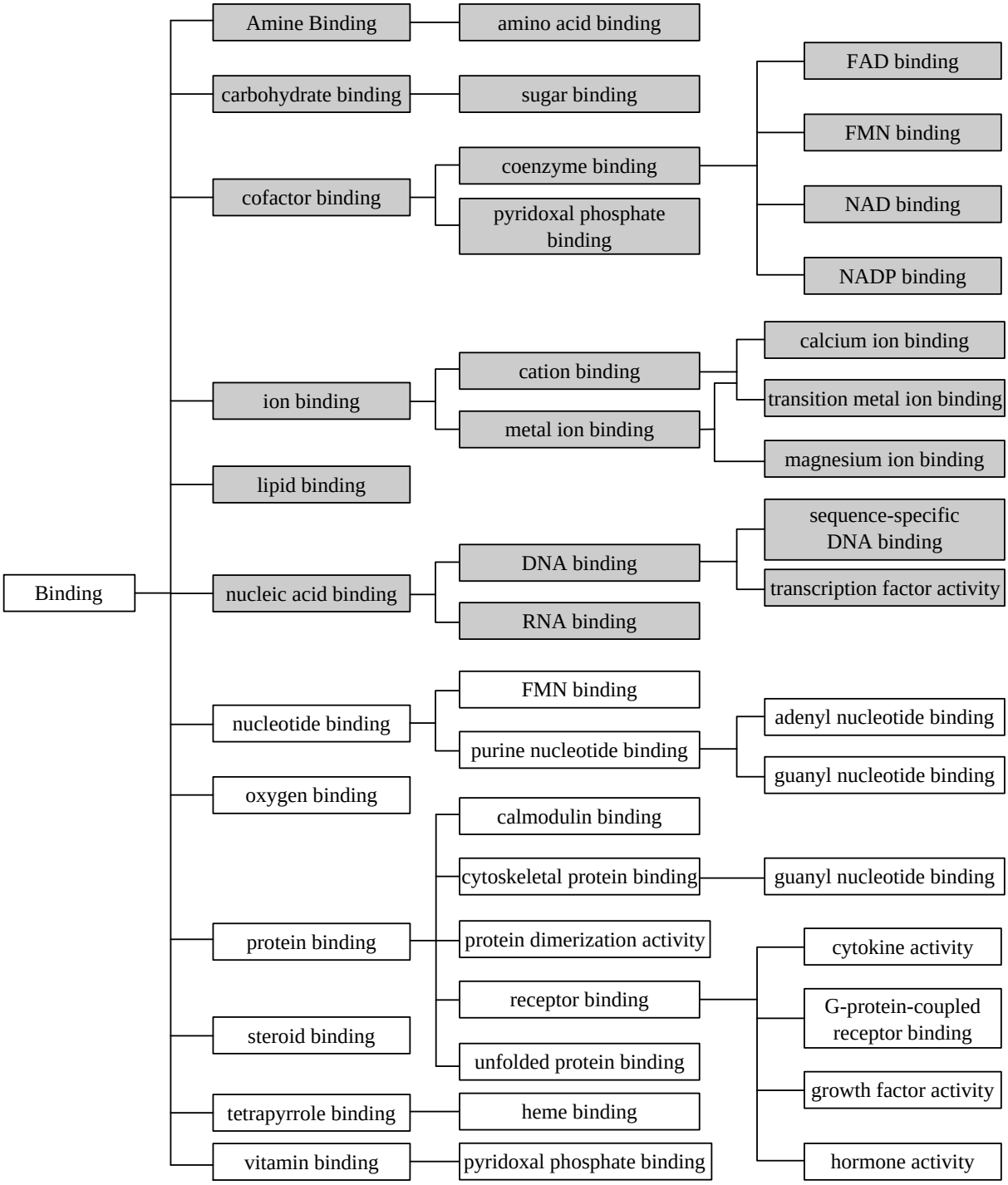


Figura 4.3: Funções de Ligação

### 4.2.2 Proteínas com Função Enzimática

A importância das atividades enzimáticas na célula foi explicada na seção 2.5.3, que abordou uma série de propriedades gerais necessárias para que o processo catalítico tenha êxito. Vale ressaltar apenas uma particularidade do processo de classificação das enzimas, que ocorre de acordo com o substrato envolvido ou, quando a enzima atua sobre mais de um substrato, de acordo com a reação bioquímica envolvida.

Dessa forma, embora alguns termos da figura 4.4 como *'peptidase activity'* pareçam descrever uma única entidade catalítica, eles englobam todas as proteínas que catalizam a hidrólise de ligações peptídicas.

Outro ponto que pode ser observado da figura 4.4 é quanto ao nome das funções enzimáticas mostradas. Percebe-se que algumas seguem um padrão de acrescentar o sufixo *-ase* ao nome do substrato específico sobre o qual agem. Assim, proteínas que hidrolizam pontes peptídicas são chamadas de *peptidase* e as que atuam sobre ácidos nucléicos são chamadas de *nuclease*.

A complexidade das reações enzimáticas impede que a figura 4.4 seja colocada de forma completa. Dessa forma, o gráfico em dendograma precisou ser "podado" nas atividades de *oxidoreductase* e *transferase* e os arcos que saem delas não são mostrados, apesar de participarem da cobertura do modelo.

### 4.2.3 Outras Funções

Outras funções que as proteínas podem exercer foram colocadas na figura 4.5. Como essas atividades são bem menos comuns em proteínas que as enzimáticas, existem menos proteínas com conformação espacial decifrada associadas a elas. A primeira vista isso parece um problema a ser enfrentado pelo modelo, mas, por outro lado, o número de proteínas decifradas a cada ano cresce gradualmente.

## 4.3 Características Extraídas das Proteínas

A presente seção trata sobre os parâmetros que participaram do modelo de aprendizagem de máquina. Entende-se por parâmetro um aspecto extraído da estrutura espacial da proteína e que representa alguma propriedade estrutural, física ou química capaz de funcionar como discriminante de sua função, tais como hidrofobicidade, energias de contato e outras (figura 4.6).

Os parâmetros utilizados no presente trabalho foram selecionados por Borro et al. (2006) utilizando métricas estatísticas e técnicas de mineração de dados

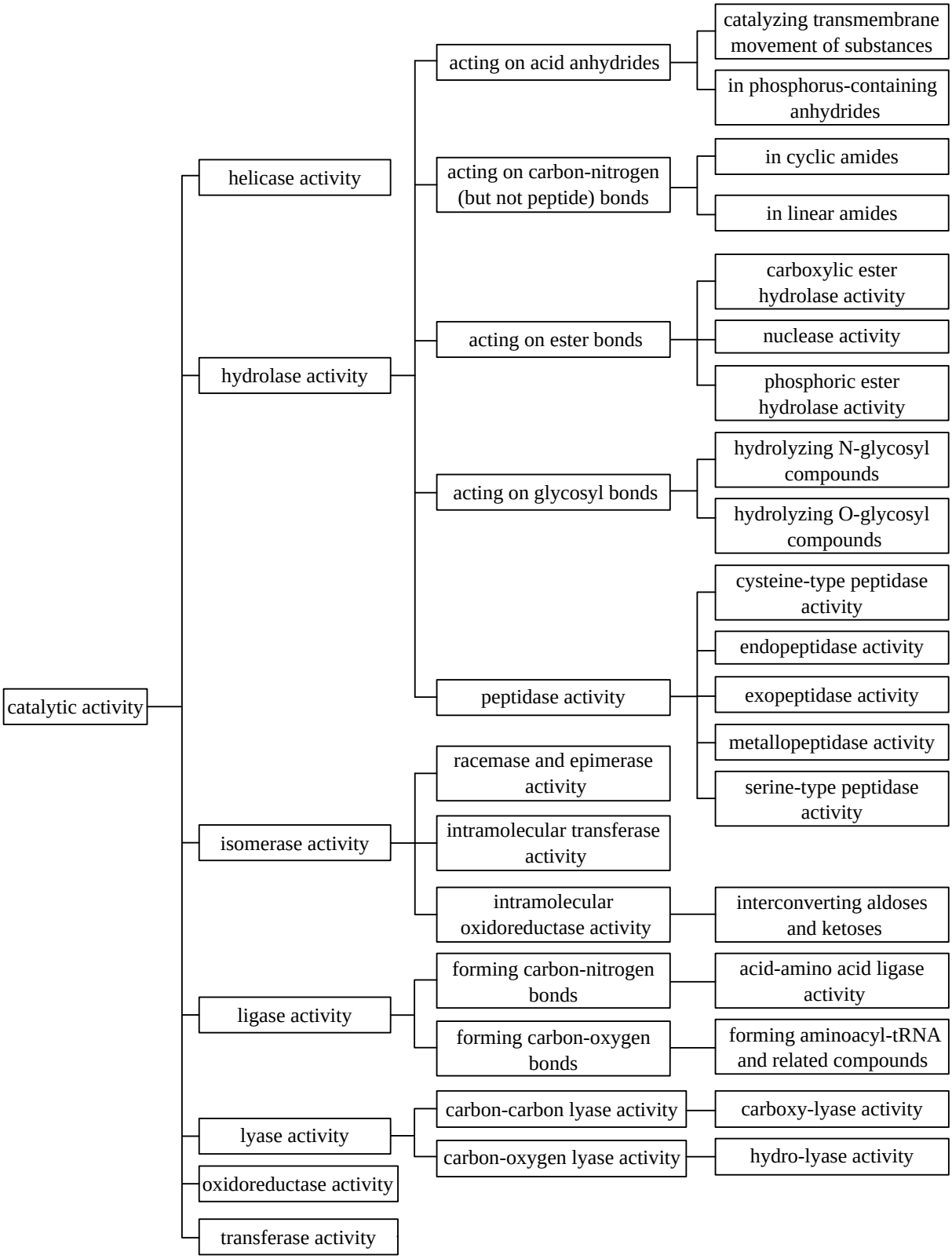


Figura 4.4: Funções Enzimáticas

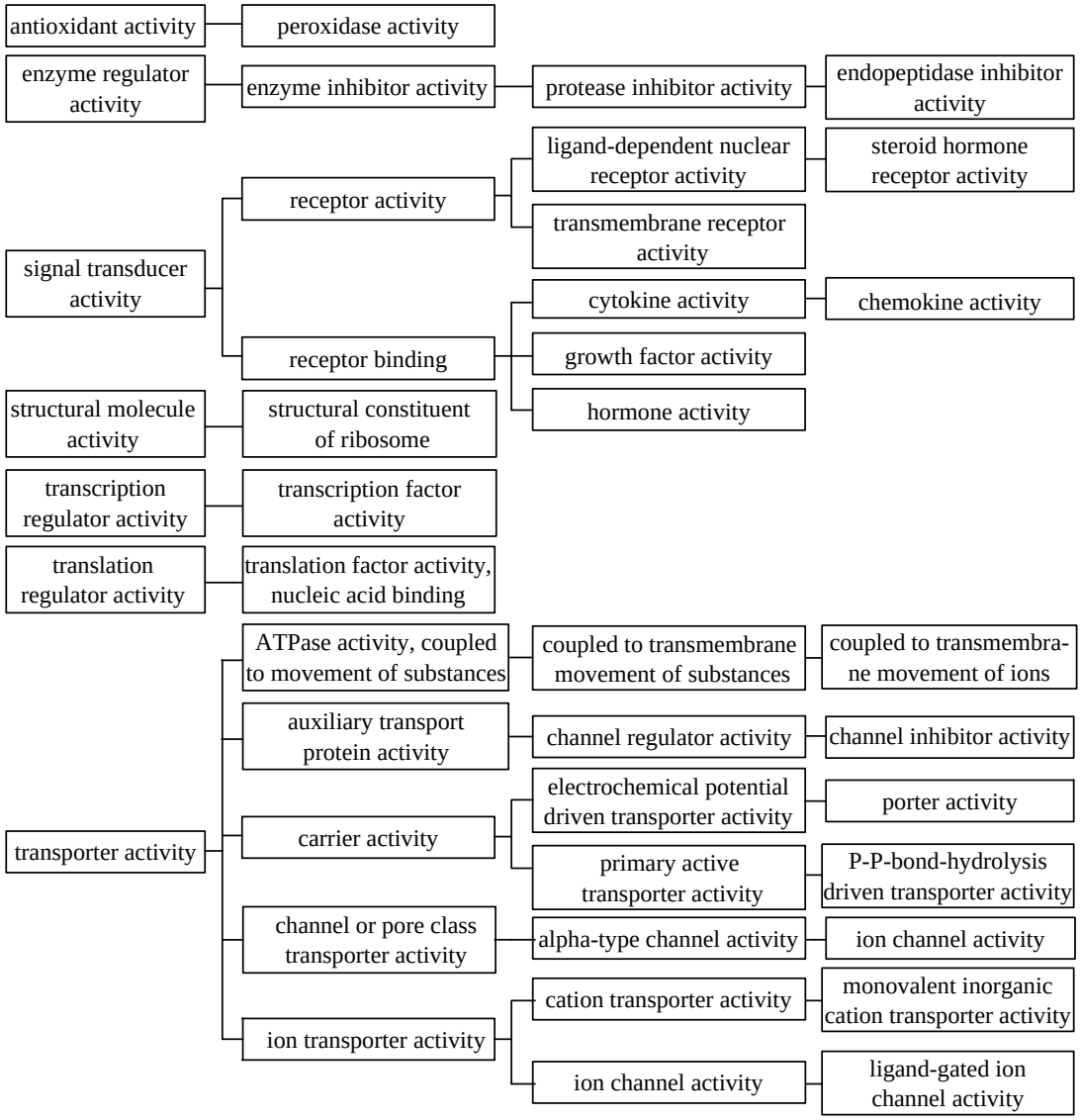


Figura 4.5: Outras Funções Abordadas

aplicadas ao banco de dados STING_DB. A vantagem de se utilizar esse subconjunto de parâmetros é que os mesmos são representativos do banco de dados. Entretanto, não é possível garantir que outros subconjuntos não seriam mais indicados.

O método de Borro et al. (2006) se baseou fortemente na identificação e remoção de informações redundantes presentes na base de dados. Para isso, o coeficiente de correlação foi a principal medida estatística utilizada para excluir da seleção de parâmetros uma série de atributos correlacionados, já que dois parâmetros altamente correlacionados ($\rho > 0.8$) possuem informações redundantes, podendo um deles ser excluído do modelo.

Existem várias outras abordagens que poderiam ser utilizadas na seleção de parâmetros, sendo esse campo já bastante consolidado na literatura pelo nome de *feature subset selection* - FSS. Em Larrañaga et al. (2006) é possível encontrar uma explicação mais detalhada sobre esse tema aplicado à área de bioinformática.

O diagrama da figura 4.6 mostra os parâmetros selecionados mantendo a estrutura dos relacionamentos de algumas das tabelas do banco de dados STING_DB. Percebe-se de início que a figura mostra um diagrama simples, que não segue nenhum tipo de convenção conhecida em banco de dados e exclui informações como chaves primárias e estrangeiras. Entretanto, para o propósito da explicação a seguir, esse diagrama supre todas as necessidades.

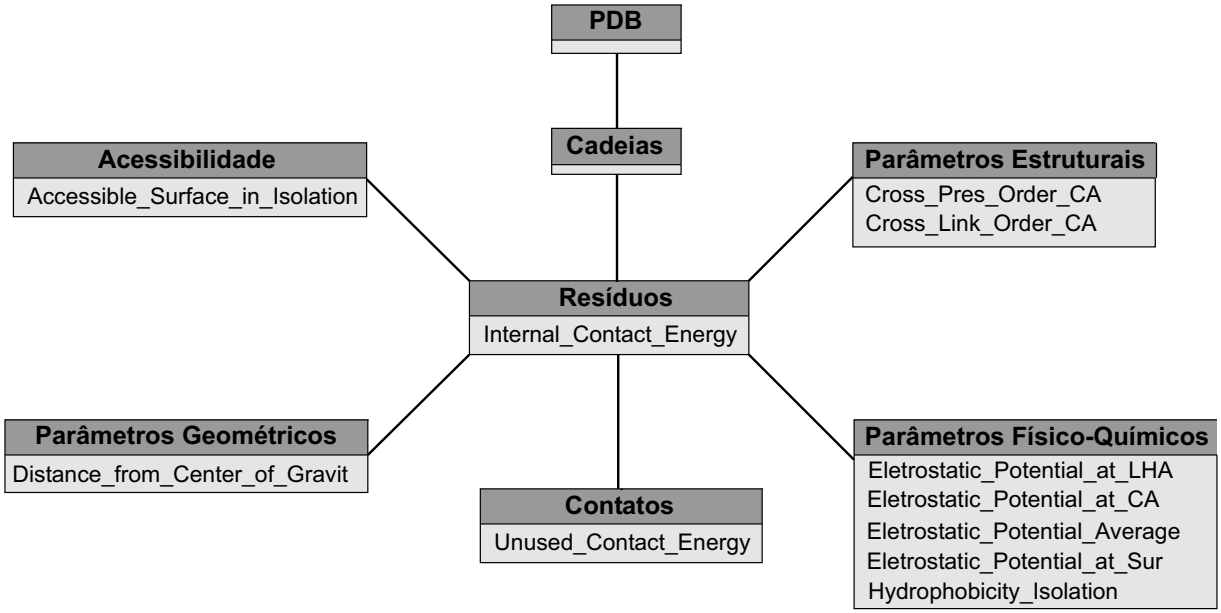


Figura 4.6: Parâmetros

O STING_DB é um banco de dados que opera como uma coleção de dados retirados de bancos de dados públicos como, por exemplo, PDB (Berman

et al. 2002) e HSSP (Dodge et al. 1998), e dados próprios como os que são explicitados na figura 4.6. O interesse maior para a dissertação são os parâmetros próprios do STING_DB e sua relação com os dados armazenados no PDB, pois, é por intermédio deste último que foram obtidas as associações com as funções da GO.

Assim, a tabela **PDB** da figura 4.6 é de suma importância, pois é uma ligação que se tem entre as funções da GO e os parâmetros estruturais do STING_DB e ajudará a criar o vetor de entrada do modelo de aprendizagem de máquina que será visto na próxima seção.

A tabela **PDB** da figura se liga diretamente com a tabela **Cadeias**, já que, as proteínas que apresentam estrutura quaternária (seção 2.3.1) são constituídas por duas ou mais cadeias polipeptídicas unidas por forças diferentes das ligações covalentes (seção 2.1.3). Assim, a relação que ocorre entre as tabelas **PDB** e **Cadeias** é de um para muitos.

Um fato que precisa ser ressaltado é que não se pode levar em conta no modelo aquelas funções que estão associadas a uma proteína como um todo e sim apenas as funções que estão associadas a uma certa cadeia. Isso ocorre para evitar atribuir um termo GO a cadeias de PDB onde o termo só seria válido para a proteína inteira correspondente.

A tabela **Cadeias** se liga diretamente com a tabela **Resíduos** em uma relação do tipo um para muitos, já que uma dada cadeia de proteína pode ter vários resíduos. O fato de uma dada cadeia possuir vários resíduos é um problema para as técnicas de redes neurais, pois o tamanho do vetor de entrada precisa ser fixado e, como as proteína possuem tamanhos diferentes, algum pré-processamento precisa ser realizado. Tal pré-processamento será tratado posteriormente na seção 4.4.1.

Para cada resíduo são calculados uma série de parâmetros como, por exemplo, os parâmetros mostrados na figura 4.6. Entretanto, nem todos esses parâmetros se adequaram ao propósito do presente trabalho, pois dois deles ('Cross_Presence_Order_CA' e 'Cross_Link_Order_CA') apresentaram uma quantidade muito grande de valores nulos ou ausentes, algo que compromete o modelo como um todo (veja seção 4.4.1).

Uma explicação simples dos parâmetros, excetuando os dois parâmetros não utilizados no presente modelo, é dada logo abaixo:

1. **Accessible_Surface_in_Isolation:** representa a área acessível a um dado solvente. Nesse parâmetro a área acessível é calculada com a cadeia em questão isolada das demais no caso de proteínas com mais de uma cadeia. O cálculo foi efetuado usando o programa surfV (Sridharan et al.

1992) baseado no algoritmo de (Shrake & Rupley 1973).

2. **Hydrophobicity_Isolation:** representa o grau de hidrofobicidade de cada aminoácido ponderada pela acessibilidade calculada no item anterior. Os valores de hidrofobicidade são mapeados de acordo com a tabela 4.1 publicada inicialmente por Radzicka & Wolfenden (1988).

| Aminoácido | Hidrofobicidade |
|---|---|
| Alanina | $1,81$ |
| Cisteina | $1,28$ |
| Ácido aspártico | $-8,72$ |
| Ácido glutâmico | $-6,81$ |
| Fenilalanina | $2,98$ |
| Glicina | $0.94$ |
| Histidina | $-4.66$ |
| Isoleucina | $4.92$ |
| Lisina | $-5.55$ |
| Leucina | $4.92$ |
| Metionina | $2.35$ |
| Asparagina | $-6.64$ |
| Prolina | $3.50$ |
| Glutamina | $-5.54$ |
| Arginina | $0.92$ |
| Serina | $-3.40$ |
| Treonina | $-2.57$ |
| Valina | $4.04$ |
| Triptofano | $2.33$ |
| Tirosina | $-0.14$ |

Tabela 4.1: Valores de hidrofobicidade por aminoácidos

A equação que calcula a hidrofobicidade para cada aminoácido $i$ é:

$$Hidrofob_i = \frac{Acessib_i}{AcessibMax_i} Radzicka$$

Onde $Acessib_i$ é o parâmetro calculado no item anterior, $AcessibMax_i$ é o valor máximo que pode ser atingido pelo resíduo $i$ e $Radzicka$ é um valor obtido em relação ao resíduo $i$ pela tabela 4.1.

3. **Internal_Contact_Energy:** calculado pela soma das energias dos contatos estabelecidos entre resíduos pertencentes a uma dada cadeia. Basicamente são calculados seis tipos de contatos diferentes mostrados na tabela 4.2.

4. **Distance_from_Center_of_Gravit:** representa a distância entre o carbono $\alpha$ de cada resíduo e o centro de massa da cadeia (baricentro).

5. **Unused_Contact_Energy:** cada resíduo pode realizar um certo número máximo de contatos interatômicos, mas o número de contatos realmente estabelecidos pode ser bem menor. Esse parâmetro é a diferença entre o número máximo de contatos e o número de contatos estabelecidos.

6. **Eletrostatic_Potential_at_CA:** potencial eletrostático pode ser resumido como uma pressão elétrica que quando varia produz um campo capaz de atrair ou repelir partículas eletricamente carregadas. Conhecer o potencial eletrostático é importante, dentre outras coisas, para saber se determinado ligante pode ser acoplado a um sítio de ligação. O parâmetro 'Eletrostatic_Potential_at_CA' representa o potencial eletrostático calculado sobre o carbono $\alpha$ de um dado resíduo.

   Esse parâmetro e os próximos são de grande importância, pois o potencial eletrostático interfere diretamente na estabilidade de um ligação entre a proteína e o seu ligante. Além disso, mostra a afinidade que certas regiões das proteínas possuem em relação a carga elétrica do ligante. O cálculo do potencial elestrostático é realizado pelo STING_DB utilizando o programa Grasp (Nicholls et al. 1991).

7. **Eletrostatic_Potential_at_LHA:** representa o potencial eletrostático calculado sobre o átomo de um resíduo conhecido como LHA (do inglês *last heavy atom*), que é definido como o átomo pertencente à cadeia lateral mais distante do carbono $\alpha$, excluindo-se dessa classificação os átomos de hidrogênio.

8. **Eletrostatic_Potential_at_Sur:** representa o potencial na região da superfície mais próxima do resíduo.

9. **Eletrostatic_Potential_Average:** é a média dos potenciais eletrostáticos calculados sobre todos os átomos do resíduo em questão.

| Tipo de Contato | Energia em Kcal/mol |
|---|---|
| Van der Waals | $0,08$ |
| Interações Hidrofóbicas | $0,6$ |
| Contatos dos anéis aromáticos | $1,5$ |
| Pontes de Hidrogênio | $2,6$ |
| Pontes Salinas | $10,0$ |
| Pontes Dissulfídicas | $85,0$ |

Tabela 4.2: Valores para energias de contato

## 4.4 Representação Vetorial das Proteínas

Esta seção completa o entendimento acerca do modelo criado. A primeira parte trata do processamento dos dados que alimentarão a rede (seção 4.4.1), posteriormente será feita uma análise de quais instância desses dados alimentarão os classificadores locais binários de acordo com a estrutura da GO. Espera-se que ao final desta seção seja possível um entendimento completo do modelo criado.

### 4.4.1 Processamento da Entrada da Rede

Como já foi mencionado na seção 4.3, cada cadeia de proteína possui vários resíduos. Entretanto, a rede neural precisa de um vetor de entrada que seja constante, sendo necessário então abstrair da proteína o maior número possível de informações relevantes em uma representação constante. Dessa forma, inspirado no trabalho de Borro et al. (2006) utilizou-se uma transformação bastante utilizada no campo de processamento de sinais e imagens, a transformação discreta do cosseno (Ahmed et al. 1974), que será apresentada na seção 4.4.1.1.

Uma outra dificuldade encontrada na confecção do modelo é a ocorrência de valores ausentes (*missing values*) para determinado resíduo, um fato que não ocorre com tanta freqüência, mas não pode ser ignorado. A primeira saída seria eliminar todos os parâmetros que possuíssem valores ausentes, entretanto alguns parâmetros com bom poder de discriminação seriam perdidos.

Para evitar perder bons parâmetros, a solução encontrada foi interpolar (Kincaid & Cheney 2002) os parâmetros ausentes usando alguma das técnicas como, por exemplo, interpolação linear, polinomial ou spline. No presente trabalho, a interpolação spline foi utilizada, principalmente, por ser menos onerosa computacionalmente.

Entretanto, a interpolação só é útil, sem comprometer o projeto, se poucos resíduos possuem valores ausentes. Assim, só participaram do modelo as proteínas que possuem menos de 1% dos resíduos com valores ausentes, nessas proteínas a interpolação possui um resultado satisfatório.

Infelizmente, esse limite para *missing values* acarreta uma nova dificuldade, pois dois dos onze parâmetros selecionados por Borro et al. (2006), 'Cross_Presence_Order_CA' e 'Cross_Link_Order_CA', eliminariam em média 30% de todas as proteínas com conformação espacial definida. Esse novo problema surge porque esses dois parâmetros são aqueles em que a ocorrência de valores ausentes é mais acentuada, concluiu-se então que tais parâmetros

são inadequados para o estudo em questão e foram excluídos do modelo.

#### 4.4.1.1 Transformação Discreta do Cosseno

O propósito da transformação discreta do cosseno (DCT) (Ahmed et al. 1974) é transformar uma seqüência de dados em outra, de modo a obter algumas características úteis como fazer com que a parte mais significativa dos dados fique contida em um pequeno número de componentes. A transformação discreta do cosseno possui apenas valores reais, ao contrário de outras abordagens como a transformação por séries de fourier que utiliza no novo domínio valores complexos.

Formalmente uma DCT é uma função $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Assim, os $n$ números reais $x_0, x_1, \cdots, x_n$ são tranformados em $n$ números reais $c_0, c_1, \cdots, c_n$ que são os coeficientes na série transformada. A DCT representa a seqüência de dados $x(n)$ de tamanho $N$ em termos de sua expansão por séries de cosseno com os coeficientes $c_k$ calculados pela equação 4.1

$$c_k = \alpha_k \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \tag{4.1}$$

Onde $\alpha_k = 1/\sqrt{N}$ para $K = 0$ e $\alpha_k = \sqrt{2/N}$ para $k = 1 \cdots N$. Para computar os coeficientes usou-se a direta aplicação das fórmulas, apesar de já existirem algoritmos que realizam a tarefa com uma complexidade de tempo inferior.

Uma característica importante da DCT é o fato de ela ser ortonormal, ou seja, $y = Cx; x = C^{-1}y$. Dessa forma, para obter a função original basta aplicar a sua inversa, dada pela equação 4.2

$$x_n = \sqrt{\frac{2}{n}} \sum_{k=0}^{N-1} \alpha_k c_k \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \tag{4.2}$$

Para exemplificar a utilização da DCT utilizou-se a função $F : \mathbb{R} \rightarrow \mathbb{R}$ dada pela equação 4.3 discretizada com $x$ recebendo apenas valores inteiros no intervalo $[0, 64]$

$$f(x) = \sin\left(\frac{4\pi x}{64}\right) + \cos\left(\frac{6\pi x}{64}\right); \tag{4.3}$$

O gráfico dessa função é mostado na figura 4.7.a e sua transformada usando a DCT é mostrada na figura 4.7.b. É importante saber que os dois gráficos representam a mesma informação em domínios diferentes e, da mesma forma como a transformada fora obtida, é possível obter a função original com uma perda de informação desprezível usando a inversa da DCT.

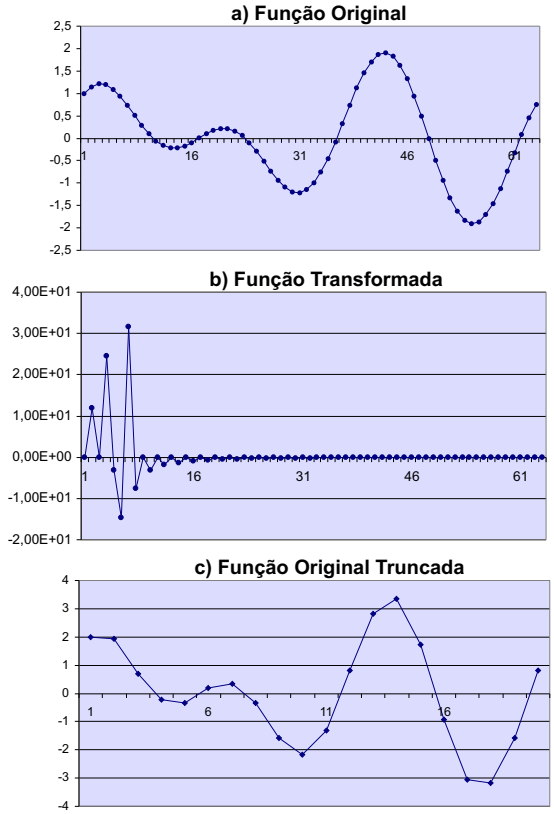


Figura 4.7: Transformação Discreta do Cosseno

Cabe ressaltar que o que motiva a utilização da DCT é que, como ilustra a figura 4.7, grande parte da informação armazenada pela transformada está armazenada nos primeiros coeficientes. Os últimos coeficientes tendem a modelar apenas as pequenas variações que ocorrem na seqüência original, possivelmente incluindo os ruídos.

Assim, selecionando-se apenas os primeiros coeficientes é possível afirmar que a parte mais relevante da informação fora obtida e para verificar essa afirmação usou-se a equação inversa nos primeiros vinte coeficientes da série transformada. O resultado desse procedimento é mostrado na figura 4.7.c, onde é possível verificar que o formato geral da função foi mantido.

A abordagem desta dissertação, então, foi escolher os dez primeiros coeficientes para os parâmetros (seção 4.3) de todas as proteínas, superando assim o problema da diversidade na quantidade de resíduos e mantendo a parte mais relevante da informação. Vale ressaltar que a escolha de apenas dez coeficientes é um compromisso entre o custo computacional envolvido e o poder de predição do sistema.

O resultado do processamento dos dados de entrada é a proteína representada em vetores de entrada com a mesma estrutura em todos os classifi-

cadores locais binários. O vetor possui um tamanho de 90 entradas, já que existem 9 parâmetros selecionadas na seção 4.3 e todos são resumidos a 10 coeficientes.

## 4.5 Organização do Classificador Global

Na seção 3.3.5, referente a alguns aspectos de máquinas de vetor de suporte, mencionou-se que SVM é uma abordagem capaz de criar classificadores binários, e quando o domínio do problema possuir várias classes é necessário criar vários desses classificadores organizados de alguma forma.

O domínio das funções das proteínas possui várias classes, que representam todas as funções pertencentes a cobertura do modelo (seção 4.2). Denomina-se classificador global o agrupamento de todos os classificadores binários presentes no modelo.

O classificador global será responsável por, dada uma proteína com estrutura espacial conhecida, organizar os classificadores binários de modo a revelar quais funções são realizadas por ela. Os classificadores locais criados um para cada função pertencente a área de cobertura do sistema devolvem uma resposta do tipo sim, caso seja inferido que a proteína executa a função representada pelo classificador, ou não, caso contrário.

Para manter a consistência do classificador global, preferiu-se utilizar uma organização que mantenha a estrutura em DAG da GO (seção 4.1). Nesse caso, a *true path rule* deve também ser observada, ou seja, caso um classificador local $n$ ofereça uma resposta positiva para uma da proteína, todas as funções ancestrais daquela representada por $n$ devem, por propagação, receber uma resposta positiva e serem adicionadas, conseqüentemente, pelo classificador global. Caso a *true path rule* não fosse observada, a resposta dada pelo sistema poderia ser inconsistente.

Por exemplo, suponha que o classificador local representativo da função molecular '*DNA Binding*' devolva uma resposta positiva para uma dada proteína $P$. Então, observando a figura 4.2, é fácil supor que pela *true path rule* todas as funções ancestrais, notadamente '*nucleic acid binding*' e '*binding*', deverão também figurar entre as funções realizadas por $P$, já que '*DNA Binding*' é um caso especial de seus ancestrais.

Duas considerações devem ser feitas sobre o classificador global, a primeira é que como ele deve utilizar todos os classificadores binários para sugerir uma lista de funções, o tempo de resposta do sistema depende linearmente da quantidade de funções existentes em sua cobertura.

A segunda consideração é que este modelo é escalável quanto a adição de novas funções moleculares, um fato importante quando se leva em conta que os termos da GO são dinâmicos. A facilidade de adicionar novos termos ao modelo se dá pelo fato de que basta criar um novo classificador binário para a função que se deseja adicionar e acoplá-lo obedecendo a estrutura hierárquica da própria GO, sem, com isso, precisar modificar os classificadores binários já existentes.

### 4.5.1 Projeto do Conjunto de Treinamento dos Classificadores Locais

Além de se preocupar com a estrutura dos dados de entrada (representação vetorial das proteínas) e com a organização do classificador global, é preciso tomar cuidado com quais instâncias devem participar do treinamento de cada classificador local binário, já que tais instâncias devem também estar de acordo com a estrutura em DAG da GO. Essa pequena seção tem o objetivo de fechar essa lacuna, que é o último passo para se entender completamente o modelo criado.

Durante o treinamento, cada classificador local deve receber um conjunto de instâncias positivas e um conjunto de instâncias negativas. Por exemplo, se o objetivo é criar um classificador local que possa inferir se uma proteína exerce a função '*DNA Binding*', ele deve automaticamente receber durante o treinamento um conjunto de proteínas que executam essa função e um conjunto de proteínas que não a executam.

Nesse caso, é preciso escolher as proteínas que participarão do conjunto. Existem Várias maneiras de se fazer essa escolha, mas utilizou-se aqui a mais adequada à estrutura hierárquica do *Gene Ontology* (Eisner 2005). Por exemplo, se uma proteína está associada explicitamente a um nó $n_i$, pela *true path rule*, todos os classificadores de funções de nós antecessores a $i$ devem utilizar essa proteína como instância positiva. Assim, o conjunto de instâncias positivas de um dado nó é formado pelas proteínas que estão associadas explicitamente àquele nó, unido com as proteínas que estão associadas de maneira implícita pela '*true path rule*'.

É preciso delimitar também o conjunto de instâncias negativas para determinado nó. A primeira vista bastaria reconhecer como negativas todas as instâncias que não foram classificadas como positivas pela regra do parágrafo anterior. Entretanto, as proteínas que estão associadas aos nós antecessores do nó $i$ não são boas instâncias negativas, pois, de acordo com a especificação da GO, as proteínas são anotadas com os termos mais específicos que os

experimentos puderam concluir.

Em outras palavras, se uma proteína é associada ao termo '*Nucleic Acid Binding*' não significa que não pertença também ao termo '*DNA Binding*', o que pode ter acontecido é que testes mais específicos ainda não foram realizados. Dessa forma, não seria totalmente correto incluir essas proteínas no conjunto de instâncias negativas.

Resumindo, o conjunto de instâncias positivas de um nó é formado por todas as proteínas associadas explicitamente ao nó e pelas proteínas que são associadas aos nós descendentes do mesmo, e o conjunto de instâncias negativas de um dado nó é formado por todas aquelas que não foram classificadas como positivas, mas excluindo-se aquelas que estão associadas aos nós antecessores do nó em questão.

# Capítulo 5

# Análise Estatística do Modelo

A análise estatística do modelo foi realizada para cada classificador binário local e para o classificador global como um todo. Os classificadores locais serão avaliados utilizando algumas medidas estatísticas tradicionais abordadas na seção 5.1.1. Entretanto, para o classificador global, as medidas estatísticas tradicionais precisam ser reformuladas para levar em conta a estrutura hierárquica da ontologia utilizada. Tal necessidade será tratada na seção 5.1.2.

Como a intenção deste capítulo é simular o comportamento do sistema diante de proteínas desconhecidas, seria um erro efetuar a avaliação utilizando as mesmas proteínas que participaram da aprendizagem. Dessa forma, foi feita uma partição nos dados de modo a tornar possível os testes em dados desconhecidos para o modelo. O particionamento gerou dois conjuntos: o conjunto de treinamento, que corresponde a 2/3 dos dados originais e o conjunto de validação, correspondente ao 1/3 restante.

## 5.1 Medidas Estatísticas

Para avaliar os classificadores binários locais isoladamente os recursos tradicionais já são suficientes. Isso ocorre porque, por exemplo, para afirmar se uma dada proteína possui ou não a função molecular '*catalytic activity*' não são necessárias quaisquer informações a respeito dos outros nós da ontologia, ou seja, a estrutura hierárquica não é um ponto relevante.

### 5.1.1 Métodos Tradicional de Medidas Estatísticas

Para avaliar objetivamente um sistema de predição, algumas métricas que quantificam a qualidade da mesma são necessárias. As métricas escolhi-

das para tanto são os conceitos de '*recall*' e '*precision*', além da equação '*f-measure*', que é aplicada a ambos para obter um valor médio.

Para entender o valor dessas medidas é preciso levar em conta que a predição fornecida por um classificador binário pode recair apenas em quatro categorias:

***True Positive* - TP:** Ocorre quando o classificador prediz uma resposta do tipo 'sim' e essa resposta está de acordo com os dados reais.

***True Negative* - TN:** Ocorre quando o classificador prediz uma resposta do tipo 'não' e essa resposta está de acordo com os dados reais.

***False Positive* - FP:** Ocorre quando o classificador prediz uma resposta do tipo 'sim'. Entretanto, na realidade a resposta deveria ser 'não'.

***False Negative* - FN:** Ocorre quando o classificador prediz uma resposta do tipo 'não'. Entretanto, na realidade a resposta deveria ser 'sim'.

A maneira mais compacta de se visualizar essas categorias é utilizando a matriz de confusão (do inglês *confusion matrix*) ilustrada na tabela 5.1. Essa visualização possui maior utilidade quando os erros de classificação, notadamente FP e FN, são igualmente indesejáveis.

| | | Predição | |
|---|---|---|---|
| | | Sim | Não |
| Realidade | Sim | **TP** | **FN** |
| | Não | **FP** | **TN** |

Tabela 5.1: Matriz de Confusão

A semântica associada à métrica '*precision*', mostrada na equação 5.1, é medir o quanto das predições positivas estão corretas, ou seja, fixando-se o espaço amostral nas predições positivas do modelo, verifica-se a probabilidade de serem corretas.

$$precision = \frac{TP}{TP + FP} \tag{5.1}$$

O conceito de '*recall*' é levemente diferente, fixando-se o espaço amostral nas instâncias positivas do mundo real (nesse caso, TP e FN pela primeira linha da matriz de confusão da tabela 5.1), verifica-se a probabilidade de o classificador avaliá-las como positivas. A equação 5.2 mostra como o '*recall*' é calculado.

$$recall = \frac{TP}{TP + FN} \tag{5.2}$$

É importante notar que, dependendo da visão de um biólogo molecular, os conceitos de '*recall*' e '*precision*' podem possuir importâncias diferentes. Por exemplo, se o biólogo está preocupado com as chances de aquelas predições positivas para uma dada função estarem corretas ele deve olhar apenas para a medida 'precision'. Entretanto, se o biólogo está mais preocupado em as funções reais estarem classificadas como positivas ele deve levar mais em consideração o valor do '*recall*'.

Um classificador binário ideal deve possuir altos valores de '*recall*' e '*precision*'. Entretanto, intuitivamente é fácil perceber que, na maioria dos casos, isso não pode acontecer na realidade, pois para aumentar o '*recall*' significa classificar mais e mais funções como verdadeiras para uma dada proteína e, nesse caso, o valor de '*precision*' seria diminuído, já que a quantidade de instâncias falsas classificadas como verdadeiras aumenta.

Analogamente, para aumentar o valor de '*precision*' é necessário ser mais rígido ao classificar uma dada função como verdade e, conseqüentemente, muitas funções verdadeiras vão ser classificadas como falsas, o que diminui o valor de '*recall*'.

Entretanto, na maioria dos casos, é necessário uma medida estatística única, por isso existe a '*f-measure*' que é uma média entre '*recall*' e '*precision*'. A '*f-measure*' é um padrão em sistemas de aprendizagem de máquina e pode privilegiar tanto o '*recall* quanto o *precision* com o parâmetro $\beta$ (equação 5.3).

$$f - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall}, \beta \in [0, \infty] \tag{5.3}$$

Quando $\beta$ recebe o valor 1, então tanto '*recall*' quanto '*precision*' possuem a mesma importância. Para valores entre 0 e 1, '*recall*' possui um peso maior e para de 1 em diante '*precision*' possui mais importância. Dessa forma, $\beta$ pode ser ajustado de acordo com o grau de importância de cada um. No presente trabalho todos os testes foram executadas com $\beta$ fixado em 1.

### 5.1.2 Medidas Estatísticas Reformuladas Segundo o DAG

O principal fator que obriga reformular os conceitos de *recall*, *precision* e *f-measure* é que, em ontologias como a GO, o acontecimento de um dado erro pode ser mais indesejável que o acontecimento de outro.

Por exemplo, a proteína expressa pelo gene *lacY* está associada à função

'*Lactose transporter activity*' (veja figura 4.2). Se o sistema atribuir a esta proteína a função '*Disaccharide transporter activity*' que está a um nó acima no DAG, pode-se afirmar que um erro ocorreu. Entretanto, esse erro não pode ser considerado igual ao erro de classificar a proteína apenas como '*Transporter activity*' que está a quatro nós acima no DAG. Esse simples exemplo ilustra a necessidade de não se tratar todos os erros como igualmente ruins.

Intuitivamente, predições que no DAG estão próximas ao que acontece realmente devem ser melhor pontuadas que aquelas que estão em lugares distantes, o objetivo, então, é efetuar os cálculos da forma mais clara e intuitiva possível.

A definição formal do método utilizado neste trabalho foi elaborada por Poulin (2004) e será explicada mais adiante. Eisner (2005), também no campo da predição da função de proteinas, incrementou a teoria com vários conceitos importantes que serão mencionados adiante nesta seção, principalmente no que diz respeito a '*true path rule*'.

Para que a predição do modelo fique de acordo com esta regra é preciso que a resposta seja propagada para os nós anteriores. Assim, suponha que $A$, $B$, $C$, $D$ e $E$ sejam funções armazenadas na GO e que seguem o DAG mostrado na figura 5.1.

Suponha-se também a existência de duas proteínas $P_1$ e $P_2$, das quais $P_1$ assume explicitamente a função molecular $B$ e $P_2$ assume explicitamente a função molecular $C$. Pela '*true path rule*', sabe-se que, após a propagação, a proteína $P_1$ estará associada ao conjunto de funções $\{A, B\}$, enquanto que a proteína $P_2$ estará associada ao conjunto $\{A, B, C\}$. Essa simples conseqüência da hierarquia é importante para tornar as novas definições de '*precision*' e '*recall*' intuitivas.

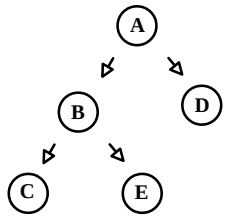


Figura 5.1: Exemplo de DAG

Suponha finalmente a existência de um classificador que, dadas as proteínas $P_1$ e $P_2$, tentará associá-las aos nós presentes no DAG da figura 5.1. Dado que o referido classificador tenha associado explicitamente à proteína $P_1$ a função $C$ e a proteína $P_2$ a função $B$, conclui-se que ambas as predições

estão incorretas.

Entretanto, com a utilização da '*true path rule*', as funções são propagadas para os nós antecessores. Assim, o que ocorre é que, na verdade, o classificador associou a proteína $P_1$ ao conjunto de funções $\{A, B, C\}$ e a proteína $P_2$ ao conjunto $\{A, B\}$. Dessa forma, percebe-se que a predição efetuada não foi completamente errada, pois, para ambas as proteínas, há intercessão entre o conjunto sugerido pelo classificador e o conjunto que ocorre na realidade.

Utilizando agora os conceitos de '*precision*' e '*recall*' vistos anteriormente, precisa-se analisar como ambos foram afetados pelos erros cometidos pelo classificador. Para a proteína $P_1$, cujo conjunto de funções é $\{A, B\}$, mas fora associada ao conjunto $\{A, B, C\}$, percebe-se que o '*recall*' não é alterado, pois fixando-se o espaço amostral nas instâncias positivas do mundo real, $\{A, B\}$, verifica-se que todas foram classificadas como positivas, ou seja, a quantidade de falsos negativos (FN) é zero. Entretanto, o valor de '*precision*' não tem o mesmo comportamento, já que ao fixar o espaço amostral nas instâncias positivas preditas pelo modelo percebe-se que apenas duas de três estão corretas.

Realizando a mesma análise para a proteína $P_2$, cujo conjunto de funções é $\{A, B, C\}$, mas fora associada ao conjunto $\{A, B\}$, percebe-se que o valor de '*precision*' é que não alterou, pois todas as funções preditas como verdadeiras o são na realidade. Entretanto, o '*recall*' é de apenas 2/3, pois das três funções, apenas duas foram preditas. A tabela 5.2 mostra de um modo resumido o exemplo.

| Proteínas | $A$ | $B$ | $C$ | TP | FN | FP | *Precision* | *Recall* |
|---|---|---|---|---|---|---|---|---|
| $P_1$ | $\otimes$ | $\otimes$ | $\times$ | 2 | 0 | 1 | 2/3 | 2/2 |
| $P_2$ | $\otimes$ | $\otimes$ | $\bigcirc$ | 2 | 1 | 0 | 2/2 | 2/3 |

$\times$ = Predição
$\bigcirc$ = Realidade

Tabela 5.2: Exemplo de '*recall*' e '*precision*' utilizando a hierarquia

Uma última observação que pode ser feita acerca do exemplo acima é que ao tentar fazer predições muito profundas no DAG corre-se o risco de reduzir o valor de '*precision*', mas sem alterar o de '*recall*'. Por outro lado, ao fazer predições próximas a raiz da árvore corre-se o risco de reduzir o valor de '*recall*', mas sem alterar o valor de precision. Esta última observação deve ser levada em conta caso se deseje alterar o valor de $\beta$ na '*f-measure*'.

Apesar de ser simples, claro e intuitivo, o método possui uma desvantagem por assumir que cada distância na hierarquia possui o mesmo peso (Eisner

2005), o que não é certo, pois não se sabe, por exemplo, como a distância entre '*binding*' e '*nucleic acid binding*' se compara com a distância entre '*binding*' e '*carbohydrate binding*'. Em Lin (1998) e Wang et al. (1999) é possível ter uma idéia de como essa diferença poderia ser quantificada.

## 5.2 Resultados

### 5.2.1 Análise dos Classificadores Locais

A performance obtida pelos classificadores locais utilizando as métricas '*precision*' e '*recall*' serão aqui mostradas. A tabela 5.3 mostra as métricas para todos os classificadores inclusos no estudo de caso, que corresponde a um subconjunto do que foi mostrado na figura 4.3.

| Funções Moleculares | *Precision* | *Recall* | *F-Measure* |
|---|---|---|---|
| Amine binding | 1,00 | 0,95 | 0,98 |
| Carbohydrate binding | 1,00 | 0,90 | 0,95 |
| Cofactor binding | 1,00 | 0,97 | 0,98 |
| Ion binding | 1,00 | 0,95 | 0,97 |
| Lipid binding | 1,00 | 0,92 | 0,96 |
| Nucleic acid binding | 0,97 | 0,91 | 0,94 |
| Amino acid binding | 1,00 | 0,92 | 0,96 |
| Sugar binding | 1,00 | 0,94 | 0,97 |
| Coenzyme binding | 1,00 | 0,98 | 0,99 |
| Pyridoxal phosphate binding | 1,00 | 0,95 | 0,97 |
| Cation binding | 0,99 | 0,95 | 0,97 |
| Metal ion binding | 1,00 | 0,95 | 0,97 |
| DNA binding | 0,92 | 0,88 | 0,90 |
| RNA binding | 1,00 | 0,91 | 0,95 |
| FAD binding | 1,00 | 0,96 | 0,98 |
| FMN binding | 1,00 | 0,96 | 0,98 |
| NAD binding | 1,00 | 0,99 | 1,00 |
| NADP binding | 1,00 | 0,97 | 0,98 |
| Calcium ion binding | 0,98 | 0,91 | 0,94 |
| Transition metal ion binding | 1,00 | 0,95 | 0,98 |
| Magnesium ion binding | 1,00 | 0,96 | 0,98 |
| Sequence-specific DNA binding | 1,00 | 0,84 | 0,92 |
| Transcription factor activity | 0,99 | 0,87 | 0,93 |

Tabela 5.3: '*Precision*' e '*recall*' dos classificadores locais

Devido ao número reduzido de funções moleculares do caso de uso, algumas análises como, por exemplo, relação das métricas estatísticas com o número de proteínas positivas e com a posição da função molecular na hierarquia ainda não podem ser consideradas conclusivas.

O que pode ser visto a partir da tabela 5.3 é que o modelo atingiu altos valores de '*Precision*' e '*recall*' nas funções moleculares testadas. De um modo geral, percebe-se que o modelo tende a classificar uma proteína como positiva apenas quando há evidências conclusivas para isso, o que gera valores de '*Precision*' maiores que os de '*recall*'.

A coluna '*F-Measure*' da tabela 5.3 apresenta apenas valores no intervalo de $0,9$ a $1$, o que demostra que em média o modelo possui respostas satisfatórias para funções não tão específicas como as usadas no presente trabalho. É importante ressaltar que essa análise não pode ser extrapolada indiscriminadamente, pois resultados de outros trabalhos mostram que, em geral, uma boa performance em funções mais altas na hierarquia não é tão difícil de se obter.

O ponto principal desta análise do modelo é mostrar que o mesmo é viável para predizer a função molecular das proteínas com parâmetros calculados a partir da conformação espacial e sem a necessidade de alinhamentos. Entretanto, esforços ainda precisam ser feitos e análises mais conclusivas ainda devem ser realizadas para que se possa comparar o presente modelo com outros existentes na literatura.

Como última instância da análise, resta calcular os valores de precision e recall para o classificador global, criado como um agrupamento dos classificadores locais apresentados nesta seção.

### 5.2.2 Análise do Classificador Global

A análise dos classificadores binários locais mostrada na seção anterior foi realizada antes de se utilizar a *true path rule*, o que não chega a ser um erro ou problema, pois, como já foi citado, os classificadores locais não precisam conhecer a estrutura hierárquica da DAG.

No caso do classificador global, ignorar a estrutura hierárquica da ontologia gera resultados inconsistentes, pois uma possível resposta seria o sistema responder, por exemplo, positivamente para uma dada função $X$ e negativamente para outra função $Y$, mesmo havendo um arco cuja origem é $Y$ e o destino é $X$.

Em outras palavras, dadas as proteínas $A$ e $B$, conforme a figura 5.1, a inconsistência ocorrerá se $A$ receber classificação negativa enquanto $B$ recebe classificação positiva. Nesse caso, a consistência será obtida tanto se a classificação positiva de $B$ for propagada em direção ao nó $A$ quanto se a classificação negativa de $A$ for propagada em direção ao nó $B$. Dentre essas duas estratégias optou-se por propagar as classificações positivas em direção

aos nós ancestrais por ter gerado resultados bastante superiores experimentalmente.

Com essa estratégia de propagação o modelo apresentou os valores de '*recall*', '*precision*' e '*f-measure*' mostrados na tabela 5.4.

| *Precision* | *Recall* | *F-Measure* |
|---|---|---|
| 0,98 | 0,93 | 0,96 |

Tabela 5.4: '*Precision*' e '*recall*' do classificador global

É possível perceber pela tabela 5.4 que o classificador global se aproxima da média de todos os classificadores locais. Além disso, a tendência em priorizar '*Precision*' em detrimento de '*recall*' permaneceu, o que já era esperado, pois o classificador global depende das análises pontuais estabelecidas pelos classificadores locais.

Ao término deste capítulo de análise estatística fica estabelecida uma justificativa para futuros esforços de pesquisa sobre o presente modelo ou sobre outros modelos que utilizem paradigmas parecidos.

# Capítulo 6

# Conclusão

## 6.1 Contribuições e Relevância

Neste trabalho, foram apresentados um modelo de aprendizagem baseado em máquinas de vetor de suporte, a representação das proteína de acordo com o referido modelo e a análise dos resultados obtidos usando um subconjunto de funções da GO.

O desenvolvimento do modelo de aprendizagem, inclusive a utilização da estrutura hierárquica da GO, obedeceu à idéia básica utilizada por Eisner (2005) de criar vários classificadores binários locais que seguem a '*true path rule*'. A diferença entre os dois métodos reside nas informações extraídas das proteínas para a confecção do modelo. Na arquitetura proposta aqui, a estrutura terciária das proteínas recebeu um papel de destaque, pois foi utilizada para calcular os parâmetros estruturais utilizados como entrada, enquanto o trabalho de Eisner (2005) limitou-se a usar a estrutura primária.

A utilização do método de máquinas de vetor de suporte baseado na estrutura hierárquica em DAG da própria ontologia fez com que a resposta do sistema se tornasse consistente. Nesse ponto, é importante observar que o projeto da camada de entrada de cada classificador local apresentada na seção 4.5.1 é de fundamental importância.

Entre as contribuições apresentadas por este trabalho estão:

- A possibilidade de se utilizar a estrutura das proteínas em um modelo de predição de função que não é baseado em similaridades seqüenciais ou estruturais.

- As funções da GO são dinâmicas e o modelo proposto é capaz de ser atualizado facilmente para acompanhar as mudanças ocorridas na ontologia

bastando, para isso, incluir um novo classificador local para cada nova função.

- Os classificadores locais permitem verificar se uma dada função molecular pode ser da alçada de uma proteína, permitindo utilizar em laboratório apenas os testes específicos que confirmem a predição e reduzindo o tempo e os custos que seriam necessários em testes laboratoriais mais abrangentes.

- O classificador global é capaz de fornecer uma lista de funções moleculares possivelmente desempenhadas por uma dada proteína.

O produto final da dissertação, o modelo, mostrou ser capaz de realizar predições nas funções escolhidas para estudo de caso, demonstrando que a idéia inicial é válida e merece especial atenção como alvo de estudos.

## 6.2 Limitações e Restrições

Devido a limitação dos objetivos deste trabalho, ficaram fora do escopo vários detalhes importantes para a consolidação do modelo proposto. Sem a idéia de exaurir estas questões, serão listadas abaixo algumas das mais importantes:

- O conjunto de funções analisadas no estudo de caso é bastante limitado, o que obriga a dizer que a análise não pode ser tratada como definitiva, mas já é suficiente para concluir que o tema possui relevância como área de estudo. Entretanto, é necessário incluir na análise muitos outros termos da GO que constam na área de cobertura do modelo, de preferência liberando a restrição feita na seção 4.2 quanto a profundidade máxima da hierarquia, pois como a maioria dos outros modelos perdem acurácia com termos mais profundos, seria interessante analisar o comportamento do presente modelo.

- Para futuras análises sugere-se, também, incluir um estudo acerca de como o modelo se comporta diante de funções biológicas correlacionadas (mencionadas na seção 4.1), para verificar se a correlação é obtida automaticamente pelo modelo, como ocorre em Eisner (2005), ou se alterações serão necessárias para que o modelo consiga assimilar a correlação.

- Sobre as características extraídas das proteínas (seção 4.3), sugere-se utilizar técnicas de seleção de atributos e verificar se o conjunto de entrada do modelo pode ser otimizado. Esse procedimento será importante

para, posteriormente, verificar quais atributos são os mais importantes no processo de predição da função de proteínas.

- Outros métodos de aprendizagem de máquina como as redes bayesianas ou outros modelos de redes neurais artificiais poderiam ser testados para verificar qual técnica apresenta melhores resultados. O método baseado em máquinas de vetor de suporte foi escolhido pelas propriedades que o tornam capazes de lidar com problemas complexos (Haykin 2001) e por ser relativamente pouco oneroso computacionalmente.

# Apêndice A

# Alinhamentos de Seqüências e Estruturas de Proteínas

Na seção 2.4.3 mencionou-se que proteínas que possuem uma certa semelhança podem ser agrupadas em famílias, onde cada membro de uma família possui características que lembram os outros. O alinhamento é o método computacional responsável por encontrar medidas associadas a similaridades entre proteínas diferentes.

A intenção principal dos alinhamentos é a comparação de dados biológicos relacionados, a fim de identificar o quanto são similares entre si. Entretanto, as definições de alinhamento mudam se os dados biológicos são seqüências ou estruturas de proteínas. Dessa forma, a seção A.1 trata do problema de comparar duas seqüências e a seção A.2 aborda o problema de alinhar a estrutura de duas macromoléculas.

## A.1 Alinhamentos de Seqüências

Alinhamentos de seqüências é o procedimento que consiste em comparar duas (alinhamento par-a-par) ou mais (alinhamento múltiplo) seqüencias em busca de padrões que se repetem na mesma ordem. No caso de seqüências de proteínas, a busca é por resíduos de aminoácidos que se repetem em ambas as proteínas.

É possível definir o alinhamento de seqüências como a inserção de buracos em pontos aleatórios de modo a fazer com que elas fiquem do mesmo tamanho. No alinhamento par-a-par de seqüências, duas cadeias (DNA ou proteína) de mesmo tamanho ou não são postas horizontalmente de forma a identificar caracteres semelhantes na mesma coluna (*match*).

Entretanto, quando a mesma coluna apresenta caracteres diferentes diz-

se que ocorreu uma substituição naquela posição (*mismatch*), que pode ser causada, por exemplo, por alguma mutação ocorrida no processo evolutivo. Os pontos onde os buracos foram colocados são normalmente chamados de *gaps*.

O processo de alinhamento fornece uma medida numérica que indica a quantidade de caracteres repetidos na mesma coluna em ambas as seqüências. Tal medida é chamada de medida de similaridade e pontua positivamente os *matches* e negativamente os *mismatches* e os *gaps*. A medida de similaridade pode ser usada para identificar proteínas relacionadas. Quando um alinhamento possui a maior medida de similaridade para duas seqüências, diz-se que um alinhamento ótimo fora obtido.

Existem vários tipos de alinhamento de seqüências, dois são mais importantes para os métodos de predição da função de proteínas: o alinhamento global e o local, que diferem entre si no modo como as proteínas são arranjadas em colunas. Para o alinhamento global, as seqüências inteiras são alinhadas de modo a incluir o maior número possível de resíduos semelhantes na mesma coluna.

O alinhamento local prioriza encontrar subregiões com alta densidade de *matches*. Tal alinhamento é mais indicado para seqüências que são similares ao longo de determinada subregião, mas dissimilares em sua grande maioria. Esse tipo de alinhamento possui grande importância biológica por ser capaz de identificar regiões funcionais (seção 2.4.2).

Até o momento explicou-se apenas o caso do alinhamento de duas seqüências, porém o alinhamento múltiplo é bem mais complexo. A idéia agora é encontrar regiões conservadas em um conjunto de seqüências de modo a definir domínios funcionais com precisão. O conjunto de regiões conservadas encontradas via alinhamento múltiplo de proteínas da mesma família pode ser considerado a assinatura que identifica essa família.

A figura A.1 exemplifica um alinhamento múltiplo entre proteínas semelhantes presentes em ratos e seres humanos. A seqüência de consenso (*consensus*) mostrada na figura representa os aminoácidos que mais se repetem. Algumas vezes, informações relevantes sobre uma nova proteína podem ser extraídas realizando-se um alinhamento par-a-par com a consenso.

Entretanto, a seqüência de consenso ignora muitas informações relevantes como, por exemplo, a ocorrência de outros caracteres em uma dada coluna é omitida. Assim, foram criadas diferentes formas de representação de padrões mais flexíveis como matrizes de pontuação específicas da posição (PSSM), que é uma matriz que armazena uma probabilidade de ocorrência de cada aminoácido, e modelos ocultos de Markov (Eddy et al. 1995), que são uma formulação

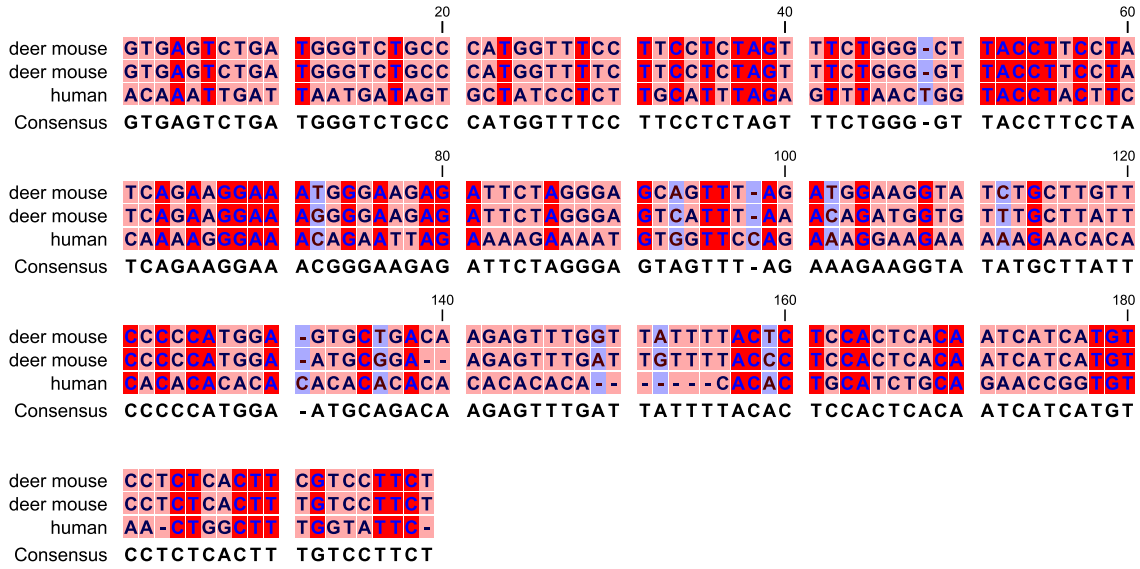


Figura A.1: Alinhamento Múltiplo

probabilística rigorosa de padrões de seqüências com uma sensibilidade bastante superior à PSSM. Como uma leitura introdutória sobre esse assunto, recomenda-se Gibas & Jambeck (2001).

## A.2 Alinhamentos de Estruturas Tridimensionais

O alinhamento de estruturas tridimensionais é um tema bem mais complexo a ser tratado, pois, além de a comparação e sobreposição de estruturas tridimensionais ser uma tarefa bem mais complexa que trabalhar com a seqüência primária, há o fato de que vários aspectos diferentes da estrutura tridimensional podem ser extraídos como, por exemplo, comprimento da ligação, polaridade e acessibilidade da superfície. Em alguns casos, é interessante fazer buscas para encontrar não as proteínas geometricamente similares, mas também as proteínas similares quimicamentes.

Em se tratando do caso da similaridade geométrica, o parâmetro mais comum é o desvio médio da raiz quadrada (RSMD), calculado em função dos átomos da cadeia central de uma proteína. Como a representação desses átomos se dá através de coordenadas cartesianas, o RSMD leva em conta a distância entre os átomos em uma estrutura e os mesmos átomos em outra. Um alinhamento ótimo entre duas estruturas é aquele que possui o menor RSMD possível.

# Referências Bibliográficas


Ahmed, N., Natarajan, T. & Rao, K. R. (1974), 'Discrete cosine transform', *IEEE Trans. Biomed. Eng* **23**, 90–93.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002), *Molecular Biology of THE CELL*, Garland Science.

Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001), 'Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking', *Journal of Molecular Biology* **311**, 395–408.

Altschul, S. F., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology* **215**, 403–410.

Andrade, M., Brown, N., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. & Sander, C. (1999), 'Automated genome sequence analysis and annotation', *Bioinformatics* **15**, 391–412.

Anfinsen (1973), 'Principles that govern the folding of protein chains', *Sceince* pp. "223–230".

Ashburner, M., Ball, C. & et al, J. B. (2000), 'Gene ontology: tool for the unification of biology', *Nature Geneticist* **25**, 25–29.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000), 'The pfam protein families database', *Nucleic Acids Research* **28**, 263–266.
*citeseer.ist.psu.edu/bateman02pfam.html

Bedell, J., Korf, I. & Yandell, M. (2003), *BLAST*, O'Reilly.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Lype, L., Jain, S. & et al (2002),

'The protein data bank', *Acta Crystallogr D Biol Crystallogr* **58 (pt 6 N 1)**, 899–907. www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html.

Binkowski, T. A., Adamian, L. & Liang, J. (2003), 'Inferring functional relationships of proteins from local sequence and spatial surface patterns', *Journal of Molecular Biology* **332**, 505–526.

Borro, L. C., Oliveira, S. R. M., Yamagishi, M. E. B., Mancini, A. L., Jardine, J. G., Mazoni, I., Santos, E. H., Higa, R. H., Kuser, P. R. & Neshich, G. (2006), 'Predicting enzyme class from protein structure using bayesian classification', *Genetic Molecular Research* **5**, 193–202.

Bowie, J. U., Luethy, R. & Eisenberg, D. (1991), 'A method to identify protein sequences that fold into a known three-dimension structure', *Science* **253**, 164–170.

Branden, C. & Tooze, J. (1991), *Introduction to protein structure*, Garland Publishing, New Yourk.

Breiman, L., Friedman, J. H. & Olshen, R. A. (1993), *Classification and Regression Trees*, Chapman and Hall.

Brenner, S. E., Crothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996), 'Understanding protein structure: Using scop for fold interpretation', *Methods in Enzymology* **266**, 635–643.

Bron, J. C. & Kerbosch, J. C. (1973), 'Finding all cliques of an undirected graph', *Communications of ACM* **16**, 575–577.

Brown, T. A. (1999), *Genética Um enfoque Molecular*, Guanabara Koogan.

Chang, C.-C. & Lin, C.-J. (2001), *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Devos, D. & Valencia, A. (2000), 'Pratical limits of function prediction', *Proteins* **41**, 98–107.

Dodge, C., Schnerder, R. & Sander, C. (1998), 'The hssp database of protein structure-sequence alignments and family proles'. *citeseer.ist.psu.edu/dodge98hssp.html

Duda, R. O. & Hart, P. (1973), *Pattern Classification and Scene Analysis*, Jon Wiley and Sons.

Eddy, S., Mitchison, G. & Durbin, R. (1995), 'Maximum discrimination hidden markov models of sequence consensus', *Journal of computational biology* **2**, 9–23.

Eisner, R. (2005), Prediction protein function using machine-learned hierarchical classifier, Master's thesis, Univervity of Alberta.

Eisner, R., Poulin, B., Szafron, D., Lu, P. & Greiner, R. (2005), 'Improving protein function prediction using the hierarchical structure of the gene ontology'.

Ferre, F., Ausiello, G., Zanzoni, A. & Helmer-Citterich, M. (2005), 'Functional annotation by identification of local surface similarities: a novel tool for structural genomics', *BMC Bioinformatics* **6**, 194.

Friedberg, I. (2006), 'Automated protein function prediction: The genomic challenge', *Briefings in Bioinformatics* .

Gerlt, J. A. & Babbitt, P. (2000), 'Can sequence determine function?', *Genome Biology* **1**, 1–10.

Gibas, C. & Jambeck, P. (2001), *Desenvolvendo Bioinformática*, Campus.

Gilks, W. R., Audit, B. & Angelis, D. (2005), 'Percolation of annotation errors through hierarchically structured protein sequence databases', *Math Biosci* **193**, 223–234.

Haykin, S. (2001), *Redes Neurais Princípios e Práticas*, 2 edn, Bookman.

Hearst, M. A., Schölkopf, B., Dumais, S., Osuna, E. & Platt, J. (1998), 'Trends and controversies - support vector machines', *IEEE Intelligent Systems* **13**, 18–28.

Hinton, G. E. (1989), 'Connectionist learning procedures', *Artificial Intelligence* **40**, 185–234.

Holm, L. & Sander, C. (1998), 'Touring protein folding space with dali/fssp', *Nucleic Acid Research* **26**(1), 316–319.

Israelachvili, J. N. (1991), *Intermolecular and Surface Forces*, second edn, Academic Press.

Kauvar, L. M. & Vilar, H. O. (1998), 'Deciphering cryptic similarities in protein binding sites', *Curr Opin Biotechnol* **9**, 390–394.

Kincaid, D. & Cheney, W. (2002), *Numerical Analysis*, 3 edn, Brooks/Cole.

Kobayashi, N. & Go, N. (1997), 'Atp binding proteins with different folds share a common atp-binding structural motif', *Nat Struct Biol* **4**, 6–7.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Lnza, L., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A. & Robles, V. (2006), 'Machine learning in bioinformatics', *Briefings in Bioinformatics* **7**(1), 86–112.

León, D. & Markel, S. (2003), *Sequence Analysis in a Nutshell*, O'Reilly.

Lesk, A. M. & Fordham, W. D. (1996), 'Conservation and variability in the structures of serine proteinases of the chymotrypsin family', *Journal of Molecular Biology* **258**, 501–537.

Lin, D. (1998), 'An information-theoretic definition of similarity', *Proceedings of the 15th International Conference on Machine Learning* pp. 296–304.

Lorena, A. C. (2006), Investigação de estratégias para geração de máquinas de vetores de suporte multiclasses, PhD thesis, Instituto de Ciências Matemáticas e de Computação - ICMC - USP.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002), 'Structural clusters of evolutionary trace residues are statistically significant and common in proteins', *Journal of Molecular Biology* **316**, 139–154.

Mayoraz, E. & Moreira, M. (1996), On the decomposition of polychotomies into dichotomies, Technical report, Dalle Molle Institute for Perceptive Artificial Intelligence.

Moody, J. E. & Rögnvaldsson, T. (1997), 'Smoothing regularizers for projective basis function networks', *Quarterly Journal of Experimental Psychology* **27**, 56–60.

Mount, D. (2001), *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press.

Murray, R. K., Granner, D. K., Mayes, P. A. & Rodwell, V. W. (1994), *Harper: Bioquímica*, Aheneu.

Nicholls, A., Sharp, K. A. & Honig, B. (1991), 'Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons', *Proteins* **11**(4), 281–296.

Norel, R., Fischer, D., Wolfson, H. J. & Nussinov, R. (1994), 'Molecular surface recognition by a computer vision-based technique', *Protein Eng* **7**, 39–46.

Pal, D. & Eisenberg, D. (2005), 'Inference of protein function from protein structure', *Structure* **13**, 121–130.

Poulin, B. (2004), Sequence-based protein function prediction, Master's thesis, University of Alberta.

Radzicka, A. & Wolfenden, R. (1988), 'Comparing the polarities of the amino-acids – side-chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous-solution', *Biochemistry* **27**, 1664–1670.

Rosenblatt, F. (1962), *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books.

Rost, B. (2002), 'Enzyme function less conserved than anticipated', *Journal of Molecular Biology* **318**, 595–608.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), 'Learning internal representations by backpropagation errors', *Nature* **323**, 533–536.

Schmitt, S., Kuhn, D. & Klebe, G. (2002), 'A new method to detect related function among proteins independent of sequence and fold homology', *Journal of Molecular Biology* **323**, 384–406.

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

Shrake, A. & Rupley, J. A. (1973), 'Environment and exposure to solvent of protein atoms. lysozyme and insulin', *Journal of molecular biology* **79**, 351–371.

Sridharan, S., Nicholls, A. & Honig, B. (1992), 'A new vertex algorithm to calculate solvent accessible surface areas', *Biophys* **61**, A174.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society Series B* **36**, 111–147.

Tipton, K. & Boyce, S. (2000), 'History of the enzyme nomenclature system', *Bioinformatics* **16**, 34–40.

Todd, A. E., Orengo, C. A. & Thornton, J. M. (20002), 'Plasticity of enzyme active sites', *Trends Biochem Sci* **27**, 419–426.

Wang, K., Zhou, S. & Liew, S. C. (1999), Building hierarchical classifiers using class proximity, *in* '25th International Conference on Very Large Databases', pp. 363–374.

Webb, E. C. (1992), *Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomeclature and Classification of Enzymes*, Academic Press.

Weigende, A. S., Rumelhart, D. E. & Huberman, B. A. (1991), 'Generalization by weight-elimination with application to forecasting', *Advances in Neural Information Processing Systems* **3**, 875–882.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L. & Lichtarge, O. (2003), 'An accurate, sensitive, and scalable method to identify functional sites in protein structures', *Journal of Molecular Biology* **326**, 255–261.

Assinatura do Aluno

Assinatura do Orientador

Assinatura do Orientador