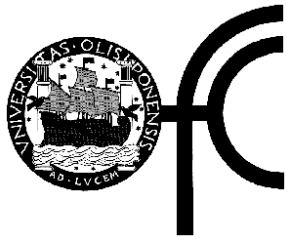


DEPARTAMENTO DE INFORMÁTICA

Faculdade de Ciências - Universidade de Lisboa
Bloco C5 - Piso 1 - Campo Grande, 1700 Lisboa
Tel & Fax: +351-1-7500084

PREVISÃO DA
ESTRUTURA SECUNDÁRIA DE PROTEÍNAS
UTILIZANDO
REDES NEURONAIS

Trabalho realizado sob a bolsa
PRAXIS XXI BM/15046/98

SARA GUILHERME OLIVEIRA DA SILVA

Dissertação apresentada na
Faculdade de Ciências da Universidade de Lisboa
para obtenção do grau de Mestre em Informática

LABORATÓRIO DE MODELOS E ARQUITECTURAS COMPUTACIONAIS

LISBOA, OUTUBRO DE 1999

Co-orientadores:

Doutor J. Félix Costa

Doutor Pedro J.N. Silva

RESUMO

Nos últimos anos, a previsão da estrutura secundária de proteínas tem sido uma das áreas de maior actividade em bioinformática. Inúmeros sistemas de previsão têm sido desenvolvidos, muitos deles utilizando redes neuronais. Baseado num dos mais bem sucedidos, o PHD, o sistema aqui desenvolvido utiliza o perceptrão multicamada como método de previsão. Foram estudadas diferentes implementações do sistema, fazendo variar o número de redes e a dimensão dos estímulos utilizados; aplicando filtros às previsões obtidas; e testando diversos métodos de separação estrutural *a priori* das proteínas a classificar, utilizando o perceptrão, o mapa de Kohonen e regras de classificação. Um índice de fiabilidade associado às previsões foi estudado e comparado com o índice utilizado no PHD. Os resultados obtidos demonstram que o sistema de previsão obtido, embora bastante mais simples do que o PHD, consegue ser pelo menos igualmente bem sucedido.

À minha tia Aldina,

possivelmente a pessoa a quem
eu mais gostaria de oferecer um
exemplar desta dissertação.

ÍNDICE RESUMIDO

ÍNDICE

PARTE III

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

PARTE I

1 INTRODUÇÃO

Desde longa data que a ciência da computação tem vindo a desenvolver arquitecturas e algoritmos baseados em mecanismos biológicos, que posteriormente se revelam adequados à realização de diferentes tarefas, nos mais variados campos científicos. É o caso das redes neuronais artificiais, particularmente bem sucedidas em tarefas de classificação, e dos algoritmos genéticos, especialmente adequados a problemas de optimização. Quando estes e outros paradigmas computacionais (*e.g.* inteligência artificial) são utilizados na resolução de problemas no âmbito das ciências biológicas, entra-se num vasto campo interdisciplinar designado por bioinformática.

A crescente abundância de dados e a forte melhoria dos recursos computacionais disponíveis, que se têm verificado nos últimos anos, provocaram o aumento drástico do número de ferramentas computacionais de processamento e simulação que complementam, ou mesmo substituem, muitas tarefas experimentais. Algumas aplicações das ferramentas da bioinformática incluem a procura de informação nas bases de dados, o reconhecimento e identificação de genes, a inferência de árvores filogenéticas, a previsão da estrutura secundária do RNA, a elaboração de alinhamentos múltiplos e a determinação da estrutura e função de proteínas [Baldi e Brunak 98, Schulze-Kremer 95]. Em particular, a previsão da estrutura secundária de proteínas tem sido uma área de intensa e competitiva actividade.

A ideia de utilizar redes neuronais na previsão da estrutura secundária de proteínas surgiu de uma forma curiosa. O sistema NETtalk, desenvolvido por Sejnowski e Rosenberg [Sejnowski e Rosenberg 87], consiste numa rede neuronal que aprende a pronunciar texto escrito em inglês – uma janela com dimensão de sete letras move-se ao longo do texto, sendo a rede treinada para pronunciar o fonema correspondente à letra central. Após uma palestra acerca do NETtalk, uma pessoa da audiência sugeriu a Sejnowski que, usando aminoácidos em vez de letras, seria possível prever a estrutura secundária de proteínas [Anderson e Rosenfeld 98]. O trabalho então publicado por Qian e Sejnowski [Qian e Sejnowski 88] demonstrou que as redes neuronais conseguiam melhores resultados do que qualquer outro método de previsão de estrutura secundária utilizado anteriormente. Seguiu-se uma longa série de trabalhos análogos, culminando naquele que parece ser o mais bem sucedido até ao momento, denominado PHD [Rost e Sander 93].

Fortemente baseado no PHD, procurou-se que o sistema de previsão aqui desenvolvido fosse, no mínimo, igualmente bem sucedido. Numa busca constante de simplicidade, tentou-se aproveitar apenas as características do PHD que lhe garantem o sucesso, desprezando aquelas cujo papel é menos óbvio.

O desenvolvimento deste sistema passou por duas fases distintas. A primeira fase consistiu na exploração das bases de dados disponíveis, de onde se pode extrair toda a informação necessária. Os ficheiros das bases de dados foram alvo de diversas rotinas de filtragem e conversão, de cuja elaboração resultou o programa responsável pela codificação e normalização dos dados a usar na fase seguinte. A segunda fase, sem dúvida a mais extensa, incluiu um trabalho intensivo de elaboração dos programas de

simulação das redes neuronais, BackProp 2.1 e Kohonen 1.0. Seguiram-se diversas etapas de treino e teste destas redes, nos conjuntos de dados obtidos na fase anterior. A segunda fase incluiu ainda o trabalho de compilação, tratamento e interpretação dos resultados.

Todos os programas foram escritos na linguagem Delphi 3.0, podendo ser utilizados apenas em ambientes Windows 95/98. Embora não sejam acompanhados por manuais de utilização ou ficheiros de ajuda, a sua disponibilização é uma opção a considerar. Não existe um programa final que realize previsões directamente a partir da informação contida nas bases de dados, mas a sua elaboração e disponibilização pertencem a um plano de trabalho adicional a realizar futuramente.

Esta dissertação encontra-se dividida em três partes. A primeira parte, na qual se insere esta introdução, inclui mais dois capítulos, dedicados aos dois temas principais deste trabalho: proteínas e redes neuronais. Neles são descritos todos os conceitos considerados necessários à compreensão do trabalho realizado.

A segunda parte é iniciada por um capítulo que descreve a origem e transformação dos dados utilizados, assim como a forma como são apresentados os resultados. Segue-se o capítulo mais longo da dissertação, que descreve todos os passos considerados importantes no estudo e desenvolvimento do sistema de previsão aqui apresentado, incluindo os respectivos resultados. Embora longo, este capítulo não inclui de modo algum todas as tentativas falhadas que ocorreram abundantemente ao longo do desenvolvimento do sistema, cuja descrição exaustiva seria certamente desprovida de interesse.

Finalmente, a terceira parte inclui somente um capítulo, que aborda alguns aspectos adicionais considerados importantes no âmbito da previsão da estrutura secundária de proteínas, terminando com uma breve conclusão.

2 PROTEÍNAS

Este capítulo descreve todos os aspectos relacionados com as proteínas considerados necessários à compreensão do trabalho realizado. A primeira secção resume o modo como as proteínas são sintetizadas, de uma forma extremamente simplista. Segue-se o tema mais pormenorizado do capítulo, a constituição e formação da estrutura das proteínas, logo seguido da descrição de uma classificação de proteínas baseada na sua estrutura. A secção seguinte tem como tema a homologia e, finalmente, a terminar o capítulo, são descritos os principais métodos de previsão da estrutura secundária de proteínas.

2.1 Síntese

O *ácido desoxirribonucleico* (DNA), presente em todas as células vivas, consiste numa longa hélice dupla formada por um esqueleto de fosfato e açúcar e por pares de moléculas denominadas *bases*. As duas metades da hélice são complementares, pois cada um dos quatro tipos de bases apenas pode emparelhar com a base do tipo complementar (figura 2.1).

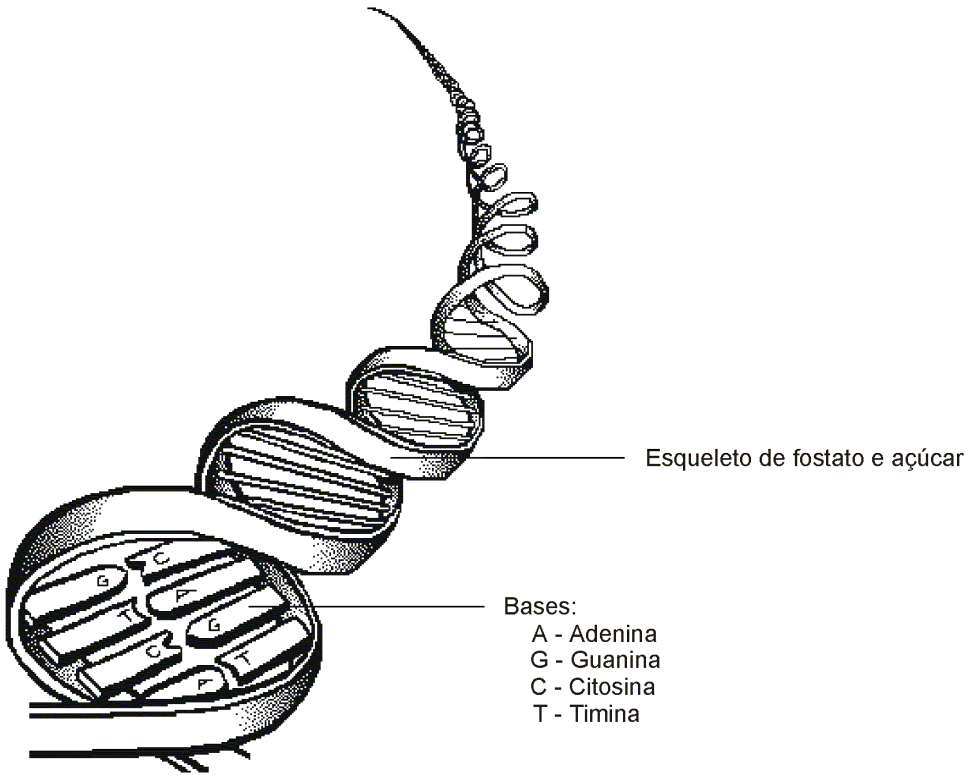


Figura 2.1 – Estrutura do DNA.

Embora a sequência de bases do DNA seja contínua, diferentes segmentos constituem unidades funcionais independentes, denominadas *genes*. São estes que contêm a informação necessária à síntese das proteínas, macromoléculas essenciais para o metabolismo dos seres vivos, de que são exemplos as enzimas, os anticorpos e várias hormonas.

Quando uma célula recebe o sinal para produzir uma proteína, uma das cadeias da hélice dupla de DNA serve de molde para a síntese de uma sequência de bases complementar, denominada *ácido ribonucleico mensageiro* (mRNA), num processo denominado *transcrição*. Cada tripleto ordenado de bases do mRNA, designado por *codão*, codifica uma de 20 moléculas, os *aminoácidos*, de que são feitas as proteínas, ou a terminação da proteína (tabela 2.1). A informação contida no mRNA é traduzida numa sequência de aminoácidos, que se vão ligando uns aos outros numa cadeia linear, denominada *cadeia polipeptídica*. Cada proteína é formada por uma ou mais cadeias polipeptídicas.

Tabela 2.1 – Código genético.

	T	C	A	G	
T	Phe	Ser	Tyr	Cys	**T**
					C
	Leu		Ter	Ter	**A**
				Trp	**G**
C	Leu	Pro	His	Arg	**T**
					C
			Gln		**A**
					G
A	Ile	Thr	Asn	Ser	**T**
					C
			Lys	Arg	**A**
	Met				**G**
G	Val	Ala	Asp	Gly	**T**
					C
			Glu		**A**
					G

Ordem dos tripletos: Esquerda – Topo – Direita
(Exemplo: ATG codifica Metionina)

Ter = Terminação

2.2 Estrutura

Ao descrever uma proteína, é costume distinguir quatro aspectos diferentes da sua estrutura: primária, secundária, terciária e quaternária. As três últimas constituem a *conformação*, ou *estrutura espacial*, da proteína.

2.2.1 Estrutura primária

2.2.1.1 Aminoácidos – composição e estrutura

Os aminoácidos são formados por um carbono central a que se ligam um hidrogénio, um *grupo carboxílico* (COOH) e um *grupo amínico* (NH_2), comuns a todos os aminoácidos, e um *grupo R*, ou *cadeia lateral*, que os distingue entre si. A figura 2.2 mostra duas fórmulas estruturais e uma representação gráfica de um aminoácido genérico, onde R denota a cadeia lateral.

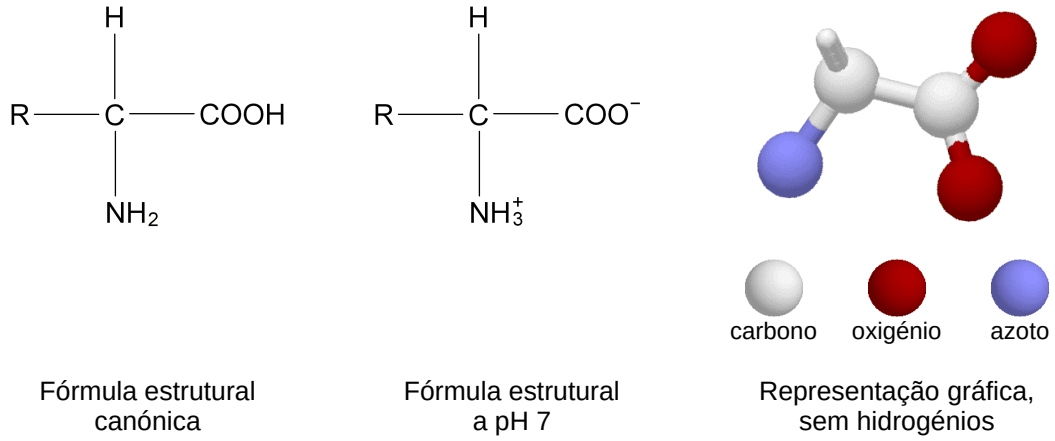


Figura 2.2 – Aminoácido genérico.

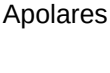
As cadeias laterais podem diferir bastante no seu tamanho, forma e propriedades químicas, sendo comum agrupar-se os aminoácidos em quatro classes, com base na sua polaridade: (1) apolares, ou hidrofóbicos, (2) polares neutros, (3) carregados positivamente, ou básicos, e (4) carregados negativamente, ou acídicos. Para além dos seus nomes, os aminoácidos podem ser designados por símbolos de um ou três caracteres. A tabela 2.2 resume algumas características dos 20 aminoácidos que constituem as proteínas.

2.2.1.2 Cadeia polipeptídica

Durante a síntese da proteína, o grupo carboxílico de um aminoácido e o grupo amínico de outro libertam uma molécula de água e formam uma ligação covalente[1] denominada *ligação peptídica*. O que resta de cada aminoácido designa-se por *resíduo de aminoácido*; muitas vezes os dois termos são usados indiscriminadamente. A cadeia polipeptídica contém de algumas dezenas a várias centenas de resíduos de aminoácidos que, ligados deste modo, formam um esqueleto em zig-zag de onde protraem as várias cadeias laterais. A figura 2.3 ilustra o processo de formação de ligações peptídicas e a cadeia polipeptídica resultante.

[1] Ligação entre dois átomos com partilha de dois ou mais electrões.

Tabela 2.2 – Características dos 20 aminoácidos que constituem as proteínas.

Classe de polaridade	Nome	Símbolos		Massa[2]	Cadeia lateral
Apolares	Alanina	Ala	A	89	CH_3 -
	Valina	Val	V	117	CH_3-CH(CH_3) -
	Leucina	Leu	L	131	CH_3-CH(CH_3)-CH_2 -
	Isoleucina	Ile	I	131	CH_3-CH_2-CH(CH_3) -
	Prolina	Pro	P	115	
	Fenilalanina	Phe	F	165	Phenyl-CH_2 -
	Triptofano	Trp	W	204	
	Metionina	Met	M	149	CH_3-S-(CH_2)$_2$ -
Polares neutros	Glicina	Gly	G	75	H -
	Serina	Ser	S	105	HO-CH_2 -
	Treonina	Thr	T	119	CH_3-CH(OH) -
	Cisteína	Cys	C	121	HS-CH_2 -
	Tirosina	Tyr	Y	181	4-OH-Phenyl-CH_2 -
	Asparagina	Asn	N	132	H_2N-CO-CH_2 -
	Glutamina	Gln	Q	146	H_2N-CO-(CH_2)$_2$ -
Básicos	Lisina	Lys	K	146	H_2N-(CH_2)$_4$ -
	Arginina	Arg	R	174	HN=C(NH_2)-NH-(CH_2)$_3$ -
	Histidina	His	H	155	
Acídicos	Ácido aspártico	Asp	D	133	HOOC-CH_2 -
	Ácido glutâmico	Glu	E	147	HOOC-(CH_2)$_2$ -

[2] Em daltons. 1 dalton = massa de um átomo de hidrogénio = 1.67×10^{-24} g.

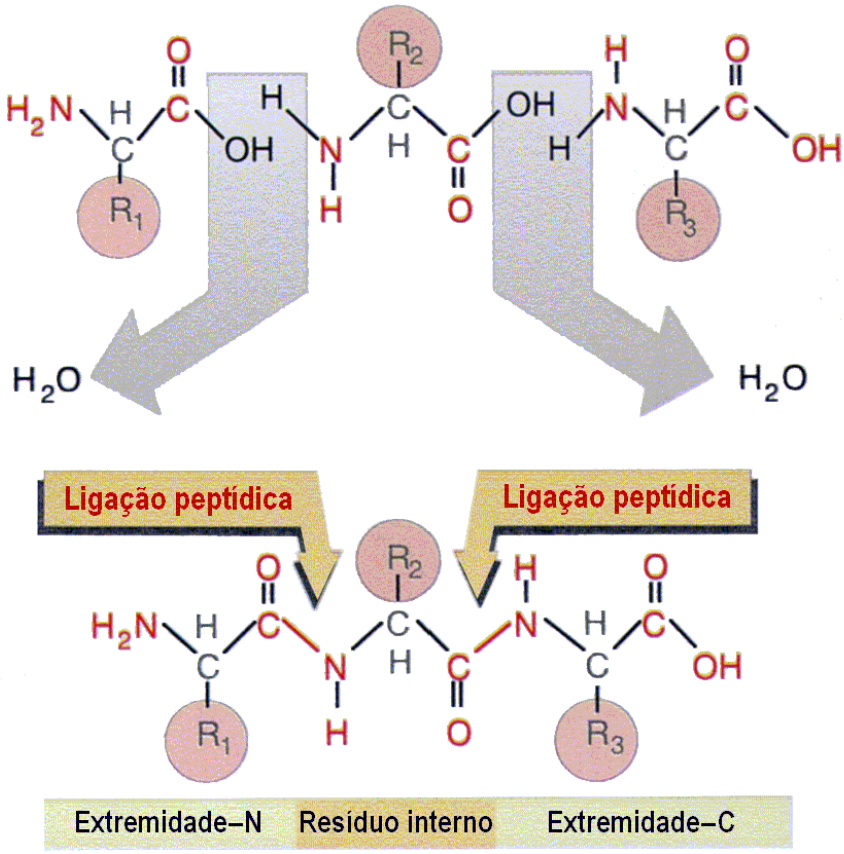


Figura 2.3 – Formação de ligações peptídicas e cadeia polipeptídica resultante. [3]

Ao primeiro aminoácido da cadeia, que tem o grupo amínico livre, chama-se *extremidade N–*, ou *amínica*; ao último, que tem o grupo carboxílico livre, chama-se *extremidade C–*, ou *carboxílica* (figura 2.3). A *estrutura primária* de uma proteína consiste na sequência de aminoácidos da sua cadeia polipeptídica, representada no sentido da extremidade N– para a extremidade C–. Caso a proteína seja formada por várias cadeias, a estrutura primária consiste nas respectivas sequências. A figura 2.4 representa a estrutura primária de uma proteína, denominada proteína G. Os resíduos destacados constituem o domínio B1.

```
          10         20         30         40         50
  1 MEKEKKVKYF LRKSAFGLAS VSAAFLVGST VFAVDSPIED TPIIRNGGEL
 51 TNLLGNSETT LALRNEESAT ADLTAAAVAD TVAAAAAENA GAAAWEAAAA
101 ADALAKAKAD ALKEFNKYGV SDYYKNLINN AKTVEGIKDL QAQVVESAKK
151 ARISEATDGL SDFLKSQTPA EDTVKSIELA EAKVLANREL DKYGVSDYHK
201 NLINNAKTVE GVKELIDEIL AALPKTDTYK LILNGKTLKG ETTTEAVDAA
251 TAEKVFKQYA NDNGVDGEWT YDDATKTFTV TEKPEVIDAS ELTPAVTTYK
301 LVINGKTLKG ETTTKAVDAE TAEKAFKQYA NDNGVDGVWT YDDATKTFTV
351 TEMVTEVPGD APTEPEKPEA SIPLVPLTPA TPIAKDDAKK DDTKKEDAKK
401 PEAKKDDAKK AETLPTTGEG SNPFFTAAAL AVMAGAGALA VASKRKED
```

Figura 2.4 – Estrutura primária.

[3] Adaptado de [Lewin 97].

2.2.2 Estrutura secundária

A cadeia polipeptídica não é de modo algum uma estrutura unidireccional. A simples presença de resíduos de prolina, devido à sua estrutura especial, e as diversas interacções químicas em que se envolvem os resíduos e o solvente, provocam inúmeras mudanças de direcção ao longo do esqueleto.

2.2.2.1 Interacções químicas

Uma *ligação de enxofre* consiste numa ligação covalente entre dois resíduos de cisteína, que podem encontrar-se bastante afastados na sequência, ou mesmo em cadeias polipeptídicas diferentes. A figura 2.5 ilustra a formação de uma ligação de enxofre entre dois resíduos de uma cadeia.

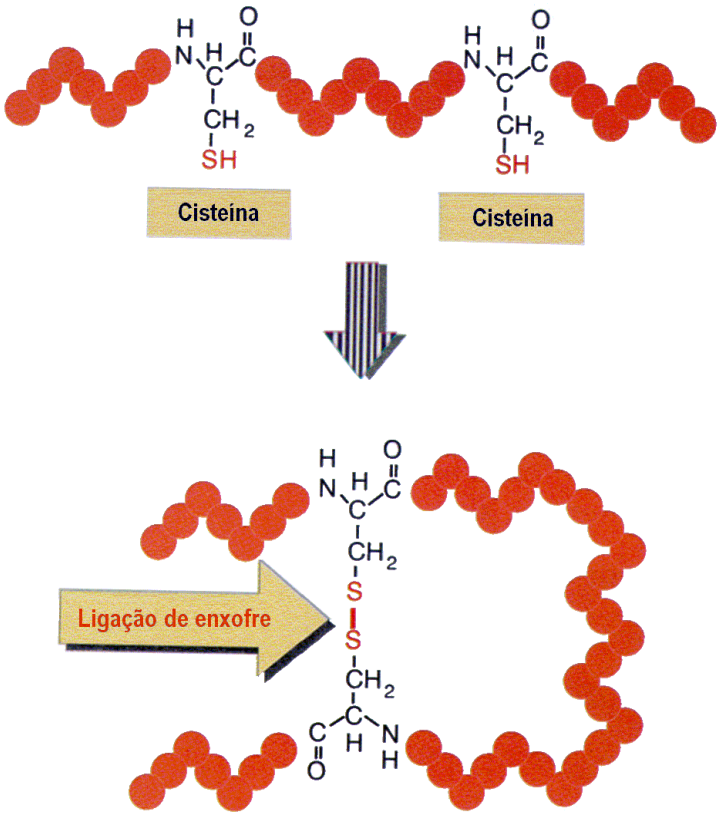


Figura 2.5 – Formação de uma ligação de enxofre. [4]

Uma *ligação de hidrogénio* é uma ligação electrostática entre um oxigénio e um hidrogénio. Ocorre entre as cadeias laterais dos aminoácidos polares, entre as cadeias laterais e o solvente (note-se que também as moléculas de água formam uma rede de ligações de hidrogénio), e no próprio esqueleto da cadeia. Embora seja uma ligação não covalente, é tão comum que contribui significativamente para a estabilidade da proteína.

[4] Adaptado de [Lewin 97].

Outras interacções não covalentes incluem as *interacções iónicas*, que ocorrem entre cadeias laterais de cargas opostas, cuja força é semelhante à das ligações de hidrogénio, e as *atracções de van der Waals*, interacções muito fracas que ocorrem entre átomos muito próximos.

Finalmente, a *hidrofobia* também influencia significativamente a conformação da cadeia polipeptídica. Como não podem participar na rede de ligações de hidrogénio da água, os resíduos hidrofóbicos são forçados a formar aglomerados cujo formato minimiza o contacto com o solvente.

2.2.2.2 Motivos comuns – hélice α e folha β

Em todas as conformações que o esqueleto pode apresentar, alguns motivos destacam-se pela frequência com que ocorrem; a sua identificação ao longo da cadeia polipeptídica constitui a *estrutura secundária* da proteína. São dois os motivos mais comuns, designados por *hélice a* e *folha b*.

2.2.2.2.1 Hélice α

Numa hélice α o esqueleto da cadeia polipeptídica forma uma estrutura helicoidal com 3.6 resíduos em cada volta, estabilizada por ligações de hidrogénio entre cada 4 resíduos, e onde todas as cadeias laterais se encontram viradas para fora. A figura 2.6 mostra três representações diferentes da hélice α.

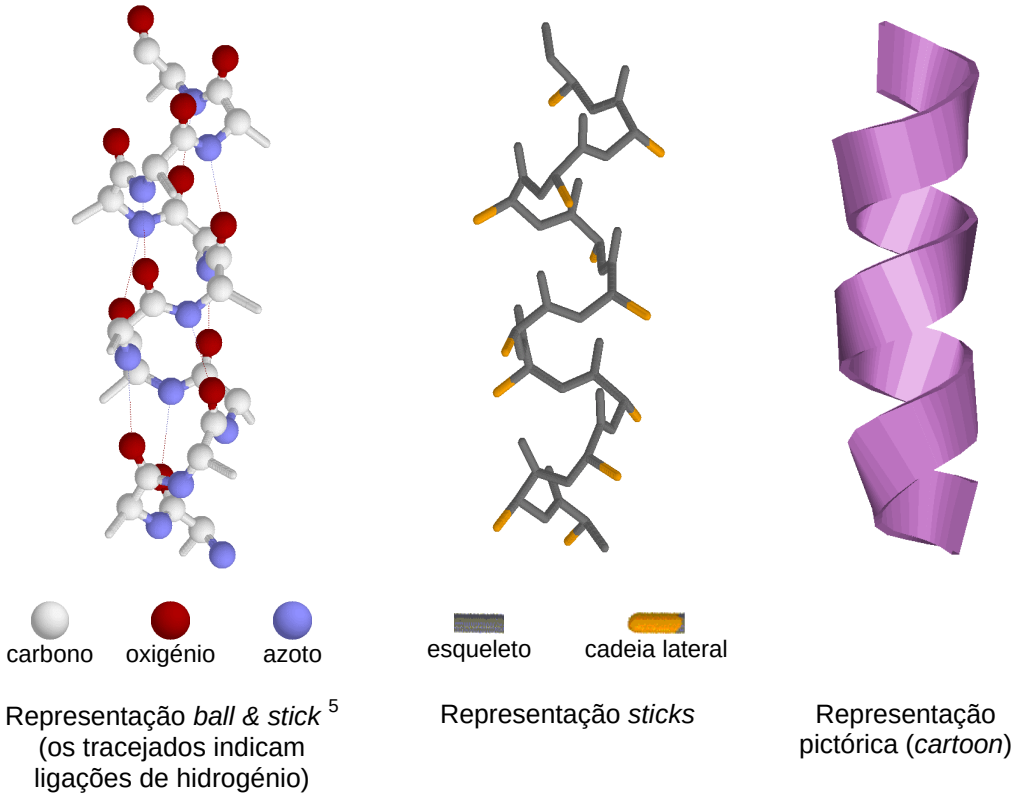


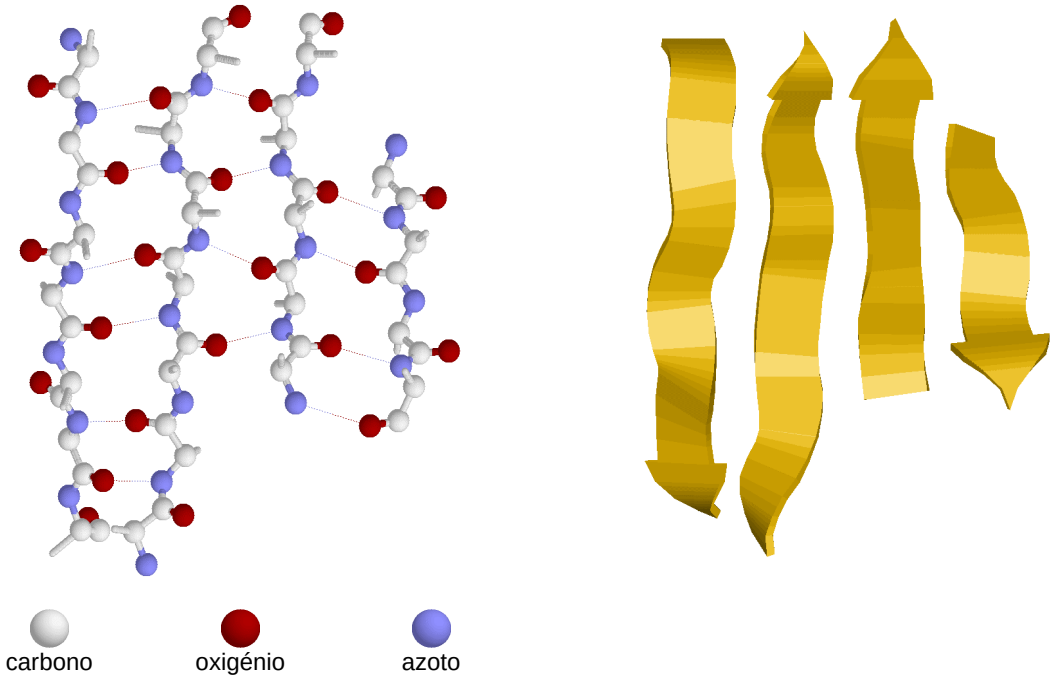
Figura 2.6 – Hélice α.

[5] *Ball & stick*, *sticks* e *cartoon* são designações de modos de representação de proteínas usadas nos programas de visualização molecular Rasmol e Chime, disponíveis no endereço http://www.umass.edu/microbio/rasmol.

Existem outros tipos de hélice, nomeadamente a *hélice p* e a *hélice* 3_{10}, menos estáveis e muito menos comuns do que a hélice α.

2.2.2.2.2 Folha β

Numa folha β, diferentes segmentos do esqueleto de uma cadeia polipeptídica, ou de cadeias diferentes, encontram-se ligados por ligações de hidrogénio em que participam todos os resíduos, formando uma estrutura planar onde as cadeias laterais se encontram viradas para cima e para baixo, e nunca interagem umas com as outras. Consoante a orientação relativa dos segmentos da folha β, esta recebe a classificação de *paralela* (segmentos todos orientados na mesma direcção), *antiparalela* (segmentos adjacentes orientados em direcções opostas) ou *mista*. A figura 2.7 mostra duas representações diferentes da folha β mista.



Representação *ball & stick*
(os tracejados indicam ligações
de hidrogénio)

Representação pictórica
(as setas indicam a direcção
da sequência)

Figura 2.7 – Folha β mista.

2.2.3 Estruturas terciária e quaternária

A *estrutura terciária* de uma proteína consiste no arranjo tridimensional de todos os átomos que a compõem. A figura 2.8 mostra, em estereoscopia[6], uma representação do domínio B1 da proteína G, cuja estrutura primária foi apresentada na figura 2.4 (página 8).

[6] Todos os pares estereoscópicos aqui apresentados devem ser visualizados usando a técnica de observação cruzada.

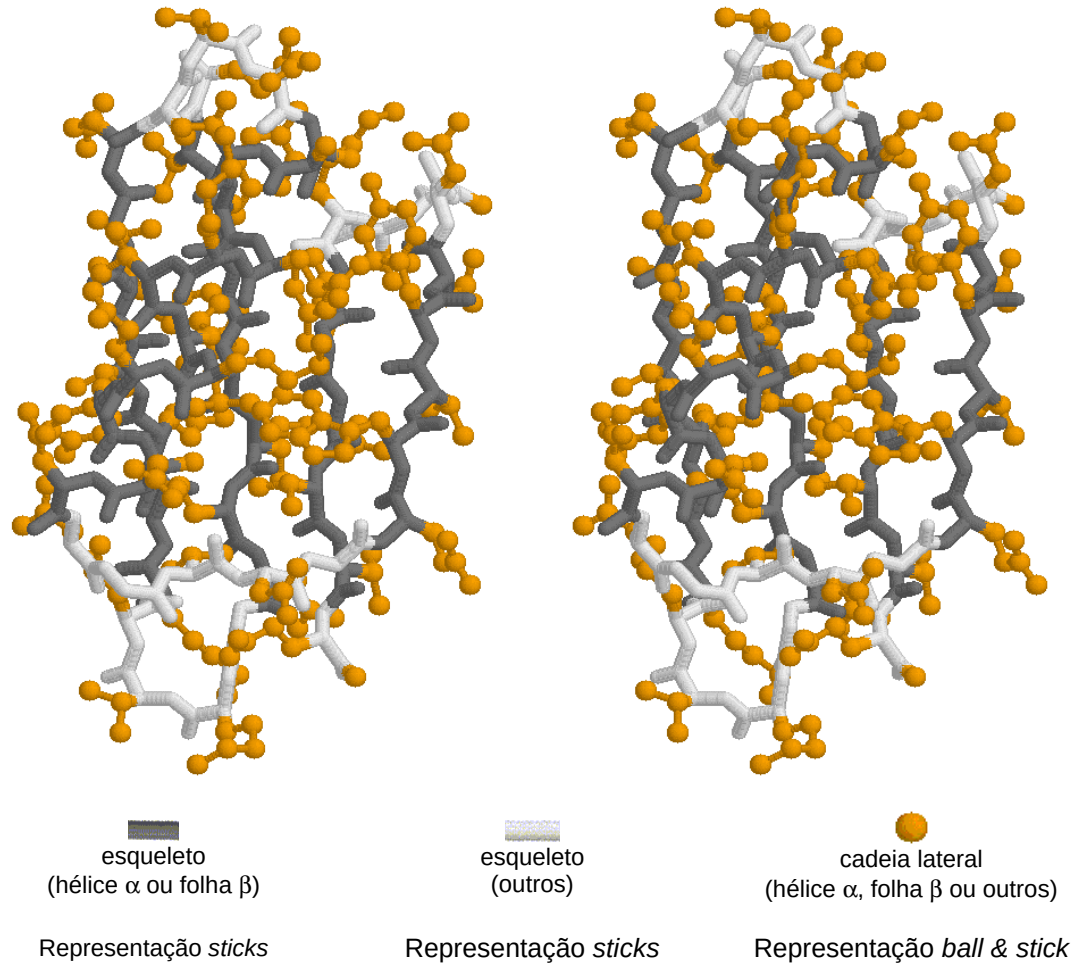


Figura 2.8 – Estrutura terciária, em estereoscopia.

Muitas vezes é extremamente útil visualizar as estruturas secundária e terciária no mesmo modelo. Para tal prescinde-se da descrição das cadeias laterais e utiliza-se as representações pictóricas da hélice α e da folha β, apresentadas nas figuras 2.6 e 2.7 (páginas 10 e 11), respectivamente. O resultado encontra-se ilustrado na figura 2.9, que representa o domínio B1 da proteína G, em estereoscopia. Esta é a forma mais comum de representação das estruturas secundária e terciária de proteínas.

A *estrutura quaternária* existe apenas quando a proteína é *oligomérica*, *i.e.*, composta por mais do que uma cadeia polipeptídica, e consiste nas suas relações e disposições relativas. Dependendo da sua estrutura terciária ou quaternária, uma proteína pode ser classificada como *fibrosa* (cadeias polipeptídicas dispostas ao longo de um eixo, formando uma estrutura alongada) ou *globular* (cadeias polipeptídicas muito compactas, formando uma estrutura esférica). A figura 2.10 mostra a conformação da hemoglobina humana, proteína globular constituída por quatro cadeias polipeptídicas.

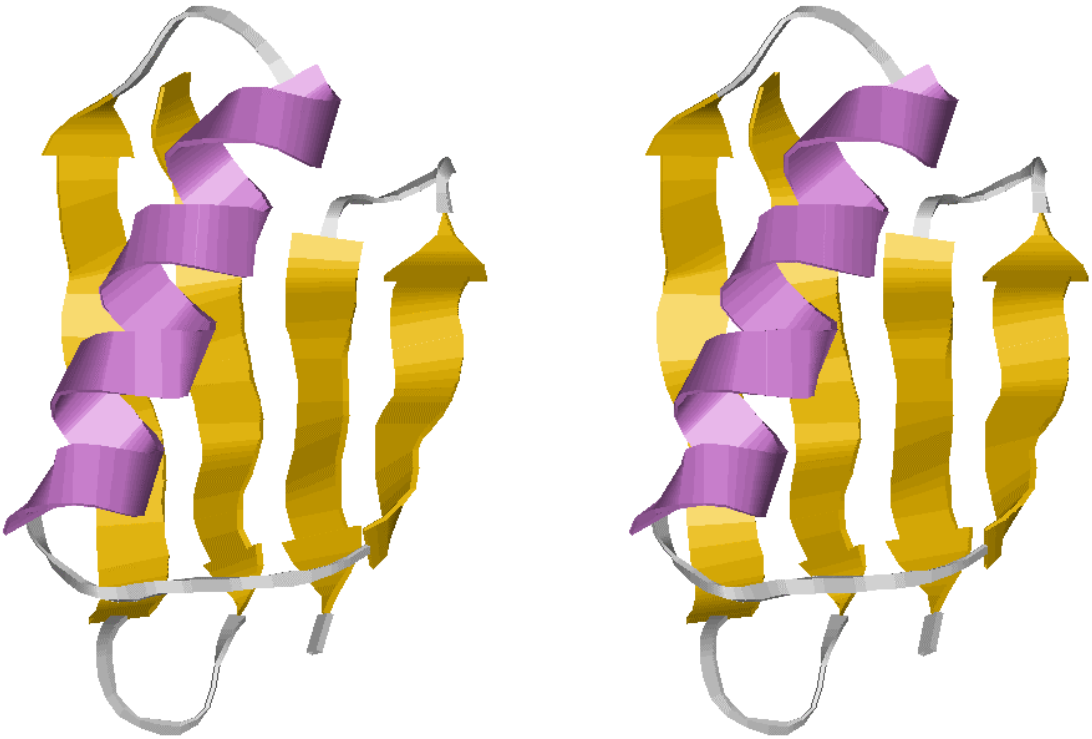


Figura 2.9 – Estruturas secundária e terciária, em estereoscopia.

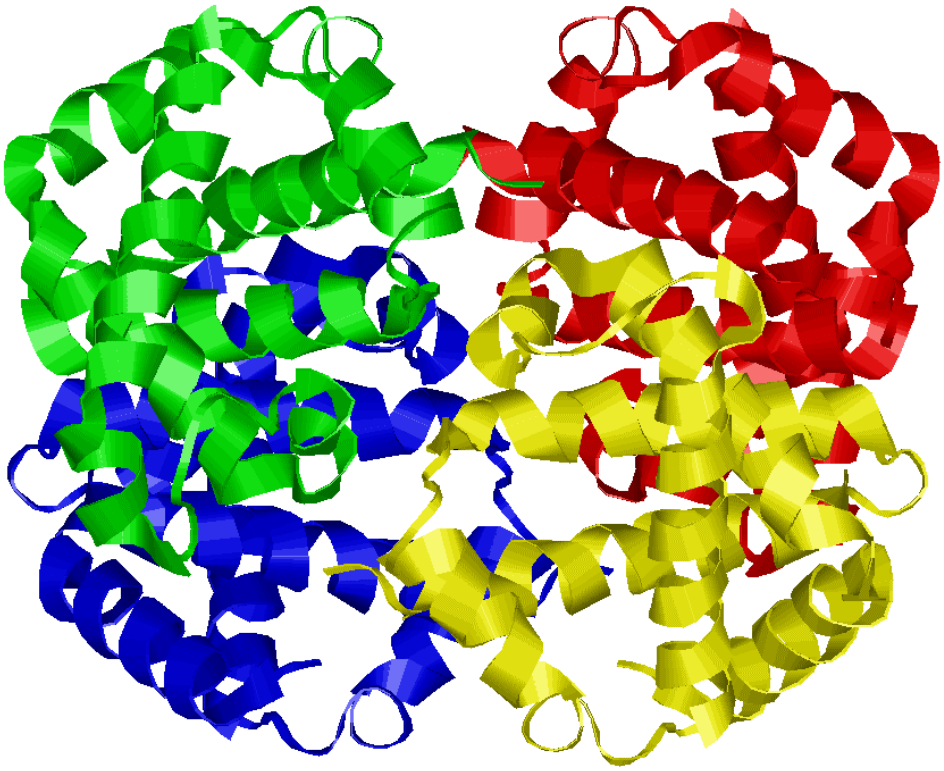


Figura 2.10 – Estrutura quaternária.

2.3 Classificação estrutural

Consoante a sua estrutura espacial, as proteínas podem ser catalogadas em quatro classes, representadas por *a/a*, *b/b*, *a/b* e *a+b*. Frequentemente, domínios diferentes da mesma proteína pertencem a classes distintas. Algumas proteínas não podem ser classificadas em nenhuma destas classes, ou porque a sua sequência é demasiado curta, ou porque no seu esqueleto não se observa praticamente nenhum motivo de estrutura secundária.

2.3.1 Classe α/α

As proteínas pertencentes à classe estrutural α/α são formadas quase exclusivamente por hélices α, com as eventuais folhas β localizadas na periferia da proteína. A hemoglobina humana, apresentada na figura 2.10 (página 13), é um bom exemplo de uma proteína α/α.

2.3.2 Classe β/β

As proteínas classificadas como β/β são constituídas quase exclusivamente por folhas β, principalmente antiparalelas, com as eventuais hélices α localizadas na periferia. A figura 2.11 representa uma proteína β/β.

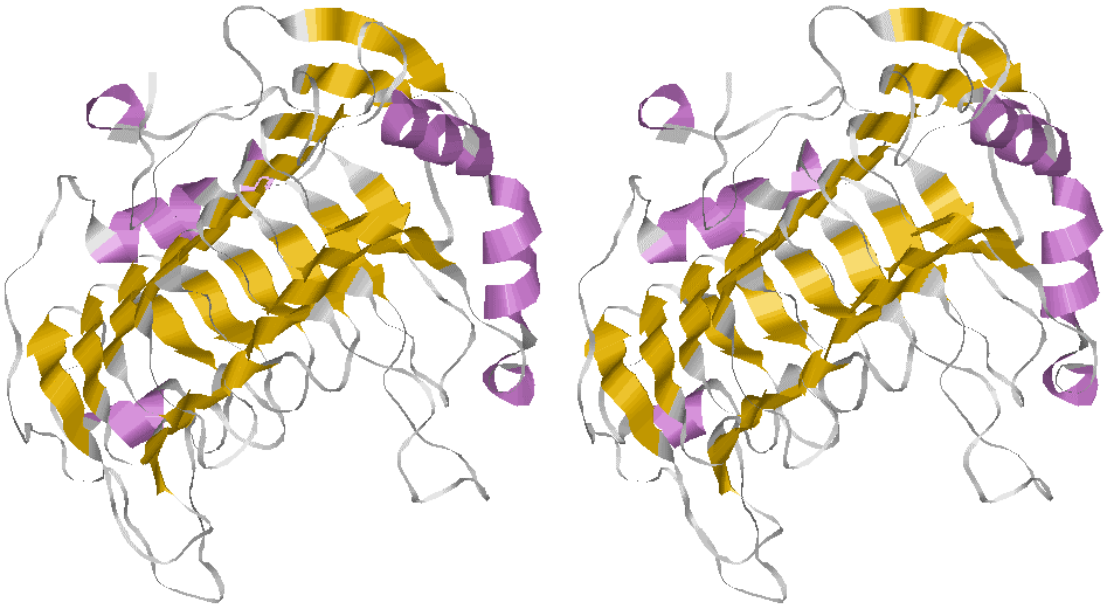


Figura 2.11 – Proteína β/β, em estereoscopia.

2.3.3 Classe α/β

As proteínas pertencentes à classe α/β apresentam uma alternância acentuada de hélices α e folhas β ao longo da sequência, dispostas de tal forma que as folhas β,

tipicamente paralelas, formam um aglomerado central rodeado por hélices α. A figura 2.12 representa uma proteína α/β.

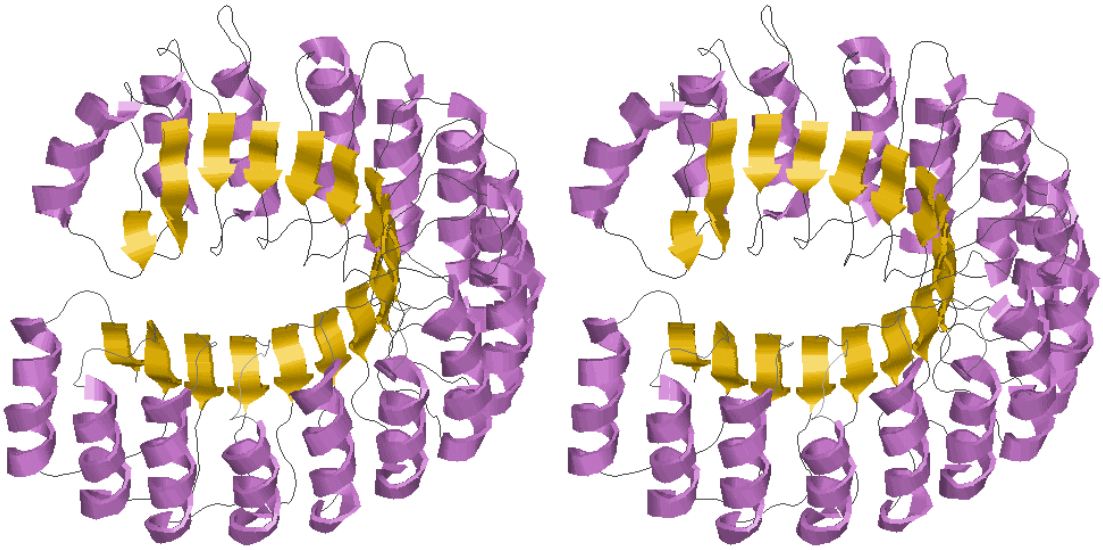


Figura 2.12 – Proteína α/β, em estereoscopia.

2.3.4 Classe α+β

A classe α+β inclui as proteínas que, sendo formadas por um número significativo de hélices α e folhas β, não são dominadas por nenhum dos motivos, nem apresentam a alternância observada na classe α/β. O domínio B1 da proteína G, apresentado na figura 2.9 (página 13), pertence à classe α+β.

2.4 Homologia

Quando os genes sofrem mutações, as proteínas que codificam podem sofrer alterações, sendo as mais comuns substituições, inserções e deleções pontuais de aminoácidos, em qualquer ponto da sequência. Algumas proteínas apresentam um grupo de aminoácidos essenciais à sua estrutura e função, denominado *centro funcional* (ou *centro activo*, nas enzimas). Uma mutação que afecte o centro funcional de uma proteína quase sempre compromete seriamente, ou mesmo inviabiliza, a sua função. Como qualquer outra mutação que provoque uma desvantagem, tende a perder-se rapidamente. Por outro lado, as substituições entre aminoácidos semelhantes raramente afectam a conformação da proteína, sendo por isso muito comuns. De um modo geral, a conformação é mais importante do que a sequência, sendo portanto mais conservada evolutivamente.

Duas proteínas dizem-se *homólogas* quando partilham um antepassado comum. É frequente afirmar-se que existe homologia quando se observa mais do que *n*% de identidade entre as sequências, sendo *n* quase sempre 20, 25 ou 30. Esta regra, embora

incorrecta, revela-se extremamente útil quando a única informação disponível é a estrutura primária, pois é de facto improvável que duas proteínas com sequências muito parecidas tenham evoluído independentemente.

Chama-se *alinhamento* de proteínas ao arranjo de sequências em que os resíduos alinhados correspondem ao mesmo resíduo num antepassado comum. Embora um alinhamento possa utilizar apenas duas sequências, um alinhamento múltiplo é mais fiável do ponto de vista biológico, e pode conter muito mais do que informação evolutiva. Nomeadamente, pode revelar a localização de centros funcionais de proteínas homólogas, identificados por um ou mais grupos de resíduos consecutivos muito conservados.

2.5 Determinação da estrutura

2.5.1 Ineficiência dos métodos experimentais

Cristalografia de raios X e espectroscopia multidimensional de ressonância magnética nuclear (NMR) são os dois métodos experimentais usados na determinação da estrutura de proteínas. No entanto, nenhum deles consegue acompanhar o rápido crescimento do número de sequências conhecidas, devido a dificuldades em conseguir purificar e cristalizar proteínas em quantidades suficientes, o que resulta numa diferença crescente entre o número de sequências e o número de estruturas disponíveis nas bases de dados públicas.

A figura 2.13 mostra os gráficos de crescimento do número de sequências anotadas disponíveis no SWISS-PROT [Bairoch e Apweiler 99], e do número de estruturas disponíveis no PDB (*Protein Data Bank*) [Bernstein *et al.* 77], entre 1986 e 1998. As edições do SWISS-PROT foram lançadas a intervalos mais ou menos regulares entre Setembro de 1986 e Dezembro de 1998. As actualizações do PDB podem sempre contemplar sobreposições e remoções de estruturas. No dia 28 de Julho de 1999, o número de estruturas era 10406 e o de sequências 80000 (edição 38), sem esquecer as 199805 sequências contidas numa base de dados suplementar ao SWISS-PROT, denominada TrEMBL, que aguardavam anotação para serem também admitidas no SWISS-PROT.

2.5.2 Métodos de previsão da estrutura secundária

Dada a grande dificuldade em conseguir determinar de forma experimental a estrutura espacial de proteínas, tem-se tentado desenvolver formas de prevê-la a partir da sequência. Partindo do pressuposto fundamental de que a conformação é determinada pela sequência, muitos dos métodos desenvolvidos até ao momento centram-se naquilo que parece ser a tarefa mais simples: a previsão da estrutura secundária.

2.5.2.1 Chou-Fasman

O primeiro método de previsão de estrutura secundária de proteínas largamente utilizado foi desenvolvido por Chou e Fasman [Chou e Fasman 74a, 74b], e baseia-se no cálculo das probabilidades de cada resíduo se encontrar numa hélice α ou numa folha β. Apesar de ser muito fácil de usar e de compreender, as estatísticas que usa

são algo duvidosas e os resultados da previsão bastante maus, com uma exactidão que não vai além dos 50%. O método de Chou-Fasman apenas utiliza a informação fornecida por cada resíduo de cada sequência, considerada independentemente dos outros resíduos e de outras sequências, e por tal recebe a designação de método de primeira geração.

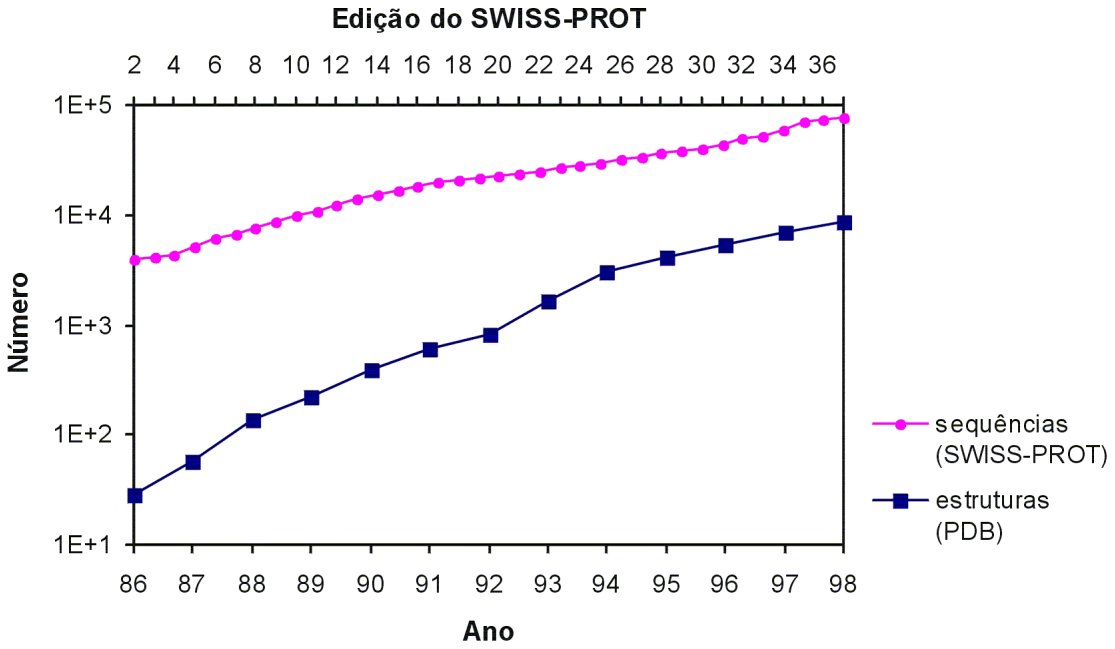


Figura 2.13 – Crescimento do número de sequências e de estruturas disponíveis.

2.5.2.2 GOR

Bastante mais bem sucedido é o método GOR (Garnier-Osguthorpe-Robson) [Garnier *et al.* 78, Gibrat *et al.* 87], baseado na ideia de que a previsão da estrutura secundária não é mais do que um processo de tradução de uma mensagem (estrutura primária) para outra (estrutura secundária). Estatísticas bem fundamentadas resultam num método robusto e teoricamente correcto, cuja terceira versão (GOR III) atinge níveis de exactidão um pouco acima dos 60%. Ao contrário de GOR I, que é um método da primeira geração, GOR III já utiliza informação sobre interacções locais entre resíduos, sendo por isso considerado um método de segunda geração.

2.5.2.3 PHD

Possivelmente o mais bem sucedido método de previsão de estrutura secundária disponível até ao momento, o PHD (*Profile network from HeiDelberg*), mais precisamente, PHDsec [Rost e Sander 93, 94], possui também o mérito de ser considerado o primeiro método de terceira geração, pois introduz informação evolutiva contida em alinhamentos na previsão de estrutura secundária de proteínas. Assim, ao receber uma sequência para classificar, a primeira prioridade do PHD consiste em obter um alinhamento múltiplo construído com base em sequências homólogas disponíveis no SWISS-PROT, operação levada a cabo pelo programa auxiliar MaxHom [Sander e Schneider 91].

O PHDsec é um programa composto por quatro níveis de processamento, dois dos quais consistem em perceptrões multicamada (ver secção 3.2) treinadas com proteínas de estrutura conhecida. O primeiro nível recebe vectores referentes a segmentos de 13 resíduos consecutivos no alinhamento, e devolve valores indicativos da verosimilhança de o resíduo central se encontrar numa hélice, folha ou outro motivo estrutural.

O segundo nível de redes neuronais recebe os valores provenientes do primeiro nível, e alguma informação global sobre a proteína em questão, e devolve valores de significado idêntico aos do primeiro nível. O valor mais elevado determina a classificação atribuída ao resíduo central, e a diferença entre os dois valores mais elevados é utilizada como índice de fiabilidade, indicando o grau de confiança que o programa tem na classificação atribuída.

Várias redes, treinadas independentemente, fazem a classificação de todos os resíduos da proteína, e o terceiro nível de computação consiste simplesmente em escolher, para cada resíduo, a classificação que apresenta a soma de índices de fiabilidade mais elevado.

Finalmente, o quarto e último nível consiste em submeter a classificação obtida a um filtro que resolve incorrecções óbvias como, por exemplo, hélices com menos de três resíduos de comprimento.

O PHD atinge uma percentagem de exactidão média de 70%, valor que sobe acima dos 80% quando somente a metade dos resíduos classificada com maior fiabilidade é considerada. Encontra-se prontamente disponível para utilização no servidor PredictProtein[7] [Rost *et al.* 94a].

[7] Acessível a partir do endereço http://www.embl-heidelberg.de/predictprotein.

3 REDES NEURONAIS

Este capítulo tem como tema os dois tipos de redes neuronais utilizados neste trabalho. Após a introdução de alguns conceitos fundamentais relacionados com redes neuronais, segue-se uma secção dedicada ao perceptrão multicamada, onde são abordados os temas da arquitectura e aprendizagem, de um modo deliberadamente abreviado. Segue-se uma secção análoga, dedicada ao mapa de Kohonen.

3.1 Fundamentos

Uma *rede neuronal artificial* é um sistema de processamento de dados, inspirado nas redes neuronais biológicas, que consiste num conjunto de unidades processadoras muito simples, denominadas *neurónios formais*, que comunicam entre si através de sinapses artificiais com impedâncias variáveis associadas, designadas por *pesos sinápticos*. A forma como os neurónios se encontram conectados, o tipo de processamento que efectuam e o modo como os pesos sinápticos são determinados, *i.e.*, o algoritmo de aprendizagem que a rede utiliza, definem modelos bem distintos, adequados à realização de diferentes tarefas como memorização, reconhecimento e classificação de dados, controlo, previsão de séries temporais e optimização, com aplicação em áreas tão diversas como biologia, medicina, robótica, telecomunicações, educação e economia.

Os diferentes modos de aprendizagem adoptados pelas redes neuronais podem classificar-se em dois tipos distintos: supervisionado e não supervisionado. O modo supervisionado pressupõe uma fase de treino, em que a rede neuronal é alimentada com pares <estímulo, resposta> representativos da forma como diferentes estímulos devem ser agrupados em diferentes classes. Terminada esta fase, a rede deve encontrar-se pronta a devolver respostas correctas perante estímulos que nunca observou, *i.e.*, deve demonstrar uma boa capacidade de generalização. No modo não supervisionado esta fase de treino é substituída por uma aprendizagem espontânea, que revela uma classificação natural dos estímulos mediante as suas semelhanças.

3.2 Perceptrão multicamada

O *perceptrão multicamada*, frequentemente designado por *rede progressiva* (*feedforward*), é uma arquitectura neuronal de aprendizagem supervisionada em que os neurónios se encontram organizados em várias camadas.

3.2.1 Arquitectura

Num perceptrão multicamada existe sempre uma camada de entrada, cujos neurónios se denominam *sensores*, uma ou mais camadas internas, e uma camada de saída, cujos neurónios se denominam *efectores*. Entre camadas sucessivas existe conexão sináptica total ou parcial. A figura 3.1 mostra a arquitectura de um perceptrão multicamada, onde os neurónios são representados por círculos e as sinapses por setas, que indicam o sentido das conexões. A camada de entrada não é contabilizada no número de camadas do perceptrão, pois os sensores não efectuam qualquer tipo de processamento, limitando-se a receber os estímulos e transmiti-los à camada seguinte. Cada um dos neurónios internos ou de saída, denominados *processadores*, tem associado um pendor, valor que actua como um peso sináptico proveniente de um neurónio fictício (frequentemente encarado como o primeiro neurónio da camada anterior) cuja resposta é sempre unitária.

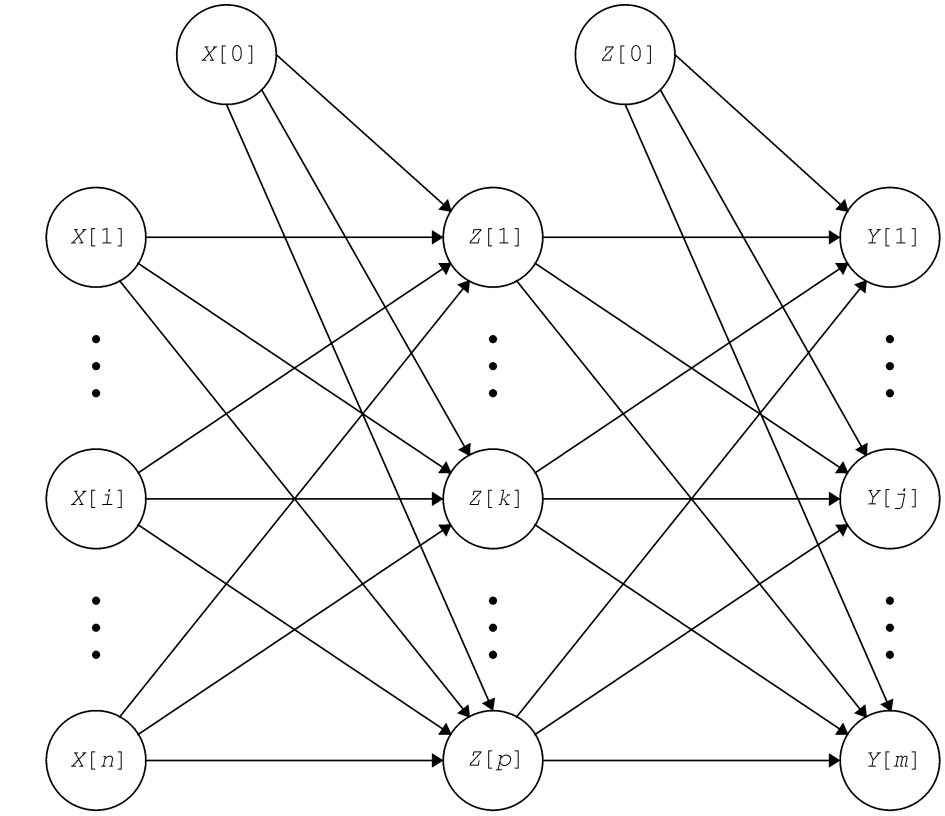
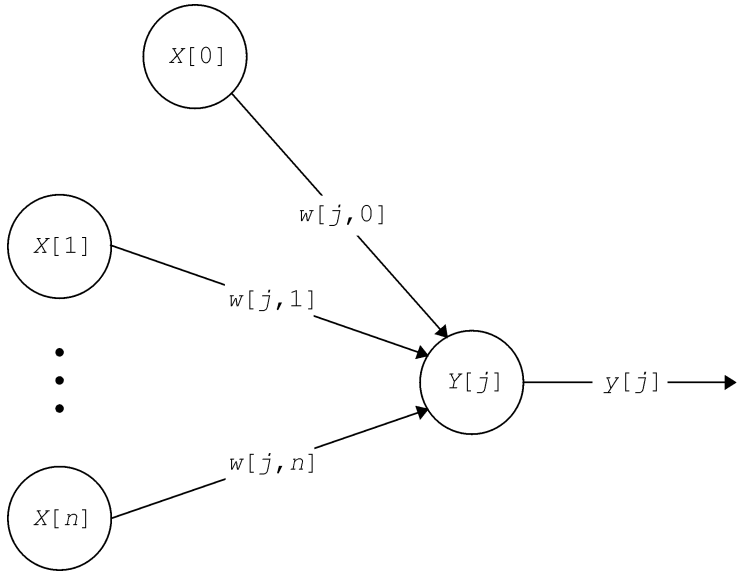


Figura 3.1 – Perceptrão multicamada.

A figura 3.2 representa um neurónio processador genérico. Os neurónios processadores calculam a sua resposta, ou sinal, efectuando a soma das respostas dos neurónios da camada anterior ponderadas pelos respectivos pesos sinápticos. Ao resultado é aplicada uma função de activação não linear e diferenciável em todo o seu domínio. Um estímulo apresentado aos sensores é assim propagado por camadas sucessivas até aos efectores, onde a rede exprime o resultado. A resposta da rede é, pois, uma função do estímulo e dos pesos sinápticos.



$y[j] = f(y_{in}[j]) = f\left(\sum_{i=0}^{n} w[j,i]x[i]\right)$ ($f \rightarrow$ função de activação)

$w[j,i] \rightarrow$ peso sináptico entre $X[i]$ e $Y[j]$ $(0 \leq i \leq n)$ ($w[j,0] \rightarrow$ pendor de $Y[j]$)

$x[i] \rightarrow$ sinal enviado por $X[i]$ $(0 \leq i \leq n)$ ($x[0] = 1$)

Figura 3.2 – Neurónio processador genérico.

3.2.2 Aprendizagem

3.2.2.1 Algoritmo

A *retropropagação* [Rumelhart *et al*. 86, Werbos 74] é o algoritmo tipicamente usado no treino do perceptrão multicamada. O seu objectivo é encontrar o conjunto de pesos sinápticos que minimizam uma função de erro, baseada na diferença entre a resposta devolvida pela rede e a resposta pretendida, para cada estímulo apresentado. Para tal, implementa um método iterativo de gradiente descendente baseado no cálculo das derivadas parciais da função de erro. Cada iteração do algoritmo é induzida pela apresentação de um par <estímulo, resposta> e processa-se em duas fases distintas: a propagação dos sinais e a retropropagação do erro, esta acompanhada pela respectiva alteração dos pesos sinápticos. Chama-se *época* a um conjunto de Q iterações sucessivas, sendo Q o número de estímulos utilizados no treino da rede. Na

especificação do algoritmo é usada a notação introduzida nas figuras 3.1 e 3.2, e os símbolos $\boldsymbol{s}$ e $\boldsymbol{t}$ para designar, respectivamente, o estímulo e a resposta pretendida.

Algoritmo de retropropagação para uma camada interna:

0. Inicializar pesos (e pendores) da camada de entrada para a camada interna

$w[k,i]$ $(k = 1..p\ ,\ i = 0..n)$

e da camada interna para a camada de saída

$v[j,k]$ $(j = 1..m\ ,\ k = 0..p)$

Inicializar valores dos parâmetros do algoritmo:

coeficiente de aprendizagem: h $(0 < h \leq 1)$

1. Repetir até certa condição de paragem ser satisfeita

Para cada par $<\boldsymbol{s},\boldsymbol{t}>$ executar de **1.1** a **1.6**

$(\boldsymbol{s} = \langle s[1],\ldots,s[n]\rangle$ e $\boldsymbol{t} = \langle t[1],\ldots,t[m]\rangle)$

Propagação dos sinais:

1.1. Activar camada de entrada:

$$x[0] = 1$$

$$x[i] = s[i] \quad (i = 1..n)$$

1.2. Activar camada interna:

$$z[0] = 1$$

$$z_{in}[k] = \sum_{i=0}^{n} w[k,i]x[i] \quad (k = 1..p)$$

$$z[k] = f(z_{in}[k]) \quad (k = 1..p)$$

1.3. Determinar resposta da rede:

$$y_{in}[j] = \sum_{k=0}^{p} v[j,k]z[k] \quad (j = 1..m)$$

$$y[j] = f(y_{in}[j]) \quad (j = 1..m)$$

Retropropagação do erro:

1.4. Calcular erro e ajustamento na camada de saída:

$$d_Y[j] = f'(y_{in}[j])(t[j] - y[j]) \quad (j = 1..m)$$

$$\Delta v[j,k] = hd_Y[j]z[k] \quad (j = 1..m\ ,\ k = 0..p)$$

1.5. Calcular erro e ajustamento na camada interna:

$$d_Z[k] = f'(z_{in}[k])\sum_{j=1}^{m} d_Y[j]v[j,k] \quad (k = 1..p)$$

$$\Delta w[k,i] = hd_Z[k]x[i] \quad (k = 1..p\ ,\ i = 0..n)$$

1.6. Ajustar pesos sinápticos:

$$v[j, k] = v[j, k] + \Delta v[j, k] \quad (j = 1..m \,,\, k = 0..p)$$

$$w[k, i] = w[k, i] + \Delta w[k, i] \quad (k = 1..p \,,\, i = 0..n)$$

2. Testar a condição de paragem.

3.2.2.2 Elementos do algoritmo

3.2.2.2.1 Função de erro

A função de erro usada na derivação do algoritmo apresentado, aqui designada genericamente por erro quadrático, é especificada por

$$E_q = \frac{1}{2} \sum_{j=1}^{m} (t_q[j] - y_q[j])^2 \,,$$

onde $\boldsymbol{t}_q$ e $\boldsymbol{y}_q$ representam a resposta pretendida e a resposta devolvida pela rede no q-ésimo estímulo, respectivamente.

Esta função define uma paisagem de erro multidimensional, na qual o algoritmo tenta convergir para o mínimo global. O ponto de partida, definido pelos pesos sinápticos iniciais, não só influencia a velocidade de convergência, como pode mesmo significar a diferença entre encontrar o mínimo global ou ficar preso num dos (provavelmente muitos) mínimos locais.

3.2.2.2.2 Inicialização dos pesos

Se os pesos sinápticos iniciais forem demasiado elevados, a função de activação, geralmente uma sigmóide, tende a saturar, e as respectivas derivadas tendem a anular-se. Se os pesos forem demasiado reduzidos, os neurónios processadores recebem sinais muito fracos das camadas anteriores. Qualquer um dos casos pode tornar a aprendizagem muito lenta. Os pesos sinápticos iniciais devem pois ser valores aleatórios uniformemente distribuídos num intervalo de valores pequenos, geralmente entre –0.5 e 0.5, ou entre –1 e 1.

Uma variação muito comum desta inicialização, proposta por Nguyen e Widrow [Nguyen e Widrow 90] para perceptrões de duas camadas, consiste na inicialização normal dos pesos, seguida de uma normalização entre as camadas de entrada e interna, o que resulta frequentemente em aprendizagens mais rápidas.

Inicialização de Nguyen-Widrow:

0.1. Inicializar pesos (e pendores) entre a camada de entrada e a camada interna

$$w[k, i] = \text{número aleatório entre } -l \text{ e } l \quad (k = 1..p \,,\, i = 0..n)$$

e entre a camada interna e a camada de saída

$$v[j, k] = \text{número aleatório entre } -l \text{ e } l \quad (j = 1..m \,,\, k = 0..p)$$

0.2. Calcular normas dos vectores $\boldsymbol{w}[k]$: $(\boldsymbol{w}[k] = \langle w[k,1], \ldots, w[k, n] \rangle)$

$$w_{norm}[k] = \|\boldsymbol{w}[k]\| = \sqrt{\sum_{i=1}^{n} (w[k,i])^2} \quad (k = 1..p)$$

0.3. Reinicializar pesos entre a camada de entrada e a camada interna

$$b = 0.7 \sqrt[n]{p}$$

$$w[k,i] = \frac{bw[k,i]}{w_{norm}[k]} \quad (k = 1..p \ , i = 1..n)$$

e pendores da camada interna

$w[k,0]$ = número aleatório entre $-b$ e b $\quad (k = 1..p)$

3.2.2.2.3 Função de activação

As funções de activação mais usadas no algoritmo de retropropagação são a sigmóide binária, representada pela função logística, e a sigmóide bipolar, geralmente representada pela tangente hiperbólica.

Função logística:	***Tangente hiperbólica:***
$f(x) = \frac{1}{1 + e^{-bx}}$	$f(x) = \tanh(bx) = \frac{e^{bx} - e^{-bx}}{e^{bx} + e^{-bx}}$
$f'(x) = bf(x)(1 - f(x))$	$f'(x) = b\left(1 - f^2(x)\right)$

O parâmetro b determina o declive da sigmóide, e consequentemente o intervalo de maior sensibilidade da função. Tipicamente $b = 1$.

3.2.2.2.4 Coeficiente de aprendizagem

Regra geral, o ajustamento de um peso sináptico atinge apenas uma determinada fracção do seu valor, especificada pelo coeficiente de aprendizagem h. Valores elevados de h podem resultar em aprendizagens rápidas, mas também podem impedir o algoritmo de convergir. Por outro lado, valores baixos aumentam o perigo de convergência num mínimo local. O valor ideal depende grandemente da natureza dos estímulos e varia consoante o ponto na paisagem de erro em que os pesos sinápticos de encontram, embora o algoritmo de retropropagação básico não explore esta possibilidade.

3.2.2.2.5 Condição de paragem

A terminação do algoritmo de retropropagação depende de uma condição de paragem, que pode ser tão elementar quanto o atingir de um determinado número de iterações, ou o erro se situar abaixo de um determinado valor. No entanto, critérios tão simples são geralmente de pouca utilidade, quando aplicados a problemas de classificação reais. Regra geral, é importante considerar a capacidade de generalização como factor indispensável ao bom desempenho da rede treinada. Isto significa que a rede deve ser

capaz de classificar estímulos que nunca lhe foram apresentados, com uma exactidão semelhante àquela atingida nos estímulos de treino.

Assim, a terminação do algoritmo deve obedecer ao procedimento proposto por Hecht-Nielsen [Hecht-Nielsen 90], que consiste na utilização de um conjunto adicional de estímulos (disjunto do conjunto de treino), denominado conjunto de teste, no qual o erro é medido periodicamente, interrompendo-se o algoritmo quando este começa a subir continuamente. Evita-se assim o ajustamento exagerado dos pesos sinápticos aos exemplos de treino, fenómeno causador da perda de capacidade de generalização da rede, conhecido por *overfitting*.

Embora o conjunto de teste não participe no treino da rede, determina de facto o ponto em que este é interrompido e, consequentemente, os valores dos pesos finais da rede. Por este motivo, e para garantir que a capacidade de generalização da rede é medida num conjunto de dados realmente independente, quando o volume de dados assim o permite utiliza-se um terceiro conjunto de dados, disjunto dos dois primeiros, denominado conjunto de validação.

O erro quadrático médio (MSE), especificado por

$$MSE = \frac{1}{2Q} \sum_{q=1}^{Q} \sum_{j=1}^{m} \left(t_q[j] - y_q[j]\right)^2 ,$$

é a medida mais frequentemente utilizada no cálculo do erro cometido pelo perceptrão multicamada num conjunto de Q estímulos.

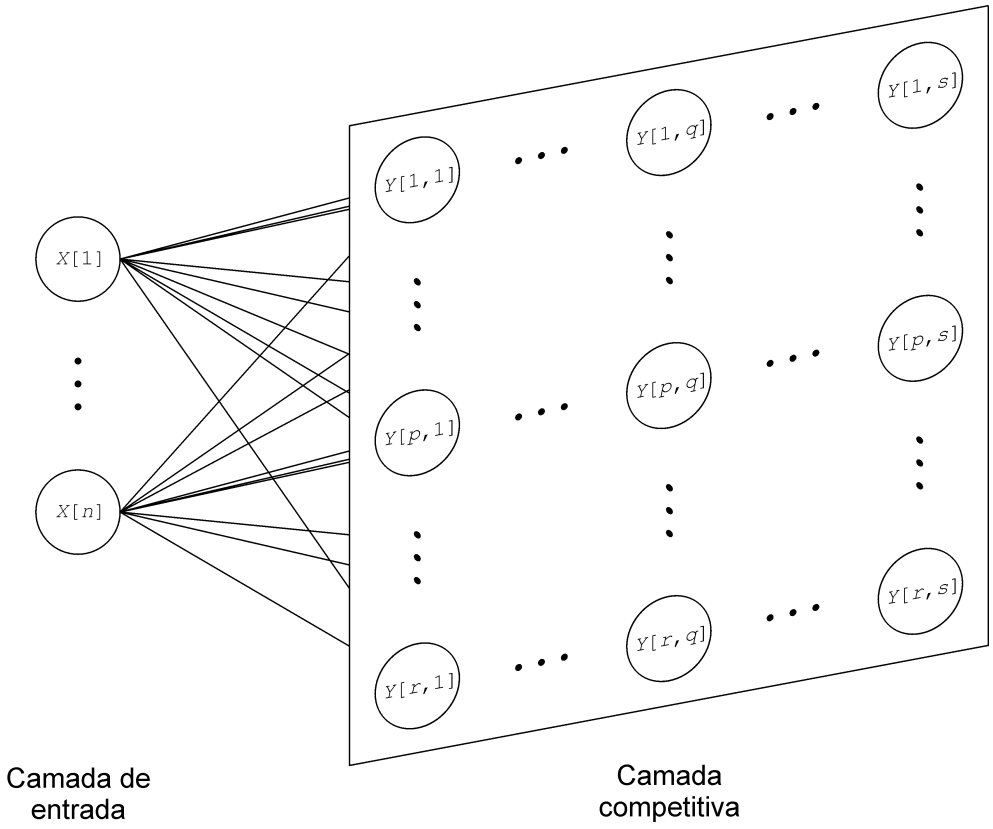
3.3 Mapa de Kohonen

O *mapa de Kohonen* [Kohonen 84] é uma arquitectura neuronal de aprendizagem não supervisionada de natureza competitiva.

3.3.1 Arquitectura

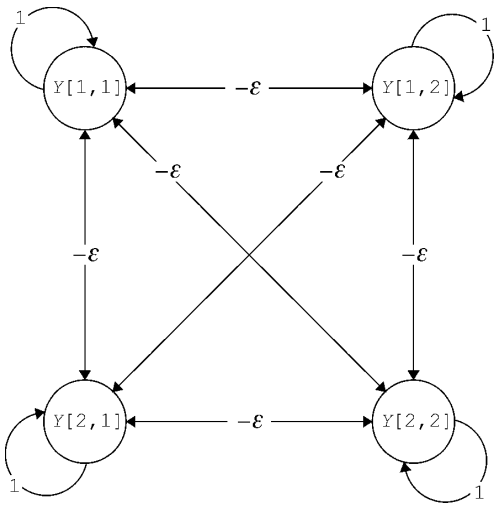
Num mapa de Kohonen existem duas camadas de neurónios, totalmente conectadas entre si: a camada de entrada, cujos neurónios se limitam a transmitir os estímulos que recebem, e a camada competitiva, organizada de modo a formar uma matriz de neurónios bidimensional. A figura 3.3 mostra um mapa de Kohonen de dimensão arbitrária.

Os neurónios da camada competitiva formam uma arquitectura MAXNET [Lippman 87], em que cada neurónio inibe os seus vizinhos, através de sinapses de peso negativo, enquanto se auto-excita, através de sinapses de peso positivo. Todos estes pesos são fixos. Um estímulo apresentado à camada de entrada é propagado a cada um dos neurónios da MAXNET através das sinapses que ligam as duas camadas. O neurónio cujos pesos sinápticos se assemelham mais ao estímulo produz o sinal mais forte, vencendo pois o jogo iterativo de inibições e excitações, no fim do qual apenas ele se encontra activo, representando a resposta da rede. A figura 3.4 mostra uma MAXNET formada por quatro neurónios.



$X[i]$ → neurónio de entrada $(0 \le i \le n)$
$Y[p,q]$ → neurónio competitivo $(1 \le p \le r\ ,\ 1 \le q \le s)$

Figura 3.3 – Mapa de Kohonen.



$-e$ → peso de sinapse inibitória $(e = 1/m)$
(m → número de neurónios da MAXNET)
1 → peso de sinapse excitatória

Figura 3.4 – MAXNET.

3.3.2 Aprendizagem

3.3.2.1 Algoritmo

O objectivo do algoritmo de aprendizagem do mapa de Kohonen é criar um mapeamento dos estímulos nos neurónios da camada competitiva. Para tal, o algoritmo implementa um método iterativo em que a apresentação de cada estímulo provoca a alteração dos pesos sinápticos entre o neurónio vencedor e os neurónios de entrada, de modo a aumentar a sua semelhança com o estímulo. Tipicamente, também os pesos dos neurónios na vizinhança do vencedor sofrem uma alteração análoga, obtendo-se assim um mapa contínuo em que neurónios topologicamente próximos representam estímulos de características semelhantes.

Na especificação do algoritmo, utiliza-se a notação introduzida na figura 3.3, mas substituindo a representação dos neurónios competitivos pela notação vectorial

$Y[j] = Y[p, q]$, com $j = s \times (p - 1) + q$.

$w[j, i]$ representa o peso sináptico da conexão entre os neurónios $X[i]$ e $Y[j]$.

Algoritmo de aprendizagem do mapa de Kohonen:

0. Inicializar pesos da camada de entrada para a camada competitiva

$w[j, i]$ $(j = 1..m , i = 1..n)$

Inicializar valores dos parâmetros do algoritmo:

formato e raio da vizinhança;

constante de decaimento do raio da vizinhança;

coeficiente de aprendizagem: h $(0 < h \leq 1)$;

tipo e factor de decaimento do coeficiente de aprendizagem

1. Repetir até certa condição de paragem ser satisfeita

Para cada estímulo ***s*** executar de **1.1** a **1.4**

1.1. Activar camada de entrada:

$x[i] = s[i]$ $(i = 1..n)$

1.2. Calcular distâncias:

$$d[j] = \sum_{i=1}^{n} (w[j, i] - x[i])^2 \quad (j = 1..m)$$

1.3. Determinar neurónio vencedor:

$Y[J]$ sendo J tal que $d[J]$ é mínima

1.4. Ajustar pesos sinápticos do neurónio vencedor e sua vizinhança:

$w[j, i] = w[j, i] + h(x[i] - w[j, i])$ $(i = 1..n)$

para todos os j tais que $Y[j]$ pertence à vizinhança de $Y[J]$

2. Reduzir coeficiente de aprendizagem.

3. Eventualmente reduzir raio da vizinhança.

4. Testar a condição de paragem.

3.3.2.2 Elementos do algoritmo

3.3.2.2.1 Inicialização dos pesos

A forma mais comum de inicializar os pesos sinápticos do mapa de Kohonen é atribuir-lhes valores aleatórios baixos, geralmente entre –0.5 e 0.5, ou entre –1 e 1. Como o algoritmo se baseia na comparação das distâncias entre o estímulo e os vectores de pesos sinápticos dos neurónios competitivos, a normalização destes vectores garante um aumento na velocidade de aprendizagem. No entanto, este procedimento não resolve um dos principais problemas que afectam a aprendizagem de natureza competitiva: a sub-utilização dos neurónios disponíveis.

Os neurónios cujos vectores de pesos iniciais são muito diferentes de qualquer um dos estímulos nunca serão vencedores, e os neurónios restantes podem não ser suficientes para permitir uma boa discriminação dos estímulos. A solução para o problema encontra-se precisamente na utilização de vizinhanças.

3.3.2.2.2 Parâmetros topológicos

A vizinhança de um neurónio competitivo constitui um elemento importante na eficiência do algoritmo de aprendizagem do mapa de Kohonen. O formato da vizinhança pode ser qualquer figura, sendo os mais comuns o círculo e o rectângulo, e costuma manter-se constante ao longo de toda a aprendizagem. O raio da vizinhança, que representa a distância (no caso do círculo, euclideana) entre o neurónio central e a fronteira da vizinhança, determina a dimensão da figura, e deve sofrer reduções periódicas durante a aprendizagem, a intervalos determinados pela constante de decaimento. A figura 3.5 representa algumas vizinhanças.

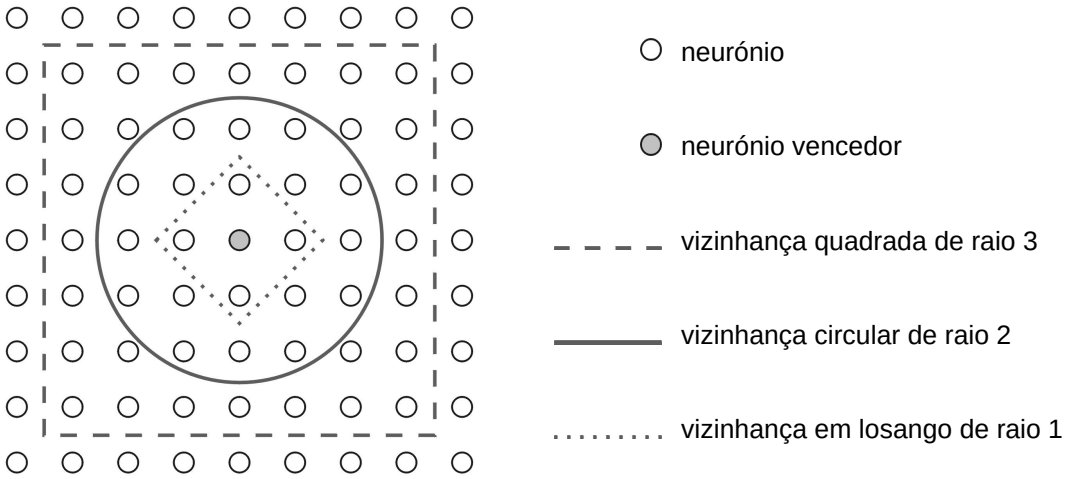


Figura 3.5 – Vizinhanças.

Um raio de vizinhança nulo determina uma vizinhança que apenas inclui o neurónio vencedor. Uma vizinhança inicial extensa resolve o problema da sub-utilização neuronal, pois garante que mesmo os neurónios não vencedores participam na alteração dos pesos sinápticos, e a sua redução em fases mais adiantadas da aprendizagem permite refinar o mapeamento dos estímulos.

3.3.2.2.3 Parâmetros de aprendizagem

Outro parâmetro de que depende a eficiência do algoritmo é o coeficiente de aprendizagem. Inicialmente, este parâmetro deve ser elevado, para permitir um rápido delineamento das características gerais do mapa, mas a convergência para uma solução estável requer o seu decaimento progressivo ao longo da aprendizagem. A escolha do tipo de decaimento (sendo os mais comuns o linear e o geométrico) e do valor do factor de decaimento pode determinar não apenas a velocidade de aprendizagem como também a qualidade e utilidade do mapa final.

3.3.2.2.4 Condição de paragem

O algoritmo de aprendizagem do mapa de Kohonen termina quando o coeficiente de aprendizagem se anula, ou atinge um valor negligenciável a partir do qual a representação dos estímulos nos pesos sinápticos para a camada competitiva praticamente não sofre alterações. Dependendo da finalidade do mapa, outras condições de paragem podem ser adoptadas, como a redução do raio da vizinhança abaixo de um certo valor, ou o atingir de um determinado número de iterações.

PARTE II

4 MATERIAIS E MÉTODOS

Este capítulo tem como temas os dados e a forma de apresentação dos resultados. A primeira secção descreve o conteúdo e formato dos ficheiros utilizados como fonte de informação, a codificação e normalização adoptadas para os estímulos, e a codificação das respostas pretendidas. A segunda secção introduz a matriz de erro e várias medidas de exactidão e erro utilizadas na descrição dos resultados apresentados no capítulo seguinte.

4.1 Origem e transformação dos dados

4.1.1 Base de dados HSSP

Todos os dados utilizados neste trabalho foram retirados da base de dados pública[8] HSSP (*Homology-derived Secondary Structure of Proteins*) [Sander e Schneider 91]. Actualizada frequentemente, no dia 28 de Julho de 1999 continha 9384 ficheiros referentes a proteínas cuja estrutura se encontra disponível na base de dados PDB, e para as quais a estrutura secundária foi determinada usando o programa[9] DSSP (*Database of Secondary Structure in Proteins*) [Kabsch e Sander 83]. Os nomes dos ficheiros HSSP são idênticos aos dos ficheiros PDB, e o seu formato encontra-se exemplificado na figura 4.1, abreviado por motivos estéticos.

O cabeçalho do ficheiro inclui informações sobre a origem e autores da determinação da estrutura da proteína, sobre o comprimento da sequência, número de cadeias da proteína e número de sequências usadas no alinhamento, e alguns parâmetros e notações utilizados, tanto na determinação da estrutura como do alinhamento.

De toda a informação contida nos ficheiros HSSP, a mais importante para este trabalho consiste na sequência, estrutura secundária e alinhamento, contidas na secção "`## ALIGNMENTS`". Os identificadores dos resíduos da sequência encontram-se localizados na coluna 15, precedidos pelo identificador da cadeia a que pertencem, na coluna 13. Quando existem dúvidas quanto à verdadeira identidade de alguns resíduos da sequência, utiliza-se símbolos diferentes daqueles introduzidos na tabela 2.2 (página 5). O símbolo "`!`" indica uma descontinuidade na sequência, geralmente a passagem de uma cadeia para outra. No entanto, este símbolo pode também indicar uma descontinuidade dentro da mesma cadeia, provocada por um erro ou uma lacuna nas coordenadas atómicas do ficheiro PDB correspondente.

[8] Acessível a partir do endereço http://www.sander.ebi.ac.uk/hssp.

[9] Acessível a partir do endereço http://www.sander.ebi.ac.uk/dssp.

```
HSSP       HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 1.0 1991
PDBID      2fiv
DATE       file generated on 14-Aug-98
SEQBASE    RELEASE 36.0 OF EMBL/SWISS-PROT WITH  74019 SEQUENCES
...
SEQLENGTH   118
NCHAIN        4 chain(s) in 2fiv data set
KCHAIN        2 chain(s) used here ; chain(s) :  A,I
NALIGN       16
...
## ALIGNMENTS    1 -   16
 SeqNo  PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  VAR
....:....1....:....2....:....3....:....4....:....5....:....6....:....7
     1    4 A V              0   0  117    7    4  VVI       VVV
     2    5 A G        +     0   0   75    7    0  GGG       GGG
     3    6 A T        -     0   0   32    7   46  TTT       VVV
     4    7 A T  E     -A  226   0A  79   10   29  TTT     E TTTE E
     5    8 A T  E     -A  225   0A  25   10   53  TTT     Y YYYL L
...
    96   99 A Q  S    S-     0   0   27    4    0  QQQ.............
    97  100 A P         -     0   0   18   11   39  PPPEEEEEDE......
    98  101 A L  E     -fH  25  34B   0   17   17  LLLVVVVVIVIIIIII
    99  102 A L  E     -f   26   0B   0   17   12  LLLLLLLLILLLLIII
   100  103 A G    >>  -     0   0    0   17    0  GGGGGGGGGGGGGGGG
...
   113  116 A M              0   0   17   17    7  MMMMMMMMLMLLLMFM
   114        !  !           0   0    0    0    0
   115  202 I X              0   0   51    0    0
   116  203 I V  E     -KO  30 238C   0    1    0
   117  204 I X  E     - O   0 237C   0    0    0
...
## SEQUENCE PROFILE AND ENTROPY
 SeqNo PDBNo   V   L   I   M   F  ...   A   P   D  NOCC NDEL NINS  ... WEIGHT
    1    4 A  86   0  14   0   0  ...   0   0   0     7    0    0  ...  1.46
    2    5 A   0   0   0   0   0  ...   0   0   0     7    0    0  ...  1.54
    3    6 A  43   0   0   0   0  ...   0   0   0     7    0    0  ...  0.66
    4    7 A   0   0   0   0   0  ... 30   0   0    10    0    0  ...  1.44
    5    8 A   0  20   0   0   0  ...  0   0   0    10    0    0  ...  0.67
...
   96   99 A   0   0   0   0   0  ...   0   0   0     4   13    0  ...  0.67
   97  100 A   0   0   0   0   0  ...   5   0   9    11    6    0  ...  0.71
   98  101 A  35  24  41   0   0  ...   0   0   0    17    0    0  ...  1.11
   99  102 A   0  76  24   0   0  ...   0   0   0    17    0    0  ...  1.36
  100  103 A   0   0   0   0   0  ...   0   0   0    17    0    0  ...  1.57
...
  113  116 A   0  24   0  71   6  ...   0   0   0    17    0    0  ...  1.41
  114          0   0   0   0   0  ...   0   0   0     0    0    0  ...  1.00
  115  202 I   0   0   0   0   0  ...   0   0   0     0    0    0  ...  1.00
  116  203 I 100   0   0   0   0  ...   0   0   0     1    0    0  ...  1.00
  117  204 I   0   0   0   0   0  ...   0   0   0     0    0    0  ...  1.00
...
```

Figura 4.1 – Formato de um ficheiro HSSP.

A coluna 18 identifica a estrutura secundária da proteína, determinada pelo DSSP. São utilizados sete símbolos para identificar motivos estruturais diferentes, segundo a tabela 4.1. A folha β isolada é uma folha β com comprimento unitário, não sendo por isso frequentemente considerada uma folha β normal; a curva com ligação de hidrogénio consiste geralmente numa fracção de hélice 3_{10} ou hélice π demasiado pequena para ser considerada uma hélice verdadeira; a ausência de símbolo indica que o resíduo não se encontra em nenhum motivo estrutural reconhecível, nem localizado numa zona de curvatura suficiente para ser considerada curva. Em caso de sobreposição de motivos, a prioridade de atribuição é a ordem de apresentação na tabela.

Tabela 4.1 – Identificadores de motivos de estrutura secundária.

Identificador	Motivo estrutural
H	hélice α
B	folha β isolada
E	folha β
G	hélice 3_{10}
I	hélice π
T	curva com ligação de hidrogénio
S	curva

Depois de alguma informação adicional sobre a estrutura e o alinhamento múltiplo, são apresentadas as sequências que participam neste último, a partir da coluna 52. Os pontos indicam deleções, e os pares de símbolos em letra minúscula indicam inserções que ocorreram entre os dois resíduos.

A secção "`## SEQUENCE PROFILE AND ENTROPY`" consiste numa matriz calculada com base no alinhamento, aqui denominada *matriz de perfil*, em que cada linha indica as percentagens de cada resíduo na respectiva posição da sequência. Devido à possível existência de resíduos não identificados, os elementos de algumas linhas podem ser todos nulos. Esta secção contém ainda outras informações, como o número de inserções e deleções ocorridas em cada posição da sequência.

4.1.2 Estímulos

4.1.2.1 Codificação

Tal como no programa PHD, neste trabalho os estímulos utilizados no treino das redes neuronais que efectuam a previsão da estrutura secundária, baseiam-se na informação contida nas matrizes de perfil dos ficheiros HSSP (ver secção 4.1.1).

Para utilizar informação sobre as interacções entre os resíduos, utiliza-se uma *janela de estímulo* de dimensão ímpar *n* que percorre a matriz de perfil, transformando cada segmento de *n* resíduos consecutivos num vector de estímulo de dimensão $n \times 20$ que contém as respectivas *n* linhas da matriz, dispostas lado a lado. A figura 4.2 ilustra este processo com uma janela de estímulo de dimensão 3, para maior clareza. Porque cada estímulo se refere apenas ao resíduo central, os estímulos referentes aos resíduos nas extremidades da sequência apresentam partes totalmente nulas, correspondentes às zonas da janela fora da sequência. No programa PHD, os estímulos têm dimensão $n \times 21$, correspondendo a posição adicional precisamente a zonas fora da sequência.

Embora os estímulos resultantes desta codificação sejam bastante esparsos, verificou-se que a tentativa de eliminar este problema, utilizando codificações que produzem estímulos mais compactos, introduz correlações falsas entre os estímulos, o que inviabiliza a sua correcta aprendizagem pelas redes neuronais.

Matriz de perfil:

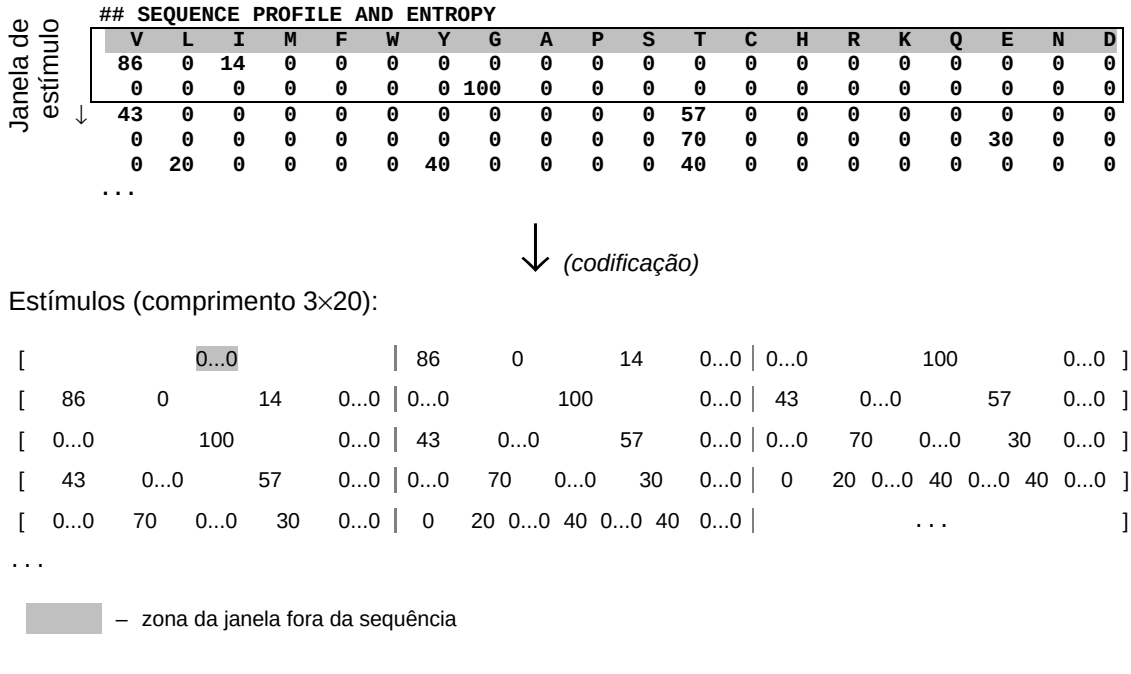


Figura 4.2 – Codificação dos estímulos.

4.1.2.2 Normalização

Normalizar os estímulos antes de os apresentar a um perceptrão multicamada pode facilitar a sua aprendizagem. No entanto, a normalização comum apenas garante a mesma magnitude (unitária) a todos os vectores, ignorando o facto de cada um ser constituído por elementos independentes – os resíduos que constituem o segmento.

Um estímulo, tal como descrito na secção anterior, pode ser encarado, não como um vector de $n \times 20$ elementos, mas como n vectores de 20 elementos cada. Assim, neste trabalho, os estímulos sofrem uma normalização em duas fases, ilustrada na figura 4.3. A primeira fase consiste na normalização de cada um dos vectores mais pequenos, o que garante a mesma magnitude a todos eles. A segunda fase consiste na normalização do vector maior resultante da primeira fase, o que lhe garante magnitude unitária. Os vectores mais pequenos ficam com magnitude $1/n$, excepto aqueles cuja codificação inclui zonas da janela fora da sequência, caso em que ficam com magnitude $1/(n - m)$, sendo m a dimensão da zona da janela fora da sequência ($m \leq n/2 - 1$). Na figura 4.2, $m = 1$.

4.1.3 Respostas

Um procedimento bastante comum na previsão da estrutura secundária de proteínas consiste em considerar apenas três classificações possíveis para os motivos estruturais em que se encontram os resíduos: hélice, folha e outros. Assim, os três tipos de hélice são identificados pelo nome *hélice*, a folha β denomina-se simplesmente *folha*, e os restantes motivos recebem a designação de *outros*. O conjunto de sete símbolos utilizados na base de dados HSSP e apresentados na secção anterior reduz-se a dois, H

(hélice) e E (folha), e a ausência de símbolo identifica os restantes. Alguns autores classificam a folha β isolada como folha [Rost e Sander 93], enquanto outros preferem atribuir-lhe a classe outros [Riis e Krogh 96]. Neste trabalho optou-se pela segunda alternativa.

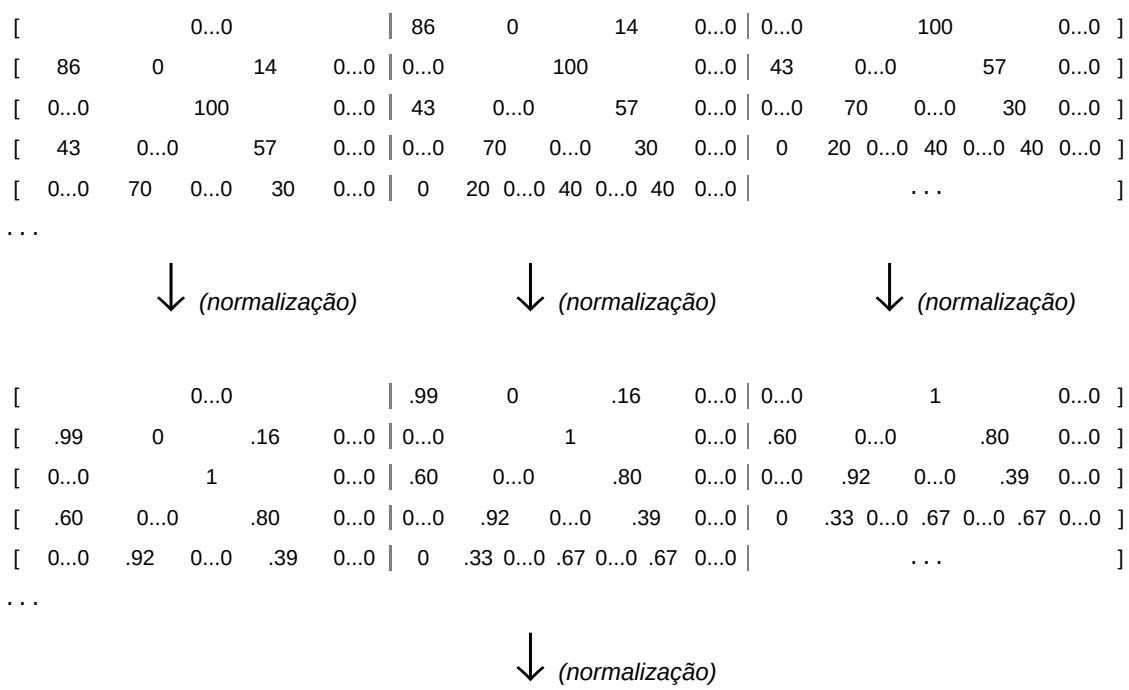


Figura 4.3 – Normalização em duas fases.

As respostas pretendidas para a rede neuronal de aprendizagem supervisionada obedecem a uma codificação binária. Cada efector representa um dos três motivos estruturais, e a sua resposta deve ser unitária quando o resíduo a classificar pertence ao motivo representado, e nula caso contrário. A resposta da rede consiste pois num vector de três elementos, sendo a classificação atribuída aquela representada pelo efector que devolveu o elemento de valor mais elevado.

4.2 Apresentação dos resultados

Os resultados das previsões da estrutura secundária podem ser apresentados sob uma grande variedade de formas. Neste trabalho utiliza-se algumas das mais comuns.

4.2.1 Matriz de erro

Uma *matriz de erro*, também denominada *matriz de confusão* ou *tabela de contingência* [Jensen 96][10], consiste numa matriz quadrada de números que indicam a quantidade de exemplos (neste caso, estímulos) classificados como pertencendo a determinada classe, relativamente à sua classe verdadeira. Embora não exista consenso sobre se as linhas se devem referir à classe verdadeira e as colunas à classificação, ou vice-versa, neste trabalho optou-se pela primeira alternativa.

Formalizando, uma matriz de erro, c, é uma matriz de dimensão $k \times k$, sendo k o número de classes consideradas e $c[i, j]$ o número de exemplos pertencentes à classe i e classificados como pertencendo à classe j (figura 4.4).

CLASSE VERDADEIRA \ CLASSIFICAÇÃO	1	...	j	...	k	Total
1	$c[1, 1]$	...	$c[1, j]$	...	$c[1, k]$	$t[1]$
...	...	...	...	...	...	...
i	$c[i, 1]$	...	$c[i, j]$	...	$c[i, k]$	$t[i]$
...	...	...	...	...	...	...
k	$c[k, 1]$	...	$c[k, j]$	...	$c[k, k]$	$t[k]$
Total	$m[1]$	...	$m[j]$	...	$m[k]$	n

Figura 4.4 – Matriz de erro.

O número total de exemplos pertencentes à classe i, $t[i]$, é dado pela soma

$$t[i] = \sum_{j=1}^{k} c[i, j].$$

O número de exemplos classificados como j, $m[j]$, consiste na soma

$$m[j] = \sum_{i=1}^{k} c[i, j].$$

Logo, o número total de exemplos, n, é dado por

$$n = \sum_{i=1}^{k} \sum_{j=1}^{k} c[i, j].$$

[10] A matriz de erro utiliza-se principalmente em problemas de classificação de imagens obtidas por detecção remota.

4.2.2 Medidas de exactidão e de erro

A informação fornecida pela matriz é utilizada na computação de uma série de medidas de exactidão e de erro [Jensen 96]. A mais simples de todas denomina-se *exactidão global* (muitas vezes representada por Q_3 em problemas relacionados com previsão de estrutura secundária em três classes), e mede a proporção de exemplos correctamente classificados em toda a amostra. Calcula-se dividindo o número total de exemplos correctamente classificados (a soma dos elementos principais da matriz) pela dimensão da amostra.

$$\text{Exactidão global} = \frac{1}{n}\sum_{i=1}^{k} c[i,j]$$

As exactidões dentro de cada classe considerada podem ser calculadas de duas formas distintas. Tradicionalmente, a exactidão na classe i calcula-se dividindo o número de exemplos correctamente classificados como i pelo número total de exemplos realmente pertencentes à classe i. Esta medida de exactidão é frequentemente designada por *exactidão do produtor*, porque quem produz a classificação deve preocupar-se em atribuir aos exemplos a sua classe verdadeira.

$$\text{Exactidão do produtor} = \frac{c[i,i]}{t[i]} \qquad (\forall\ 1 \le i \le k)$$

Deste modo, a exactidão do produtor indica a proporção de exemplos correctamente classificados na sua classe verdadeira. Esta medida encontra-se directamente relacionada com a medida oposta, denominada *erro de omissão*.

$$\text{Erro de omissão} = \frac{t[i] - c[i,i]}{t[i]} \qquad (\forall\ 1 \le i \le k)$$

O erro de omissão indica a proporção de exemplos de uma dada classe, incorrectamente classificados noutra classe, *i.e.*, a proporção de exemplos que a classificação omitiu da sua verdadeira classe. Note-se que a soma da exactidão do produtor com o erro de omissão iguala a unidade.

A segunda forma de calcular as exactidões por classe produz uma medida designada por *exactidão do utilizador*. Calcula-se dividindo o número de exemplos correctamente classificados numa dada classe pelo número total de exemplos classificados – correctamente ou não – nessa classe.

$$\text{Exactidão do utilizador} = \frac{c[i,i]}{m[i]} \qquad (\forall\ 1 \le i \le k)$$

Esta medida indica a proporção de exemplos que realmente pertencem à classe que lhes foi atribuída. O seu nome deriva do facto de o utilizador da classificação se preocupar em que a classe atribuída a cada exemplo seja de facto a sua classe verdadeira. As exactidões do produtor e utilizador podem revelar-se extremamente diferentes, pelo que é sempre aconselhável calcular as duas medidas. Analogamente ao caso anterior, a exactidão do utilizador e o *erro de comissão* são complemento uma da outra.

$$\text{Erro de comissão} = \frac{m[i] - c[i,i]}{m[i]} \qquad (\forall\ 1 \le i \le k)$$

O erro de comissão indica a proporção de exemplos incorrectamente classificados numa classe diferente da sua classe verdadeira.

4.2.3 Medidas utilizadas

A matriz de erro raramente é apresentada, pois as medidas de exactidão e de erro dela derivadas geralmente fornecem toda a informação considerada importante sobre a qualidade das previsões. Em contrapartida, os erros de omissão e de comissão aparecem quase sempre, o que dispensa a apresentação das exactidões do produtor e do consumidor.

Quanto à exactidão global, esta raramente surge na forma como foi descrita. Regra geral, uma previsão inclui várias proteínas; a exactidão global, que considera todos os resíduos conjuntamente, independentemente da proteína ou cadeia a que pertencem, não fornece qualquer pista quanto à exactidão obtida em cada uma. Assim, a medida de exactidão mais frequentemente apresentada consiste na média e desvio padrão dos valores de exactidão medidos em cada cadeia polipeptídica. Esta informação é muitas vezes acompanhada pelo respectivo histograma. As medidas de exactidão e de erro são expressas em valores percentuais.

Porque a capacidade de generalização de um perceptrão treinado é geralmente menor do que a capacidade de ajustamento aos dados de treino, optou-se por não apresentar os resultados medidos no conjunto de treino, salvo quando a comparação entre ajustamento e generalização é importante. Embora os dados disponíveis tenham sido divididos em conjuntos de treino, teste e validação sempre que possível, os resultados obtidos nos conjuntos de teste e validação revelaram-se sempre tão semelhantes, que se optou por apresentá-los em conjunto, excepto nos casos em que não existe conjunto de validação. Todos os valores apresentados são arredondados às unidades.

5 ESTUDO DE UM SISTEMA DE PREVISÃO

Este capítulo descreve o longo processo de estudo e desenvolvimento de um sistema de previsão da estrutura secundária de proteínas. Embora tenha sido baseado no programa PHD, ao longo do seu desenvolvimento foram testadas várias ideias diferentes, a maioria das quais infelizmente mal sucedidas. As secções que se seguem descrevem apenas as etapas consideradas mais importantes em todo este processo. O número de redes a utilizar na previsão, a dimensão da janela de estímulo e a aplicação de um filtro aos resultados obtidos são os três temas iniciais. Segue-se uma extensa secção dedicada à separação estrutural das proteínas como meio de tentar melhorar a qualidade das previsões, que inclui os resultados da experimentação de diferentes métodos. O capítulo termina com a descrição do índice de fiabilidade aqui desenvolvido e sua comparação com o índice utilizado no programa PHD.

Alguns parâmetros mantiveram-se constantes ao longo das várias etapas. Todas as aprendizagens supervisionadas foram realizadas com perceptrões de duas camadas com conexão sináptica total, salvo indicação em contrário. A sua arquitectura é especificada para cada caso, assim como o número de épocas de aprendizagem. A inicialização dos pesos foi sempre efectuada segundo o processo de Nguyen-Widrow e a função de activação usada a logística, com $b = 1$. A apresentação dos estímulos foi sempre aleatória com reposição e efectuada de forma a garantir um treino equilibrado, em que a escolha dos estímulos garante que todas as classes têm igual probabilidade de serem escolhidas, independentemente do número de exemplos que contêm. O coeficiente de aprendizagem utilizado foi constante e igual a 0.1, e a condição de paragem adoptada foi sempre o procedimento de Hecht-Nielsen.

Não foram efectuados muitos testes com o objectivo de determinar qual o número óptimo de neurónios internos do perceptrão multicamada. O principal motivo foi o facto dos recursos computacionais disponíveis para a realização deste trabalho se terem revelado claramente insuficientes para um volume de dados tão grande. Para além disso, este e outros factores relativos à arquitectura neuronal adoptada, assim como a diversas opções de aprendizagem, parecem não afectar grandemente a capacidade de aprendizagem do perceptrão [Rost e Sander 93].

5.1 Número de redes

5.1.1 Introdução

Alguns autores defendem que a utilização de tantas redes neuronais quantos os motivos estruturais a discriminar, cada uma treinada para reconhecer um único motivo, melhora os resultados da previsão da estrutura secundária [Riis e Krogh 96]. No entanto, a maioria dos autores utilizam apenas uma única rede, treinada para reconhecer todos os motivos estruturais, o que mantém o processo de classificação

relativamente simples. Na tentativa de esclarecer qual a melhor opção, procedeu-se à realização de duas previsões do mesmo conjunto de dados, adoptando-se métodos diferentes em cada uma.

5.1.2 Lista de cadeias PDB_SELECT

Muitos resultados publicados sobre previsão de estrutura secundária de proteínas apresentam valores de exactidão bastante superiores aos obtidos pelo programa PHD, apesar deste ser considerado o melhor programa de previsão de estrutura secundária conhecido até ao momento. A contradição deriva do facto de muitos autores testarem os seus métodos em proteínas homólogas às utilizadas no desenvolvimento dos mesmos, o que devolve resultados enganadores quanto à real capacidade de generalização do sistema.

Muitas das cadeias utilizadas no treino e teste do programa PHD, um conjunto de 156, não se encontram disponíveis na base de dados HSSP. Por este motivo, e pelo facto do número de estruturas conhecidas estar a aumentar rapidamente, neste trabalho recorreu-se a um conjunto de dados diferente. Publicamente disponível[11] e actualizada algumas vezes por ano, a lista PDB_SELECT [Hobohm *et al.* 92, Hobohm e Sander 94] consiste numa selecção representativa de cadeias, contendo cinco ou seis vezes menos sequências do que a base de dados PDB. O algoritmo que efectua a selecção garante que todas as sequências da lista apresentam menos de 25% de identidade entre si (embora outras listas sob o mesmo nome sejam mais permissivas), esperando-se assim reduzir a homologia a um nível negligenciável. As sequências desta lista consistem em identificadores formados por duas partes: quatro caracteres que identificam o ficheiro (PDB, HSSP ou outro) que contém a proteína, e um carácter que identifica a cadeia pretendida.

Neste trabalho foi utilizada a lista PDB_SELECT de Agosto de 1998. Das 947 cadeias nela indicadas, muitas não foram encontradas na base de dados HSSP, não continham informação completa sobre a estrutura secundária, ou apresentavam descontinuidades, restando 727 cadeias, aqui designadas por *conjunto PDB_SELECT*. Estas foram divididas em conjuntos de treino, teste e validação (ver secção 3.2.2.2.5), contendo o conjunto de validação 10% das cadeias, o de teste 20% das restantes 90%, e o de treino todas as restantes.

5.1.3 Uma rede *versus* três redes

O conjunto PDB_SELECT representa um volume de dados demasiado grande para permitir a realização de diversos testes em tempo útil. Tendo em conta a relação que existe entre o número de alinhamentos (leia-se o número de sequências utilizadas no alinhamento múltiplo) e a qualidade da previsão [Rost e Sander 93], o conjunto foi reduzido de forma a aproveitar somente as cadeias com 20 ou mais alinhamentos (pois é mais fácil escolher entre duas previsões boas do que entre duas previsões más), que representam pouco mais de 30% do conjunto inteiro. Note-se que, deste modo, os resultados apresentados são melhores do que seria de esperar para um conjunto de dados normal, onde a maioria das cadeias tem um número de alinhamentos mais reduzido. Foi utilizada uma janela de estímulo de dimensão 7, o que representa mais

[11] Acessível a partir do endereço http://swift.embl-heidelberg.de/pdbsel.

uma limitação ao volume de dados, assim como à dimensão das redes necessárias ao seu processamento.

Para realizar a primeira previsão, um perceptrão multicamada (ver secção 3.2) constituído por 140 sensores, 35 neurónios internos e 3 efectores foi treinado para discriminar entre hélices, folhas e outros, durante 80 épocas. A figura 5.1 apresenta as medidas de exactidão e erro da previsão realizada por esta rede, assim como um histograma das exactidões obtidas nas várias cadeias. Recorda-se que estes valores referem-se aos conjuntos de teste e validação.

Exactidão (%):

Média	69
Desvio padrão	9

Erros (%):	hélice	folha	outros
Omissão	23	32	36
Comissão	28	46	22

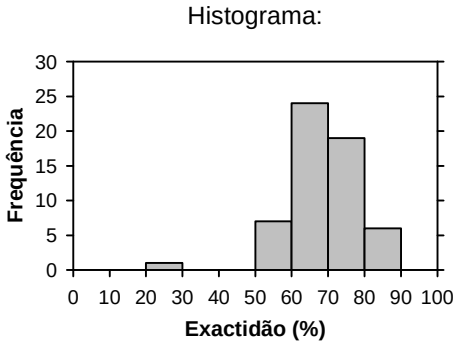


Figura 5.1 – Resultados: uma rede para três motivos estruturais, com janela de dimensão 7.

O maior erro que se pode observar nos resultados é o erro de comissão da classe folha. Uma das cadeias, a proteína 1cfh, foi extremamente mal classificada, com apenas 30% de exactidão. A observação desta cadeia revela que contém um número de resíduos extremamente baixo (47) e a sua estrutura secundária é constituída por apenas uma pequena hélice.

Para a segunda previsão, três perceptrões constituídos pelo mesmo número de sensores e de neurónios internos e somente um efector, foram treinados independentemente para reconhecer cada um dos motivos estruturais, durante 85 (hélice), 73 (folha) e 92 (outros) épocas. As respostas dos efectores destas três redes foram combinadas como se fossem na realidade as respostas dos três efectores de uma única rede, e interpretadas exactamente da mesma forma. A figura 5.2 apresenta os resultados da previsão por elas realizada.

Nesta previsão, o erro de comissão da classe folha baixou ligeiramente, assim como o erro de omissão da classe outros. A proteína 1cfh foi classificada com 38% de exactidão.

5.1.4 Conclusão

A previsão realizada pelas três redes treinadas em motivos estruturais diferentes aparenta ser ligeiramente melhor do que a realizada pela rede única. O motivo pode dever-se ao facto das aprendizagens dos diferentes motivos estruturais necessitarem de um número diferente de épocas até atingirem a condição de paragem, ou ao facto das redes treinadas apenas num motivo estrutural disporem de um número de neurónios internos mais elevado em relação ao número de classes a discriminar. No

entanto, a ligeira melhoria conseguida pelas três redes pode ter sido casual, e dificilmente compensa o aumento de complexidade relativamente a uma única rede.

Exactidão (%):

Média	70
Desvio padrão	9

Erros (%):	hélice	folha	outros
Omissão	24	34	31
Comissão	27	42	23

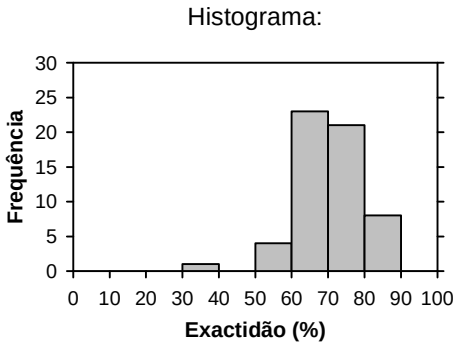


Figura 5.2 – Resultados: uma rede para cada motivo estrutural, com janela de dimensão 7.

5.2 Dimensão da janela de estímulo

5.2.1 Introdução

Os primeiros autores a publicar resultados sobre previsão de estrutura secundária de proteínas efectuada com redes neuronais [Qian e Sejnowski 88] utilizaram uma janela de estímulo de dimensão 13, concluindo ser esta a dimensão mais adequada. Não obstante, outros autores utilizaram janelas de dimensão superior, como 17 [Holley e Karplus 89], ou mesmo 51 [Bohr *et al*. 88].

Embora os recursos computacionais disponíveis para este trabalho obriguem a utilizar apenas janelas de estímulo de dimensões comedidas, urge verificar, no mínimo, se a alteração de 7 para 13 provoca diferenças significativas na qualidade da previsão.

5.2.2 Dimensão 7 *versus* dimensão 13

Utilizando os mesmos conjuntos de dados da secção anterior, convertidos em estímulos por uma janela de dimensão 13, foram efectuadas duas previsões, à semelhança da secção anterior: a primeira utilizando somente uma rede, e a segunda utilizando uma rede por cada um dos três motivos estruturais a discriminar. As redes utilizadas diferem das anteriores apenas no número de sensores, que aumenta para 260. A rede de três efectores foi treinada durante 67 épocas, e as três redes de efector único foram treinadas em 61 (hélice), 70 (folha) e 102 (outros) épocas. As figuras 5.3 e 5.4 apresentam os resultados das previsões realizadas pela rede única e pelas três redes, respectivamente.

Em ambos os casos, não só a exactidão média subiu alguns pontos percentuais, como praticamente todas as percentagens de erro baixaram consideravelmente, muitas apresentando diferenças de cinco ou mais pontos percentuais. A proteína 1cfh tornou-se menos problemática, tendo sido classificada com 45% e 49% de exactidão, respectivamente pela rede única e pelas três redes.

Exactidão (%):

Média	72
Desvio padrão	8

Erros (%):	hélice	folha	outros
Omissão	23	29	30
Comissão	23	41	24

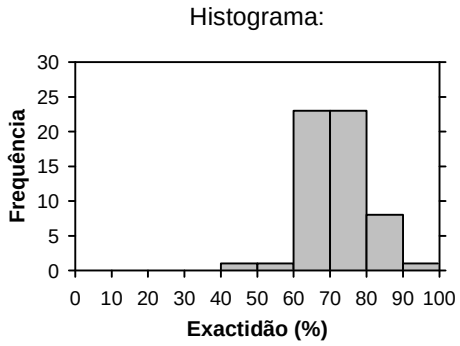


Figura 5.3 – Resultados: uma rede para três motivos estruturais, com janela de dimensão 13.

Exactidão (%):

Média	73
Desvio padrão	8

Erros (%):	hélice	folha	outros
Omissão	24	32	26
Comissão	21	39	25

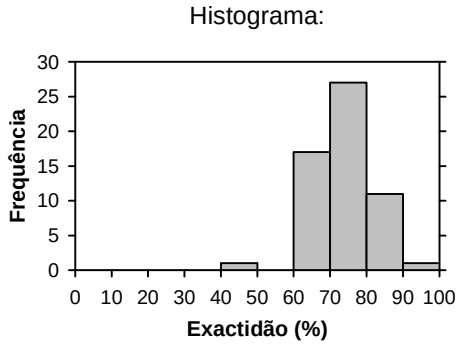


Figura 5.4 – Resultados: uma rede para cada motivo estrutural, com janela de dimensão 13.

5.2.3 Conclusão

A utilização de janelas de estímulo de dimensão 13 permite obter previsões consideravelmente melhores do que as conseguidas com janelas de dimensão 7, confirmando a importância que as interacções não locais entre os resíduos têm na determinação da estrutura secundária da proteína. O treino de redes neuronais utilizando janelas de dimensão superior a 13 constitui uma operação computacionalmente cara, motivo provável pelo qual esta opção parece não ter sido estudada. Mais uma vez se verificou uma ligeira vantagem na utilização de uma rede por cada motivo estrutural, levando a crer que as melhorias introduzidas por esta opção nos resultados da secção anterior não foram casuais.

5.3 Filtro

5.3.1 Introdução

O segundo nível computacional do programa PHD (ver secção 2.5.2.3) é composto por redes neuronais que actuam como filtros, recebendo informação proveniente das redes do nível anterior, adicionando-lhe alguma informação de carácter global, e devolvendo uma classificação melhorada dos resíduos da proteína. O quarto nível computacional também consiste num filtro, prenunciando que este tipo de processamento introduz de facto melhorias na qualidade das previsões. Para tentar confirmar estas suspeitas, aplicou-se um filtro às classificações obtidas nas secções anteriores, inspirado no segundo nível computacional do PHD, embora bastante mais simples, por não utilizar mais nenhuma informação para além das respostas dos efectores da primeira rede. A utilização de um filtro com janela de dimensão superior a 13 pode resolver alguns problemas provocados pelo desprezar de informação relativa a interacções entre resíduos mais distantes na sequência. Talvez consiga mesmo resolver muitos erros resultantes da utilização de janelas de estímulo de dimensão mais reduzida.

5.3.2 Filtragem de resultados anteriores

O filtro foi aplicado às classificações produzidas pelas redes únicas descritas nas duas secções 5.1 e 5.2. As respostas dos três efectores, para 17 resíduos consecutivos na sequência, foram dispostas em vectores de 51 elementos cada, sem qualquer normalização, e estes apresentados a perceptrões multicamada com 51 sensores, 17 neurónios internos, e 3 efectores. O primeiro filtro, aplicado à classificação obtida com a janela de estímulo de dimensão 7, foi treinado durante 37 épocas, produzindo os resultados apresentados na figura 5.5. O segundo filtro, aplicado à classificação obtida com a janela de dimensão 13, foi treinado durante 65 épocas, produzindo os resultados apresentados na figura 5.6.

Exactidão (%):

Média	74
Desvio padrão	9

Erros (%):	hélice	folha	outros
Omissão	21	31	26
Comissão	20	35	24

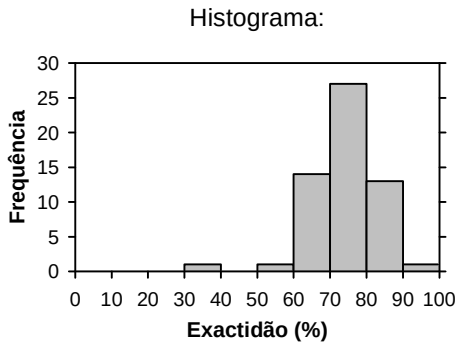


Figura 5.5 – Resultados: filtro aplicado à classificação produzida com janela de dimensão 7.

Em relação aos resultados apresentados nas secções 5.1 e 5.2, as melhorias são óbvias. Tanto num caso como noutro, a exactidão média subiu alguns pontos percentuais e os

erros de omissão e comissão decresceram consideravelmente, especialmente no primeiro caso, onde as diferenças chegam a exceder dez pontos percentuais.

A proteína 1cfh continua a ser mal classificada, com exactidões de 34% e 49% no primeiro e no segundo caso, respectivamente. Curiosamente, a proteína 1ppt, que contém um número ainda mais pequeno de resíduos (37), e cuja estrutura secundária é constituída por uma longa hélice, foi classificada com exactidões de 97% e 100%, respectivamente. Excluindo algumas diferenças mais notórias, as duas previsões apresentadas nesta secção revelam-se espantosamente semelhantes.

Exactidão (%):

Média	75
Desvio padrão	8

Erros (%):	hélice	folha	outros
Omissão	21	29	26
Comissão	19	35	25

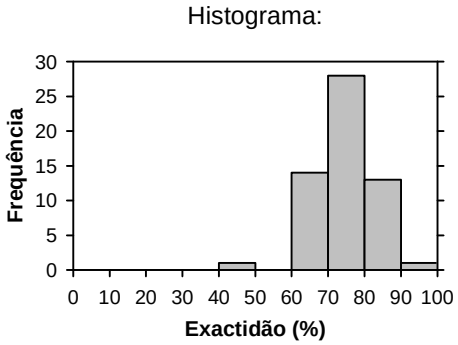


Figura 5.6 – Resultados: filtro aplicado à classificação produzida com janela de dimensão 13.

5.3.3 Conclusão

A aplicação de filtros às classificações produzidas pelas redes iniciais melhora francamente a qualidade das previsões. A previsão que havia sido efectuada com janela de dimensão 7 beneficia de tais melhoramentos que as diferenças de qualidade em relação à previsão efectuada com janela de dimensão 13 se esbatem quase totalmente. Confirma-se pois que os erros resultantes da utilização de uma janela de dimensão reduzida podem ser resolvidos desta forma, e levanta-se a dúvida sobre a necessidade de usar janelas de dimensão superior.

5.4 Separação em classes estruturais

5.4.1 Introdução

Levitt e Chothia [Levitt e Chothia 76] foram os primeiros autores a realizar uma classificação de proteínas em classes estruturais. Com base na observação visual da sucessão de motivos estruturais na cadeia polipeptídica, dividiram um conjunto de 37 proteínas globulares em quatro classes estruturais distintas, com designações semelhantes, mas definições ligeiramente diferentes, daquelas adoptadas mais tarde. Mas o conjunto utilizado era muito reduzido, e à medida que as bases de dados de estruturas conhecidas foram crescendo, tornou-se evidente que os critérios de

classificação então utilizados não eram muito eficazes, principalmente na discriminação entre as proteínas α/β e $\alpha+\beta$.

Vinte anos mais tarde, Michie, Orengo e Thorton [Michie *et al.* 96] apresentaram um sistema de classificação automática, baseado nas percentagens de hélices α e folhas β, paralelas e antiparalelas, que constituem a proteína, nos contactos[12] observados entre resíduos pertencentes aos diferentes motivos estruturais, e numa medida de alternância de motivos ao longo da sequência. Ajustado a um conjunto de 197 cadeias e testado num conjunto independente de 43 cadeias, classificou correctamente cerca de 90% das cadeias, em ambos os conjuntos, deixando as restantes por classificar. Este sistema considera a existência de cinco classes: quatro classes correspondentes às definições apresentadas na secção 2.3, mais uma classe, aqui denominada *outros*, para as sequências que praticamente não apresentam motivos de estrutura secundária reconhecíveis, *i.e.*, hélices ou folhas.

Determinar a classe estrutural de uma proteína, observando a sua conformação, não é uma tarefa fácil; prevê-la, tendo como única informação a estrutura primária, ainda menos. Muitos autores têm abordado este problema, afirmando que o conhecimento da classe estrutural de uma proteína pode facilitar grandemente a previsão da sua estrutura secundária, pois permite a utilização de métodos especializados nas características particulares de cada classe [Cohen e Cohen 94]. Outros referem que a previsão da classe estrutural com base na estrutura primária pouco ou nada facilita a previsão da estrutura secundária, apresentando como motivo principal a dificuldade em discriminar entre as diferentes classes estruturais [Rost e Sander 93].

Partindo do pressuposto de que a divisão do volume de dados disponível em conjuntos mais homogéneos pode facilitar a previsão da estrutura secundária, procedeu-se ao estudo de diferentes formas de realizar e utilizar a separação em classes estruturais.

5.4.2 Vantagens do conhecimento da classe estrutural

Das 240 cadeias utilizadas por Michie, foram dispensadas as que o sistema automático não classificou. Outras não foram encontradas na base de dados HSSP, não continham informação completa sobre a estrutura secundária, ou apresentavam descontinuidades, restando 191 cadeias, aqui designadas por *conjunto de Michie*. A classe outros, contendo as sequências com percentagens muito reduzidas de motivos estruturais reconhecíveis, também foi dispensada por conter apenas sete cadeias. As restantes 184 cadeias, aqui designadas por *conjunto de Michie reduzido*, foram divididas em dois conjuntos, de treino e teste, contendo respectivamente 80% e 20% das cadeias pertencentes a cada classe estrutural. Não foi utilizado conjunto de validação, devido ao reduzido volume de dados disponível por cada classe. Utilizou-se uma janela de estímulo de dimensão 7.

A tabela 5.1 especifica as percentagens de hélice, folha e outros encontradas em cada classe estrutural considerada, e no total do conjunto.

[12] Considera-se que existe contacto entre dois resíduos quando estes se encontram a uma distância inferior a um dado limiar, em Ângströms (Å), que depende dos motivos estruturais envolvidos: entre hélice e hélice, 8 Å; entre hélice e folha, 10 Å; entre folha e folha, 21 Å. 1 Å = 10^{10} m.

Tabela 5.1 – Percentagens dos motivos estruturais no conjunto de Michie reduzido.

(%)	hélice	folha	outros
α/α	66	1	33
β/β	7	43	50
α/β	42	16	42
α+β	27	26	47
global	32	24	44

Foram efectuadas duas previsões, utilizando perceptrões multicamada constituídos por 140 sensores, 35 neurónios internos e 3 efectores cada. Aos resultados obtidos foram aplicados filtros consistindo em perceptrões com 51 sensores, 17 neurónios internos e 3 efectores cada, tal como descrito na secção 5.3. A primeira previsão foi efectuada por um perceptrão, treinado durante 90 épocas, e um filtro, treinado durante 55 épocas; a segunda previsão contou com quatro perceptrões, cada um treinado somente numa das quatro classes estruturais consideradas, durante 134 (α/α), 129 (β/β), 73 (α/β) e 91 (α+β) épocas, e quatro filtros, treinados durante 23, 20, 31 e 27 épocas, respectivamente. A tabela 5.2 apresenta os resultados produzidos por ambas, apenas no conjunto de teste, especificando as exactidões por classe estrutural e globais. Estes, ao contrário dos resultados apresentados nas secções anteriores, já constituem uma estimativa correcta do que se pode esperar obter na previsão da estrutura secundária de proteínas, pois foram obtidos num conjunto com qualquer número de alinhamentos.

Tabela 5.2 – Resultados: com e sem separação em classes estruturais, no conjunto de Michie reduzido.

Exactidão (média ± desvio padrão):

(%)	α/α	β/β	α/β	α+β	global
sem separação	76±9	64±11	80±4	69±10	70±11
com separação	78±8	73±7	80±9	57±13	71±12

As previsões das cadeias β/β revelam-se de qualidade bastante superior quando classificadas pelas redes especializadas nestas classes, e apenas ligeiramente melhores no caso α/α. As previsões efectuadas pelo método generalista (sem separação em classes estruturais) produzem melhores resultados na classe α+β. A classe α/β é muito bem classificada em ambos os casos, preferindo-se no entanto o método mais generalista, em que os resultados de exactidão apresentam um desvio padrão menor. Utilizando o método especializado nas classes α/α e β/β, e o método generalista nas classes α/β e α+β, obtém-se os resultados apresentados na figura 5.7.

Exactidão (%):

Média	74
Desvio padrão	9

Erros (%):	hélice	folha	outros
Omissão	24	31	25
Comissão	19	36	26

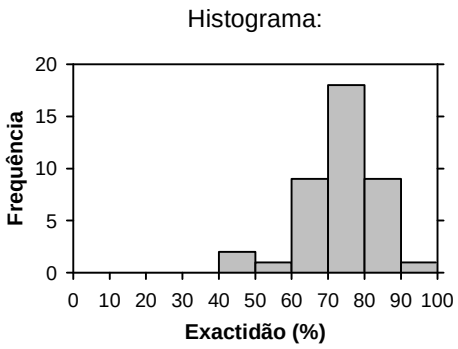


Figura 5.7 – Resultados: separação das classes estruturais α/α e β/β, no conjunto de Michie reduzido.

5.4.3 Atribuição de classes não supervisionada

A separação em classes estruturais, descrita na secção anterior, resulta em conjuntos cuja homogeneidade se centra nas respostas pretendidas, e não necessariamente nos estímulos. Para obter conjuntos de homogeneidade centrada nos estímulos, a separação deve basear-se na composição das cadeias polipeptídicas, por exemplo, no conhecimento das frequências dos aminoácidos que as constituem.

Foram calculadas as frequências de aminoácidos das 191 cadeias do conjunto de Michie. Os vectores de frequências foram apresentados a um mapa de Kohonen de dimensões 2×2 que, com uma vizinhança quadrada de raio 1 e constante de decaimento 50, coeficiente de aprendizagem 0.5 e decaimento linear de factor 0.001, produziu um mapeamento em 500 iterações.

A tabela 5.3 mostra uma matriz que relaciona as classes estruturais a que realmente pertencem as cadeias com os aglomerados resultantes do mapeamento dos estímulos. Embora as duas classificações sejam extremamente diferentes, a distribuição revela que diferentes classes estruturais preferem ou evitam diferentes aglomerados. Foram utilizados diversos outros parâmetros topológicos e de aprendizagem, mas o mapeamento demonstrou ser praticamente insensível a essas alterações.

Tabela 5.3 – Distribuição das cadeias no conjunto de Michie.

	α/α	β/β	α/β	$\alpha+\beta$	outros
aglomerado 1	7	31	4	8	1
aglomerado 2	4	12	0	9	4
aglomerado 3	21	18	5	26	1
aglomerado 4	19	4	10	6	1

A tabela 5.4 especifica as percentagens de hélice, folha e outros encontradas nos aglomerados formados pelo mapa de Kohonen, onde se pode identificar claramente a predominância de um ou outro motivo, consoante o aglomerado. Isto prova que existe de facto uma relação entre a homogeneidade nos estímulos e as características das

respostas pretendidas, embora não relacionadas com a separação em classes estruturais verdadeiras.

Tabela 5.4 – Percentagens dos motivos estruturais nos aglomerados do mapa de Kohonen.

(%)	hélice	folha	outros
aglomerado 1	19	32	49
aglomerado 2	18	22	60
aglomerado 3	34	23	43
aglomerado 4	50	14	36

Embora a separação efectuada de modo não supervisionado não apresente muitas semelhanças com a separação em classes estruturais, isso não significa que não facilite de igual modo a aprendizagem dos perceptrões multicamada. Assim, foram utilizados quatro perceptrões com a mesma arquitectura utilizada na secção anterior, cada um especializado num dos aglomerados formados pelo mapa de Kohonen, treinados durante 94 (1), 13 (2), 90 (3) e 78 (4) épocas, e quatro filtros, também análogos aos da secção anterior, e treinados durante 10, 9, 17 e 15 épocas, respectivamente. A tabela 5.5 apresenta os resultados obtidos com e sem filtro, apenas no conjunto de teste, especificando as exactidões por aglomerado e globais.

Tabela 5.5 – Resultados: com e sem filtro, com separação em aglomerados.

Exactidão (média ± desvio padrão):

(%)	aglomerado 1	aglomerado 2	aglomerado 3	aglomerado 4	global
sem filtro	64±6	70±9	65±7	60±7	65±7
com filtro	67±9	62±11	70±9	63±9	66±9

Curiosamente, no aglomerado 2, cuja rede utilizou um número de épocas de aprendizagem anormalmente reduzido, o filtro tem uma acção prejudicial. Este fenómeno já havia sido observado, embora não de forma tão dramática, na previsão da classe $\alpha+\beta$ da secção anterior, também utilizando redes especializadas (resultados não apresentados). Com ou sem filtro, as previsões efectuadas usando a atribuição de classes não supervisionada são de qualidade bastante inferior às da secção anterior.

5.4.4 Previsão da classe estrutural

Uma vez que a separação em classes estruturais revelou permitir melhores previsões que a separação efectuada pelo mapa de Kohonen, procedeu-se ao estudo de formas de efectuar a separação estrutural com base na estrutura primária.

5.4.4.1 Frequências de aminoácidos

As frequências de aminoácidos das 184 cadeias do conjunto de Michie reduzido foram divididas em dois conjuntos, de treino e de teste, e apresentadas a diversos perceptrões multicamada, com o objectivo de discriminar as quatro classes estruturais

α/α, β/β, α/β e α+β. Devido à conhecida dificuldade em separar as classes α/β e α+β, estas foram também utilizadas como um único conjunto, denominado αβ, na discriminação de somente três classes estruturais.

A melhor separação em três classes foi obtida por um perceptrão constituído por 20 sensores, 10 neurónios internos e 3 efectores, com conexão sináptica parcial a 75% entre as diversas camadas (após a inicialização dos pesos, 25% das conexões são removidas). A melhor separação em quatro classes foi conseguida por um perceptrão com o mesmo número de sensores e efectores, e 15 neurónios internos, também com conexão parcial a 75%. Treinados durante 237 e 157 épocas, respectivamente, produziram os resultados apresentados na figura 5.8. As matrizes e tabelas de erro referem-se apenas ao conjunto de teste.

Exactidão (%):

Treino	65
Teste	57

Matriz de erro:

	α/α	β/β	αβ
α/α	6	1	3
β/β	0	6	7
αβ	2	3	9

Erros (%):	α/α	β/β	αβ
Omissão	40	54	36
Comissão	25	40	53

Exactidão (%):

Treino	61
Teste	62

Matriz de erro:

	α/α	β/β	α/β	α+β
α/α	7	1	1	1
β/β	1	8	1	3
α/β	2	0	1	1
α+β	1	2	0	7

Erros (%):	α/α	β/β	α/β	α+β
Omissão	30	38	75	30
Comissão	36	27	67	42

Figura 5.8 – Resultados: separação do conjunto de Michie reduzido em três e quatro classes estruturais, com frequências de aminoácidos.

Os resultados da separação em três classes não apresentam quaisquer surpresas: as classes α/α e β/β confundem-se bastante com a classe αβ, mas muito pouco entre si. Na separação em quatro classes, a maior dificuldade consiste na discriminação entre as classes β/β e α+β. Curiosamente, as classes α/β e α+β praticamente não se confundem. Ambos os resultados são, no entanto, bastante maus.

5.4.4.2 Frequências de pares de aminoácidos

Na tentativa de melhorar os resultados da separação em classes estruturais, foram calculadas as frequências de pares de aminoácidos do conjunto de Michie reduzido, e utilizadas de forma análoga.

A melhor separação em três classes foi obtida por um perceptrão constituído por 400 sensores, 10 neurónios internos e 3 efectores, com conexão parcial a 75% entre as diversas camadas. A melhor separação em quatro classes foi conseguida por um perceptrão com o mesmo número de sensores e efectores, e 15 neurónios internos, com conexão total entre as camadas. Treinados durante 123 e 109 épocas

respectivamente, produziram os resultados apresentados na figura 5.9, onde as matrizes e tabelas de erro se referem apenas ao conjunto de teste.

Exactidão (%):

Treino	96
Teste	59

Matriz de erro:

	α/α	β/β	αβ
α/α	7	1	2
β/β	2	7	4
αβ	2	4	8

Erros (%):	α/α	β/β	αβ
Omissão	30	46	43
Comissão	36	42	43

Exactidão (%):

Treino	93
Teste	54

Matriz de erro:

	α/α	β/β	α/β	α+β
α/α	7	3	0	0
β/β	1	10	1	1
α/β	1	1	2	0
α+β	1	8	0	1

Erros (%):	α/α	β/β	α/β	α+β
Omissão	30	23	50	90
Comissão	30	55	33	50

Figura 5.9 – Resultados: separação do conjunto de Michie reduzido em três e quatro classes estruturais, com frequências de pares de aminoácidos.

Apesar de ambos os resultados de treino serem extremamente bons, a generalização apresenta os mesmos problemas presentes na classificação obtida usando as frequências de aminoácidos, com a agravante de demonstrar uma confusão acrescida entre as classes α/α e β/β. Na separação em quatro classes, a discriminação entre α/β e α+β foi perfeita, o que contraria abertamente a ideia generalizada de que estas duas classes são as mais difíceis de discriminar. Apesar de bastante promissores neste aspecto, estes resultados também são de qualidade francamente medíocre.

5.4.4.3 Regras de classificação

Perante a aparente impossibilidade de conseguir uma boa separação em classes estruturais com base na composição das cadeias polipeptídicas, recorreu-se a uma abordagem diferente, referida por alguns autores [Cohen e Cohen 94]. Esta baseia-se na observação de uma previsão preliminar da estrutura secundária da proteína que, se for de qualidade elevada, pode fornecer uma estimativa apurada das frequências de hélice, folha e outros, assim como da sua alternância ao longo da sequência. Estes dados foram utilizados na inferência de regras de classificação inspiradas nas regras utilizadas pelo sistema de classificação automática de Michie *et al.*, embora este utilize também informação referente à conformação da proteína, como os contactos entre resíduos e o tipo predominante de folha β, que não pode ser obtida em nenhuma previsão de estrutura secundária, por muito exacta que seja.

As regras utilizam a *medida de alternância de Michie*, utilizada pelo sistema de classificação automática, e outra mais simples, aqui designada precisamente por *medida de alternância simples*. A medida de Michie consiste na média entre duas medidas diferentes, calculadas em direcções opostas da sequência, como mostra a

figura 5.10. A medida simples consiste na soma da pontuação calculada em qualquer direcção (mesma figura).

			⟵ C– para N– ⟵							
sequência de motivos estruturais	H	E	H	E	H	E	E	E	E	E
pontuação	0	1	1	1	1	1	0	0	0	0
bónus	0	1	2	3	4	5	4	3	2	1
soma corrente	0	2	5	9	14	20	24	27	29	30
total										30
total normalizado										**3**

			⟶ N– para C– ⟶							
sequência de motivos estruturais	H	E	H	E	H	E	E	E	E	E
pontuação	1	1	1	1	1	0	0	0	0	0
bónus	1	0	-1	-2	-3	-4	-3	-2	-1	0
soma corrente	-10	-12	-13	-13	-12	-10	-6	-3	-1	0
total										-10
total normalizado										**-1**

H – segmento em hélice
E – segmento em folha

Medida de alternância de Michie	$(3 + (-1))/2 = 1$
Medida de alternância simples	$1 + 1 + 1 + 1 + 1 = 5$

Figura 5.10 – Cálculo das medidas de alternância. [13]

Na especificação das regras de classificação em classes estruturais, *Sscore* denota a medida simples, *Mscore* a medida de Michie, e *%H* e *%E* as percentagens de resíduos pertencentes a hélices e folhas, respectivamente, no total de resíduos pertencentes a um dos dois motivos estruturais (não incluindo a classe outros). As duas percentagens são, pois, complementares. As regras devem ser aplicadas pela ordem de apresentação.

Regras de classificação (primeira versão):

α/α:

$(\%E \leq 0.1) \vee ((\%H \geq 85) \wedge (\%E \leq \%H - 40) \wedge (\%E \leq 15) \wedge (Sscore \leq 7))$

β/β:

$(\%H \leq 0.1) \vee ((\%E - Mscore \geq 70) \wedge (\%H + Mscore \leq \%E - 45))$

α/β:

$(Mscore \geq 4) \vee ((Sscore \geq 8) \wedge (Mscore \geq 2) \wedge (\%E/\%H \leq 0.6))$

$\alpha+\beta$:

(restantes)

[13] Segundo [Michie *et al.* 96].

Utilizando a informação sobre a estrutura secundária real das proteínas do conjunto de Michie reduzido, estas regras mostraram ser capazes de separar as quatro classes estruturais com uma exactidão de 97%, medida nos conjuntos de treino e teste. Tal qualidade era esperada, uma vez que as regras foram inferidas usando precisamente essa informação. Mas ao utilizar, não a informação sobre a estrutura secundária real, mas apenas as estimativas fornecidas por uma previsão (apresentada na secção 5.4.2, obviamente sem separação estrutural), a exactidão desce para 77%.

A segunda versão das regras resultou de um ajustamento de alguns parâmetros às estimativas do conjunto de treino. Os limites de decisão foram alterados, e as percentagens de resíduos em hélice e folha foram reduzidas por um factor de certeza, denominado *índice de fiabilidade*. Este, ligeiramente diferente do índice de fiabilidade do programa PHD, calcula-se para cada resíduo classificado e consiste na diferença entre as duas respostas mais elevadas dos efectores do perceptrão, multiplicada pelo valor da resposta mais elevada. A utilidade deste índice alarga-se muito além da inferência de regras de classificação estrutural, e constitui tema de uma secção posterior.

Na especificação das novas regras de classificação, utiliza-se a notação introduzida anteriormente, e *%h* e *%e* denotam as percentagens *%H* e *%E* multiplicadas pelo índice de fiabilidade médio nas hélices e folhas, respectivamente.

Regras de classificação (segunda versão):

α/α:

$$(\%e \le 0.1) \vee \begin{pmatrix} (\%h - Mscore \ge 30) \wedge (\%e + Mscore \le \%h - 25) \wedge \\ \wedge\ (\%e + Mscore \le 15) \wedge (Sscore \le 7) \end{pmatrix}$$

β/β:

$$(\%h \le 0.1) \vee ((\%e - Mscore \ge 30) \wedge (\%h + Mscore \le \%e - 20))$$

α/β:

$$(Mscore \ge 4) \vee ((Sscore \ge 8) \wedge (Mscore \ge 2) \wedge (\%e/\%h \le 0.6))$$

α+β:

(restantes)

A figura 5.11 mostra os resultados obtidos por aplicação destas regras às estimativas do conjunto de teste. As matrizes e tabelas de erro referem-se ao conjunto de teste. Embora a exactidão global tenha subido para 80% na separação em quatro classes estruturais, a generalização (teste) revela-se bem pior que o ajustamento (treino). A classe α/β confunde-se com a α+β, ao contrário do que havia acontecido com a utilização de frequências de pares de aminoácidos. No entanto, porque a classe α/β contém apenas quatro cadeias, os respectivos valores não podem ser considerados, em qualquer dos casos, muito significativos. Ao considerar as classes α/β e α+β em conjunto, embora utilizando o mesmo conjunto de regras, os resultados melhoram consideravelmente, situando-se muito acima dos valores obtidos anteriormente.

Exactidão (%):

Treino	87
Teste	76

Matriz de erro:

	α/α	β/β	αβ
α/α	6	0	4
β/β	0	9	4
αβ	0	1	13

Erros (%):	α/α	β/β	αβ
Omissão	40	31	7
Comissão	0	10	38

Exactidão (%):

Treino	84
Teste	68

Matriz de erro:

	α/α	β/β	α/β	α+β
α/α	6	0	3	1
β/β	0	9	0	4
α/β	0	0	2	2
α+β	0	1	1	8

Erros (%):	α/α	β/β	α/β	α+β
Omissão	40	31	50	20
Comissão	0	10	67	47

Figura 5.11 – Resultados: separação do conjunto de Michie reduzido em três e quatro classes estruturais, com regras de classificação.

5.4.5 Utilização das regras de classificação

As 727 cadeias do conjunto PDB_SELECT foram divididas em quatro classes estruturais, utilizando a primeira versão das regras de classificação apresentadas na secção 5.4.4.3, com o objectivo de verificar se as vantagens da separação estrutural se mantêm num conjunto maior. A tabela 5.6 especifica as percentagens de hélice, folha e outros encontrada em cada classe estrutural considerada.

Tabela 5.6 – Percentagens dos motivos estruturais nas classes estruturais do conjunto PDB_SELECT.

(%)	hélice	folha	outros
α/α	63	1	36
β/β	7	43	50
α/β	40	19	41
α+β	33	22	45
global	35	21	44

Por comparação com a tabela 5.1 (página 47), que especifica as percentagens encontradas nas classes do conjunto de Michie reduzido, existem motivos para acreditar que a separação resultante da aplicação das regras de classificação não é muito diferente da que seria efectuada pelo sistema de classificação automática de Michie, dado que os valores de ambas as tabelas são muito semelhantes.

À semelhança do procedimento adoptado na secção 5.4.2, foram efectuadas duas previsões, utilizando perceptrões multicamada constituídos por 140 sensores, 35 neurónios internos e 3 efectores cada. Aos resultados obtidos foram aplicados filtros

consistindo em perceptrões com 51 sensores, 17 neurónios internos e 3 efectores cada. A primeira previsão foi efectuada por um perceptrão, treinado durante 75 épocas, e um filtro, treinado durante 26 épocas; a segunda previsão contou com quatro perceptrões, cada um treinado somente numa das quatro classes estruturais consideradas, durante 154 (α/α), 45 (β/β), 44 (α/β) e 52 (α+β) épocas, e quatro filtros, treinados durante 16, 22, 9 e 36 épocas, respectivamente. A tabela 5.7 apresenta os resultados produzidos por ambas, nos conjuntos de teste e validação, especificando as exactidões por classe estrutural e globais.

Tabela 5.7 – Resultados: com e sem separação estrutural, no conjunto PDB_SELECT.

Exactidão (média ± desvio padrão):

(%)	α/α	β/β	α/β	α+β	global
sem separação	78±11	69±8	75±6	71±7	73±9
com separação	80±10	67±8	74±6	69±8	72±9

Curiosamente, apenas a previsão das cadeias α/α beneficia da separação em classes estruturais. As cadeias β/β, cuja previsão na secção 5.4.2 havia melhorado mais que todas as outras com a separação estrutural, aqui sofrem um decréscimo na exactidão quando classificadas pelas redes especializadas. As previsões nas classes α/β e α+β continuam a ser de qualidade superior quando classificadas pelo método generalista. No entanto, a qualidade da previsão das cadeias α+β, quando efectuada pelas redes especializadas, é bastante superior ao resultado obtido na secção 5.4.2, provavelmente devido ao aumento drástico do volume de dados disponíveis para aprendizagem. Utilizando o método especializado apenas na classe α/α, e o método generalista nas restantes, obtém-se os resultados apresentados na figura 5.12.

Exactidão (%):

Média	73
Desvio padrão	9

Erros (%):	hélice	folha	outros
Omissão	22	39	25
Comissão	23	32	28

Histograma:

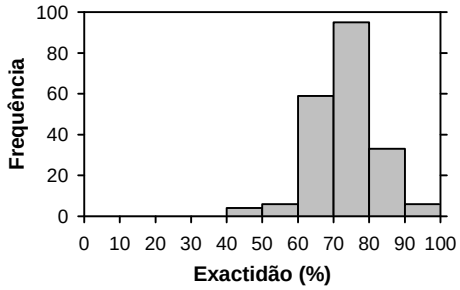


Figura 5.12 – Resultados: separação da classe estrutural α/α, no conjunto PDB_SELECT.

5.4.6 Conclusão

A separação em classes estruturais permite aumentar ligeiramente a qualidade da previsão da estrutura secundária das proteínas α/α. As proteínas α/β e α+β não beneficiam deste procedimento, e nas β/β os resultados são inconclusivos, pois

embora num conjunto de dados a separação tenha melhorado bastante a qualidade da previsão, noutro apenas a piorou.

A atribuição de classes efectuada de modo não supervisionado, utilizando frequências de aminoácidos, resultou numa separação muito divergente da separação em classes estruturais verdadeiras. Embora os aglomerados obtidos com este método mostrem agrupar diferentes características das proteínas, esta separação não permite melhorar a qualidade da previsão da sua estrutura secundária.

Com base apenas na estrutura primária, verificou-se ser possível separar as proteínas α/α e β/β das restantes, usando regras de classificação inferidas a partir de uma separação estrutural conhecida. A utilização de frequências de aminoácidos e frequências de pares de aminoácidos devolveu valores de exactidão menores, mas em que a discriminação entre as classes α/β e $\alpha+\beta$ foi efectuada com relativa facilidade, demonstrando que separar estas duas classes não é, afinal, assim tão difícil.

Considerando que a previsão da classe estrutural, quando baseada somente na estrutura primária, não permite uma separação estrutural perfeita, os erros daí decorrentes quase certamente anulariam as possíveis melhorias introduzidas pela utilização de redes especializadas. Sem mais resultados a que recorrer, conclui-se que a separação em classes estruturais, embora possível, não compensa a utilização de um número acrescido de redes neuronais, num processo cujas directrizes incluem a busca da simplicidade.

5.5 Índice de fiabilidade

5.5.1 Introdução

O PHD calcula, para cada resíduo classificado, um índice de fiabilidade que indica a confiança que o programa tem na classificação atribuída. Foi demonstrado que este índice, aqui denominado *índice PHD*, e a exactidão da previsão, estão relacionados de forma linear, e que a exactidão ultrapassa os 80% quando é considerada somente a metade dos resíduos classificados com maior fiabilidade [Rost e Sander 93]. O índice PHD consiste na diferença entre os valores das duas respostas mais elevadas dos efectores da rede neuronal. Neste trabalho utiliza-se um índice ligeiramente diferente, que consiste no índice PHD multiplicado pelo valor da resposta mais elevada. Ao contrário do índice PHD, que toma o mesmo valor independentemente das magnitudes das respostas, desde que as diferenças sejam iguais, este índice considera os dois factores. A multiplicação reduz inevitavelmente o valor de fiabilidade, pelo que este índice pode ser considerado mais pessimista, ou mais exigente, recebendo por isso o nome de *índice exigente*. Resta verificar se mantém as mesmas propriedades desejáveis do índice PHD. Todos os valores de fiabilidade foram convertidos para o intervalo entre 0 e 9.

5.5.2 Fiabilidade *versus* exactidão

5.5.2.1 Por proteína

Foram calculados os índices de fiabilidade médios de todas as proteínas do conjunto de Michie reduzido, sem separação estrutural e com separação estrutural das classes α/α e β/β (ver secção 5.4.2). A figura 5.13 mostra, para o índice exigente, a sua relação com as exactidões obtidas, onde *r* denota o coeficiente de correlação linear de Pearson.

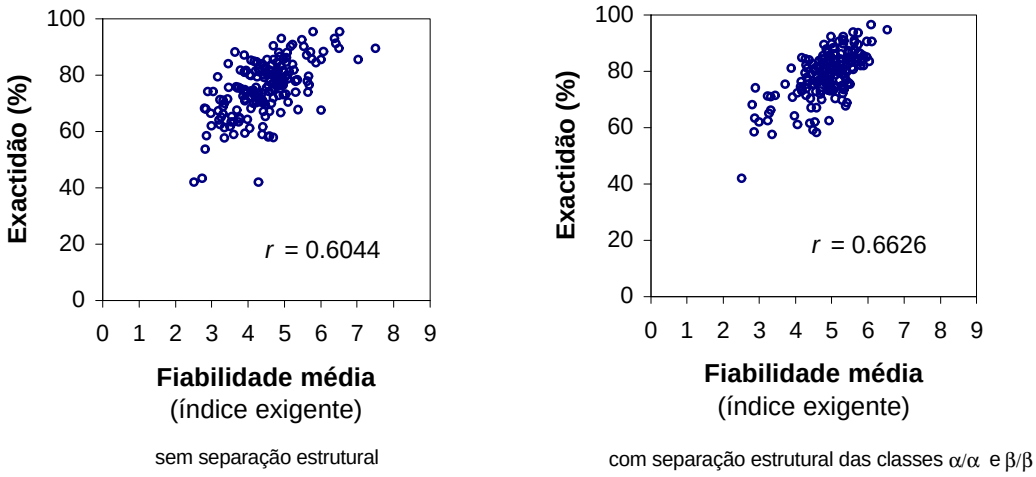


Figura 5.13 – Fiabilidade média *versus* exactidão, para o índice exigente.

Verifica-se que o coeficiente de correlação linear entre fiabilidade e exactidão, por proteína, aumenta com a separação estrutural.

A tabela 5.8 contém as correlações para os índices exigente e PHD, calculados nos conjuntos de treino e teste (184 proteínas) e apenas no conjunto de teste (37 proteínas), com separação estrutural.

Tabela 5.8 – Correlação linear entre fiabilidade média e exactidão, com separação estrutural.

	treino + teste	teste
Índice exigente	0.6626	0.8272
Índice PHD	0.6647	0.8305

Observa-se que as diferenças entre os dois índices são mínimas. Em ambos a correlação é bastante mais elevada quando se considera apenas o conjunto de teste, uma propriedade bastante agradável.

5.5.2.2 Por resíduo

Foram calculados os índices de fiabilidade (arredondados às unidades) de todos os resíduos de teste do conjunto de Michie reduzido (5333 resíduos), com separação estrutural. Para cada valor de fiabilidade, foi calculada a proporção de resíduos correctamente classificados sob a perspectiva do utilizador, *i.e.*, cuja classificação atribuída corresponde de facto ao seu motivo estrutural verdadeiro. Esta proporção representa a exactidão obtida apenas nos resíduos que apresentam esse valor de fiabilidade. Os gráficos da figura 5.14 mostram a relação entre os valores de fiabilidade e as exactidões calculadas deste modo, para ambos os índices.

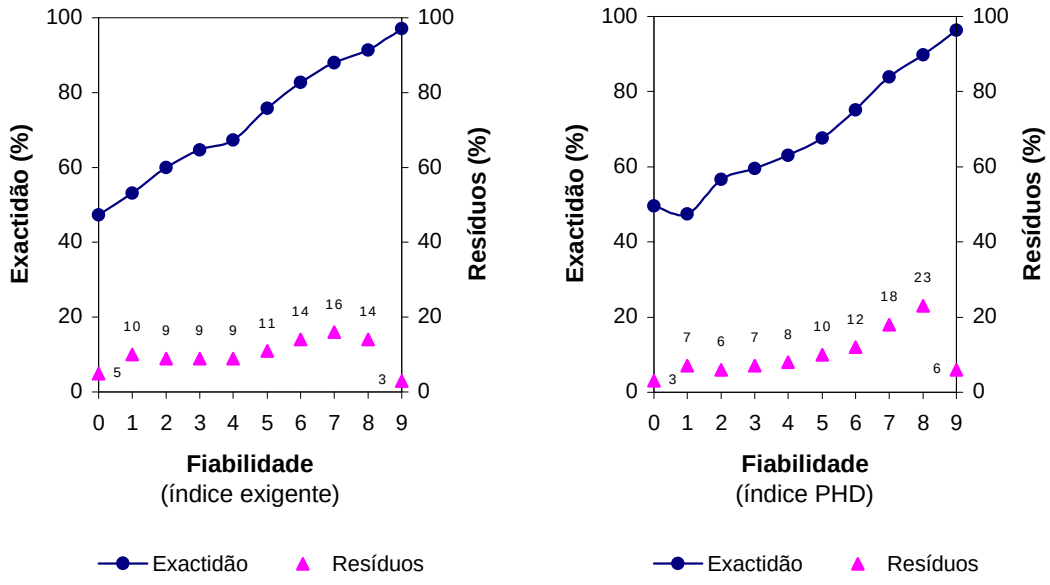


Figura 5.14 – Valor de fiabilidade *versus* exactidão para esse valor.

Apesar da elevada linearidade evidente em ambos os gráficos, existe uma pequena incoerência no caso do índice PHD, onde a exactidão dos resíduos com fiabilidade nula (50%) é mais elevada do que a exactidão dos resíduos com fiabilidade unitária (47%). Tal não se verifica no índice exigente. A distribuição dos valores de fiabilidade pelos resíduos é mais uniforme no caso do índice exigente. Tal como esperado, verifica-se o índice exigente tende a tomar valores mais baixos do que o índice PHD (42% dos resíduos apresentam valores de fiabilidade inferiores a 5 no caso exigente, contra apenas 31% no caso PHD).

A tabela 5.9 especifica os coeficientes de correlação calculados para todos os resíduos, resíduos em hélice e resíduos em folha. Verifica-se que os valores obtidos com os dois índices são, como era de esperar pela observação da figura anterior, extremamente elevados e bastante semelhantes entre si. A diferença mais significativa ocorre nos resíduos em folha, sendo no entanto inferior a 0.025.

Tabela 5.9 – Correlação linear entre valor de fiabilidade e exactidão para esse valor.

	hélice	folha	global
Índice exigente	0.9840	0.9918	0.9973
Índice PHD	0.9746	0.9687	0.9844

5.5.3 Fiabilidade mínima

Os gráficos da figura 5.15 mostram as relações observadas entre o índice de fiabilidade mínimo admitido na classificação, a percentagem de resíduos classificados e a exactidão neles obtida, para o índice exigente, em todos os resíduos de teste do conjunto de Michie reduzido.

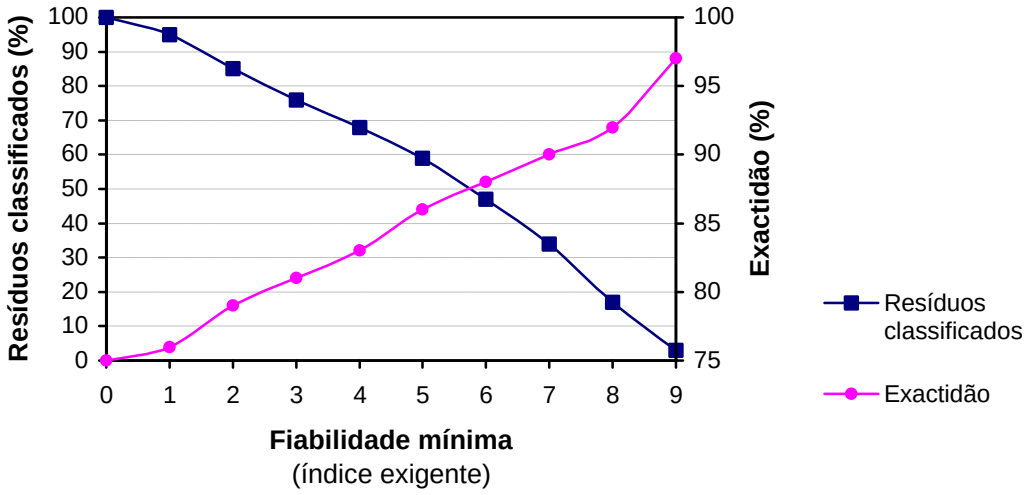


Figura 5.15 – Fiabilidade mínima *versus* resíduos classificados *versus* exactidão.

Observa-se que, caso sejam classificados apenas metade dos resíduos, a sua exactidão situa-se entre 85 e 90%, contra os cerca de 83% conseguidos pelo programa PHD no seu conjunto de teste [Rost e Sander 93]. Relembra-se, no entanto, que as relações entre as três medidas podem variar muito entre as diferentes proteínas.

5.5.4 Conclusão

O índice de fiabilidade exigente mantém as mesmas propriedades desejáveis observadas no índice utilizado no programa PHD, nomeadamente a relação linear entre o valor de fiabilidade atribuído aos resíduos e a exactidão neles obtida. Classificar apenas os resíduos com maior fiabilidade permite assim obter previsões parciais com exactidão bastante elevada.

PARTE III

6 CONSIDERAÇÕES FINAIS

Este capítulo lida com alguns aspectos importantes relacionados com a previsão da estrutura secundária de proteínas que não foram focados nos capítulos anteriores. A primeira secção aborda a relação entre a existência de homólogas conhecidas à proteína cuja estrutura secundária se pretende prever e a exactidão que nela se pode esperar obter utilizando métodos de terceira geração. Segue-se a descrição de alguns obstáculos que se colocam à previsão de estrutura de proteínas, seguindo-se um alerta para a impossibilidade de medir eficazmente a qualidade de uma previsão utilizando as medidas de exactidão mais usuais. A dissertação termina com uma breve conclusão relativamente aos resultados obtidos neste trabalho.

6.1 Homologia e exactidão

A utilização de informação evolutiva, sob a forma de alinhamentos múltiplos, permitiu aos métodos de terceira geração aumentar em mais de seis pontos percentuais a exactidão da previsão da estrutura secundária de proteínas [Rost e Sander 93]. No entanto, na eventualidade de não serem conhecidas homólogas à proteína cuja estrutura se quer prever, a ausência de alinhamentos praticamente anula essa vantagem, mesmo utilizando um método treinado com alinhamentos, como o PHD [Rost e Sander 93]. O sucesso dos métodos de terceira geração depende da existência de homólogas, o que constitui uma propriedade bastante indesejável.

No entanto, verifica-se que as novas sequências muitas vezes encontram homólogas nas bases de dados, que crescem muito rapidamente. Em particular, à medida que os projectos de sequenciação de genomas completos vão terminando, espera-se que a grande maioria das novas sequências venham a ter homólogas conhecidas. Deste modo, o calcanhar de Aquiles dos métodos de terceira geração pode vir a ser gradualmente eliminado.

6.2 Limitações

O objectivo inicial da previsão da estrutura secundária de proteínas era identificar todos os motivos de estrutura secundária com total exactidão. No entanto conclui-se que, utilizando informação sobre segmentos de resíduos, que abrangem apenas parte da sequência, esse objectivo é inatingível. O principal obstáculo consiste no facto de o mesmo segmento poder apresentar conformações diferentes, quando encontrado em proteínas diferentes. Na realidade, até mesmo sequências completas iguais podem apresentar diferenças na sua conformação que dependem, por exemplo, das propriedades do solvente. Verifica-se ainda que a variação da estrutura secundária de proteínas homólogas, embora se concentre principalmente nas extremidades das

sequências, atinge valores superiores a 10%, o que estabelece um limite superior de menos de 90% na exactidão que alguma vez se pode esperar obter com métodos de terceira geração [Rost *et al.* 94b].

Outras limitações incluem os abundantes erros que, embora tenham vindo a ser sistematicamente detectados e eliminados, podem ainda ser encontrados em algumas bases de dados públicas, nomeadamente o PDB. Estes propagam-se, não apenas à dedução da estrutura secundária, como também à construção de alinhamentos, tarefa já de si bastante difícil. Incorrecções nestes dois tipos de informação podem prejudicar o desempenho dos métodos de terceira geração, mesmo aqueles baseados em redes neuronais, apesar da sua conhecida robustez perante erros nos dados que utilizam.

6.3 Medidas de exactidão

Uma boa previsão da estrutura secundária não significa apenas um elevado grau de concordância entre os motivos estruturais verdadeiros e previstos de cada um dos resíduos. A previsão tem que ser, acima de tudo, realista, uma propriedade que as medidas de exactidão mais frequentemente utilizadas, nomeadamente o Q_3, não conseguem medir.

Uma previsão realista identifica e posiciona correctamente os motivos estruturais na sequência, mesmo que ligeiramente desfazados da sua localização verdadeira, e atribui-lhes um comprimento que se aproxima da verdade. No entanto, as medidas de exactidão mais habituais podem atribuir a uma previsão deste tipo um valor de exactidão mais baixo do que aquele obtido por uma previsão que identifica hélices onde existem folhas, e vice-versa, ou prevê padrões impossíveis de alternância entre os motivos estruturais. Existem algumas medidas de exactidão baseadas na sobreposição de segmentos de estrutura secundária [Rost *et al.* 94b], mas são ainda pouco divulgadas, motivo pelo qual não foram também utilizadas neste trabalho.

6.4 Conclusão

Durante 25 anos tem-se tentado prever a estrutura secundária de proteínas com base apenas na sua sequência. Uma tarefa aparentemente simples, mas que revelou resistir a sucessivas vagas de métodos de previsão que inicialmente conseguiam uma exactidão de 50%, e somente há cerca de cinco anos conseguiram atingir a quase mítica barreira dos 70%. Utilizando um sistema mais simples do que aquele considerado o melhor sistema de previsão de estrutura secundária disponível até ao momento, conseguiu-se neste trabalho, e com relativa facilidade, ultrapassar este valor. Embora o sistema aqui desenvolvido não tenha sido sujeito a testes rigorosos, conseguiu obter valores de exactidão semelhantes em conjuntos de dados muito diferentes. Admitindo que esses valores representam fielmente aquilo que se pode esperar obter na previsão da estrutura secundária de qualquer conjunto de proteínas, fica por determinar se um sistema mais complexo, mais parecido com o PHD, permitiria obter melhores resultados.

REFERÊNCIAS

[Anderson e Rosenfeld 98]
Anderson, J.A. e Rosenfeld, E., coords. (1998). *Talking nets: an oral history of neural networks*. Cambridge, MA: MIT Press.

[Bairoch e Apweiler 99]
Bairoch, A. e Apweiler, R. (1999). "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999." *Nucleic Acids Res.*, 27: 49-54.

[Baldi e Brunak 98]
Baldi, P. e Brunak, S. (1998). *Bioinformatics: the machine learning approach*. Cambridge, MA: MIT Press.

[Bernstein *et al.* 77]
Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. e Tasumi, M. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." *J. Mol. Biol.*, 112: 535-542.

[Bohr *et al.* 88]
Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nørskov, L., Olsen, O.H. e Petersen, S.B. (1988). "Protein secondary structure and homology by neural networks. The α-helices in rhodopsin." *FEBS Lett.*, 241: 223-228.

[Chou e Fasman 74a]
Chou, P.Y. e Fasman, G. (1974). "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins." *Biochemistry*, 13: 211-222.

[Chou e Fasman 74b]
Chou, P.Y. e Fasman, G. (1974). "Prediction of protein conformation." *Biochemistry*, 13: 222-245.

[Cohen e Cohen 94]
Cohen, B.I. e Cohen, F.E. (1994). "Predictions of protein secondary and terciary structure." In Douglas W. Smith, coord., *Biocomputing – Informatics and Genome Projects*. San Diego, CA: Academic Press. 203-232.

[Garnier *et al.* 78]
Garnier, J., Osguthorpe, D.J. e Robson, B. (1978). "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." *J. Mol. Biol.*, 120: 97-120.

[Gibrat *et al.* 87]
Gibrat, J.F., Robson, B. e Garnier, J. (1987). "Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs." *J. Mol. Biol.*, 198: 425-443.

[Hecht-Nielsen 90]
Hecht-Nielsen, R. (1990). *Neurocomputing*. Reading, PA: Addison-Wesley.

[Hobohm e Sander 94]
Hobohm, U. e Sander, C. (1994). "Enlarged representative set of protein structures." *Protein Sci.*, 3: 522-524.

[Hobohm *et al.* 92]
Hobohm, U., Scharf, M., Schneider, R. e Sander, C. (1992). "Selection of a representative set of structures from the Brookhaven Protein Data Bank." *Protein Sci.*, 1: 409-417.

[Holley e Karplus 88]
Holley, L.H. e Karplus, M. (1989). "Protein secondary structure prediction with a neural network." *Proc. Natl. Acad. Sci. USA*, 86: 152-156.

[Jensen 96]
Jensen, J.R. (1996). *Introductory digital image processing: a remote sensing perspective* (2ª ed.). Englewood Cliffs, NJ: Prentice-Hall.

[Kabsch e Sander 83]
Kabsch, W. e Sander, C. (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers*, 22: 2577-2637.

[Kohonen 84]
Kohonen, T. (1984). *Self-organization and associative memory*. Springer Series in Information Science, vol. 8. Nova Iorque, NY: Springer-Verlag.

[Levitt e Chothia 76]
Levitt, M. e Chothia, C. (1976). "Structural patterns in globular proteins." *Nature*, 261: 552-558.

[Lewin 97]
Lewin, B. (1997). *Genes VI*. Oxford, UK: University Press.

[Lippman 87]
Lippman, R.P. (1987). "An introduction to computing with neural nets." *IEEE ASSP Mag.*, 4: 4-22.

[Michie *et al.* 96]
Michie, A.D., Orengo, C.A. e Thorton, J.M. (1996). "Analysis of domain structural class using an automated class assignment protocol." *J. Mol. Biol.*, 262: 168-185.

[Nguyen e Widrow 90]
Nguyen, D. e Widrow, B. (1990). "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights." In *Proceedings of the International Joint Conference on Neural Networks, San Diego*, vol. 3. Ann Arbor, MI: IEEE. 21-26.

[Qian e Sejnowski 88]
Qian, N. e Sejnowski, T.J. (1988). "Predicting the secondary structure of globular proteins using neural network models." *J. Mol. Biol.*, 202: 865-884.

[Riis e Krogh 96]
Riis, S.K. e Krogh, A. (1996). "Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments." *J. Comp. Biol.*, 3: 163-183.

[Rost e Sander 93]
Rost, B. e Sander, C. (1993). "Prediction of protein secondary structure at better than 70% accuracy." *J. Mol. Biol.*, 232: 584-599.

[Rost e Sander 94]
Rost, B. e Sander, C. (1994). "Combining evolutionary information and neural networks to predict protein secondary structure." *Proteins*, 19: 55-72.

[Rost *et al.* 94a]
Rost, B., Sander, C. e Schneider, R. (1994). "PHD - an automatic mail server for protein secondary structure prediction." *CABIOS*, 10: 53-60.

[Rost *et al.* 94b]
Rost, B., Sander, C. e Schneider, R. (1994). "Redefining the goals of protein secondary structure prediction." *J. Mol. Biol.*, 235: 13-26.

[Rumelhart *et al.* 86]
Rumelhart, D.E., Hinton, G.E. e Williams, R.J. (1986). "Learning internal representations by error propagation." In D.E. Rumelhart *et al.*, coords., *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1. Cambridge, MA: MIT Press. 318-362.

[Sander e Schneider 91]
Sander, C. e Schneider, R. (1991). "Database of homology-derived protein structures and the structural meaning of sequence alignment." *Proteins*, 9: 56-68.

[Schulze-Kremer 95]
Schulze-Kremer, S. (1995). *Molecular bioinformatics: algorithms and applications*. Nova Iorque, NY: Walter de Gruyter.

[Sejnowski e Rosenberg 87]
Sejnowski, T.J. e Rosenberg, C.R. (1987). "Parallel networks that learn to pronounce English text." *Complex Systems*, 1: 145-168.

[Werbos 74]
Werbos, P.J. (1974). “Beyond regression: new tools for prediction and analysis in the behavioral science.” Ph.D. Thesis, Harvard University, Cambridge, MA.

ÍNDICE REMISSIVO

(*itálico*: autores; **bold**: figuras)

S

T

V

W