UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
APLICADA - PIPCA

Extração de Regras de Redes Neurais Artificiais Aplicadas ao Problema da Previsão da Estrutura Secundária de Proteínas

por

EDUARDO BATTISTELLA

Dissertação submetida à avaliação
como requisito parcial para a obtenção do grau de
Mestre em Computação Aplicada

Prof. Dr. Adelmo Luis Cechin
Orientador

São Leopoldo, abril de 2004.

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Battistella, Eduardo

Extração de Regras de Redes Neurais Artificiais Aplicadas ao Problema da Previsão da Estrutura Secundária de Proteínas / por Eduardo Battistella. — São Leopoldo: Ciências Exatas e Tecnológicas da UNISINOS, 2004.

119 f.: il.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos. Ciências Exatas e TecnológicasPrograma Interdisciplinar de Pós-Graduação em Computação Aplicada - PIPCA, São Leopoldo, BR–RS, 2004. Orientador: Cechin, Adelmo Luis.

1. Rede Neural Artificial. 2. Extração de Regras. 3. Data Minning. 4. Dobramento de Proteínas. I. Cechin, Adelmo Luis. II. Título.

"É mais freqüente que a confiança seja gerada pela ignorância do que pelo conhecimento: são os que conhecem pouco, e não os que conhecem muito, os que afirmam tão positivamente que este ou aquele problema nunca será solucionado pela ciência."

Charles Darwin

Agradecimentos

Inicialmente, desejo agradecer, e muito, ao Adelmo. Pelo profissionalismo, pelo apoio, pela paciência, pela camaradagem e, por algo que deveria ser comum no meio acadêmico... por personificar a idéia de que ensinar é mostrar aos outros que eles sabem tanto quanto quem ensina.

Não podia deixar de agradecer ao meu pai e minha mãe, pelo $uporte nesta fase da minha vida; ao meu amigo Motta, pela "mão amiga nas horas difíceis"; e por fim, mas não menos importante, a Adriana, pela sua inigualável companhia.

Agradeço também aos professores Ney Lemke e Fernando Osório, pela conviência diária, pela contribuição dada nas disciplinas do mestrado e no seminário de andamento; e, finalmente, à CAPES e à UNISINOS, pela bolsa de estudos concedida.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

DSSP	Define Secondary Structure of Proteins
FAGNIS	Fuzzy Automatically Generated Neural Inferred System
HMM	Hidden Markov Model
HP	Hydropathic Index
k-FCV	k-Fold-Cross-Validation
PDB	Protein Data Bank
PESP	Previsão da Estrutura Secundária de Proteínas
pI	Isoeletric point
RMS	Root Mean Square
RNA	Rede Neural Artificial
RNAMM	Rede Neural Artificial Multi-Modal
RNM	Ressonância Nuclear Magnética
RProp	Resilient Propagation
SVM	Support Vector Machine

Resumo

Extração de Regras de Redes Neurais Artificiais Aplicadas ao Problema da Previsão da Estrutura Secundária de Proteínas apresenta o estudo feito sobre a extração de conhecimento de Redes Neurais na forma de regras difusas. Na aplicação desta técnica, foi utilizado o problema da classificação da estrutura secundária de proteínas, em *alfa, beta* e *coil*, a partir da estrutura primária.

Serão apresentadas as implementações feitas para viabilizar esta tarefa. Dentre elas: a implementação de recursos adicionais ao software de extração de regras; a definição de uma metodologia de extração de regras; a implementação desta metodologia; e a análise das regras extraídas.

Dentre os recursos implementados no processo de extração, será visto que o foco principal foi o de embasar o conhecimento extraído sobre um suporte estatístico e disponibilizar medidas complementares para a sua avaliação.

Na definição da metodologia, será visto que cuidados devem ser tomados na preparação da base de dados e na definição da estrutura da rede. Cuidados que foram seguidos e devidamente apresentados na fase de implementação.

Será visto que a implementação da metodologia e a análise das regras procurou validar qualitativamente o processo, comparando o conhecimento extraído com um já existente bem como fornecendo um novo. Neste processo, foi utilizada a abordagem tradicional de codificação ortogonal dos resíduos e uma proposta de utilização de atributos físico-químicos (*i.e.* ponto isoelétrico e índice hidropático) considerados válidos e importantes ao contexto.

Por fim, espera-se que o conhecimento disponibilizado seja analisado por profissionais de competência para tal, visando fomentar a discussão sobre a sua validade e do processo pelo qual foram geradas. E que o retorno da análise sirva para o aprimoramento da técnica utilizada.

Palavras-chave: Rede Neural Artificial, Extração de Regras, Data Minning, Dobramento de Proteínas.

TITLE: "RULE EXTRACTION FROM ARTIFICIAL NEURAL NETWORKS APPLIED TO THE PROBLEM OF PROTEIN SECONDARY STRUCTURE PREDICTION"

Abstract

This work presents a study about knowledge extraction from Neural Networks in the form of fuzzy rules. In the application of this technique, it was investigated the problem of classification of the protein secondary structure (alpha, beta and coil) from its primary structure.

The implementations that make possible this task will be presented. Amongst them: the implementation of new features in the rule extraction software; the definition of a methodology for the rule extraction process; the implementation of this methodology; and the analysis of the extracted rules.

Amongst the implemented features in the rule extration process, it will be noticed that the main point was to provide a statistical support for the knowledge extracted and to make available additional resources to measure this information.

In the definition of this methodology, it will be seen that some considerations must be observed in the database preparation and in definition of the network structure. Observations that had been followed and properly presented in the implementation phase.

It will be observed that the implementation of this methodology and the analysis of the rules took in consideration a qualitative validation of the process by comparing the extracted knowledge against a previous one as well supplying a new one. This process used a traditional orthogonal codification of the residues and proposed the use of physical-chemical attributes (i.e. isoeletric point and hidrophatic index) considered important in this context.

Finally, it is expected that qualified professionals analyze the knowledge obtained, aiming at to instigate a debate about its validity and the process by which they had been generated. And that the results of this analysis serves for the improvement of the used technique.

Keywords: Artificial Neural Network, Rule Extraction, Data Minning, Protein Folding.

Capítulo 1

Introdução

As proteínas estão no centro da ação nos processos biológicos. Sua importância e sua notável gama de atividades podem ser exemplificadas em funções como catálise enzimática, movimento muscular, sustentação mecânica, proteção imunitária, geração e transmissão de impulsos nervosos e várias mais.

Conforme Voet *et al.* [1], uma lista completa de funções conhecidas das proteínas teria dezenas de itens, incluindo proteínas que transportam outras moléculas. Uma tal lista deixaria de fora milhares de proteínas cujas funções ainda não estão inteiramente elucidadas ou, em muitos casos, são mesmo completamente desconhecidas.

Uma peça chave para decifrar a função de uma dada proteína é o entendimento da sua estrutura. Assim como as outras principais macromoléculas biológicas (*i.e.* ácidos nucléicos e polissacarídeos), as proteínas são polímeros formados por moléculas menores. Porém, ao contrário dos ácidos nucléicos, as proteínas não apresentam estruturas regulares e uniformes. Isso se deve, em parte, ao fato de os 20 aminoácidos dos quais as proteínas são feitas apresentarem propriedades físicas e químicas muito distintas. Examinando-se como estes aminoácidos estão enfileirados em uma proteína, pode-se tentar entender as propriedades físicas e químicas das proteínas e, finalmente, seus mecanismos de ação nos seres vivos.

Atualmente, as proteínas têm sua estrutura determinada através de técnicas de raio-X ou Ressonância Nuclear Magnética (RNM). A utilização de raio-X requer a cristalização da proteína, o que pode ser uma tarefa relativamente difícil. Adicionalmente, os padrões de difração podem não ser interpretáveis. Entretanto, mesmo os padrões sendo interpretáveis, pode-se levar até meses para determinar a estrutura de uma única proteína.

A RNM, por sua vez, não requer a cristalização da proteína. Trata-se de uma alternativa útil para a determinação da estrutura de pequenas proteínas, uma vez que seu custo computacional

é elevado.

Uma análise superficial dos métodos convencionais expostos leva a uma constatação simples: métodos computacionais eficientes para a predição da estrutura de proteínas são altamente desejáveis devido ao alto custo, financeiro e de tempo, dos métodos de laboratório.

Desde que, em 1973, o estudo de Anfinsen [2] lançou a hipótese de que a estrutura de uma proteína pode ser unicamente determinada a partir da estrutura primária (*i.e.* seqüência de aminoácidos), uma série de trabalhos têm sido desenvolvidos na tentativa de corroborar tal afirmação.

Conforme Mount [3], um dos maiores objetivos da Bioinformática é entender a relação entre a seqüência de aminoácidos de uma proteína e a sua estrutura. Se esta relação for conhecida, então a estrutura de uma proteína pode ser prevista de modo seguro. Lamentavelmente, a relação entre a seqüência e estrutura não é tão simples. Um grande progresso tem sido feito na categorização de proteínas com base na sua seqüência e este tipo de informação é muito útil na modelagem de proteínas.

Conforme Holbrook *et al.* [4], um dos maiores focos na previsão da estrutura secundária de uma proteína é puramente empírico. Baseado em bancos de dados de cujas proteínas se conhece a seqüência e a estrutura. Esta abordagem espera encontrar características comuns nestes bancos de dados que possam ser generalizadas a ponto de fornecer modelos estruturais para estas proteínas.

Na busca por métodos computacionais que implementem tal solução, algumas linhas de pesquisa se desenvolveram mais notadamente. Dentre elas, podem ser citadas a determinação da estrutura através da análise de seqüências homólogas, as técnicas de simulação que usam princípios *ab initio* e as técnicas de aprendizado de máquina como as Redes Neurais Artificiais (RNAs), redes Bayesianas, *Hidden Markov Models* (HMMs) e, mais recentemente, as *Support Vector Machines* (SVMs).

No contexto da determinação da estrutura de proteínas, onde o número de seqüências é muito superior ao número de estruturas já determinadas, a utilização de RNAs mostra-se uma técnica adequada. A explicação para tal reside no fato das RNAs aprenderem a partir dos exemplos recebidos e exibirem uma capacidade de generalizar este aprendizado para dados ainda não vistos. Esta característica faz desta técnica uma ótima candidata a processar dados de domínios para os quais se tenham pouco ou incompleto conhecimento do problema a ser resolvido, mas onde existam dados de treinamento suficientes para a projeção de um modelo.

O problema da determinação da estrutura envolve, como será visto no decorrer deste estudo, basicamente a identificação de 3 formas estruturais (classes). Duas destas formas, apresentam

padrões repetitivos sendo denominadas de *alfa* e *beta*. A terceira, denominada de *coil*, não apresenta um padrão de comportamento. Embora possam ser encontrados na literatura exemplos de RNAs que trabalhem com a determinação de mais classes (*i.e* especializações/combinações dos citados), a maioria tenta determinar apenas estas.

Uma análise dos trabalhos envolvendo a aplicação de RNAs no problema da predição da estrutura secundária de proteínas (PESP) apresenta um ponto em comum que chama a atenção. Notadamente pode ser observado que há um direcionamento específico na busca por melhores percentuais de exatidão do resultado obtido.

Na tentativa de uma melhor determinação da estrutura das proteínas, variações no trabalho de Anfinsen [2] têm proposto, com relativo sucesso, a utilização de atributos físico-químicos de cada aminoácido da cadeia ao invés de uma codificação ortogonal dos mesmos. Por exemplo, atributos como hidrofobicidade, massa, volume e outros vêm sendo combinados e testados. Estes testes visam, unicamente, um melhor resultado na predição da estrutura.

A busca pelos 100% de exatidão, inatingívies pois o sentimento "comum" dos pesquisadores é de que as RNAs deverão chegar a no máximo 90%, tem se tornado incessante. A utilização de RNAs cada vez mais complexas, como as recorrentes bi-direcionais, e o uso mais intenso de SVMs visam, unicamente, a busca pelo ótimo. Perseguir melhores resultados é uma meta válida, mas não é a única.

Outro ponto que chama a atenção nos trabalhos hoje existentes, de aplicação de RNAs ao problema da PESP, é a não exploração da aquisição de conhecimento adquirido pelas redes. Mesmo com a derrubada gradual do rótulo de “caixa-preta” atribuído às RNAs, através da disponibilização de métodos de extração do conhecimento aprendido, isto não tem sido feito.

Como conseqüência, perde-se uma grande oportunidade de utilizar este conhecimento em prol da ciência. Há muito, biólogos moleculares, e outros profissionais da área, já poderiam estar tirando proveito destas informações. Mesmo uma negação do conhecimento disponibilizado pelas RNAs seria útil, pois ajudaria a melhorar o processo de extração do conhecimento da mesma.

Dentro deste contexto, o problema da PESP foi tratado como uma oportunidade para a disponibilização do conhecimento aquirido pelas RNAs. Nesta disponibilização de conhecimento, dois pontos principais foram definidos. Primeiro, como resultado do processamento da rede, foi feita a determinação das classes (*i.e. alfa, beta* e *coil*). Segundo, foram utilizados os atributos físico-químicos denominados ponto isoelétrico (*isoeletric point (pI)*) e índice hidropático (*hydropathic index (HP)*) para representar a cadeia de aminoácidos. A partir do que foi definido, pretendeu-se obter um conhecimento que caracterizasse o comportamento dos atributos utilizados na formação das classes estruturais.

Para viabilizar esta tarefa, uma série de procedimentos foram executados. Primeiro, uma base de dados foi devidamente preparada. Cuidados relativos a confiabilidade e integridade dos dados foram ser observados. Neste processo, foi elencado um conjunto de atributos físico-químicos que representaram cada aminoácido da seqüência que compõe a proteína. Adicionalmente, procurou-se utilizar seqüências de proteínas não-homólogas visando uma maior exploração do espaço de entrada e conseqüente eficiência da rede treinada.

Segundo, a arquitetura da rede foi definida. Cuidados foram observados com relação a codificação das entradas no que diz respeito a utilização da técnica de janelamento (vista em detalhes no item 3.8). Testes foram feitos visando identificar o número de neurônios da camada oculta que atendam as necessidades deste estudo. No processo de treinamento, foi empregada uma rede tipo *Multilayer Perceptron* (MLP) *feedforward*.

Terceiro, a partir de uma análise do que foi aprendido pela rede, regras difusas foram extraídas utilizando-se o sistema *Fuzzy Automatically Generated Neural Inferred System* (FAGNIS), proposto por Cechin[5]. O sistema FAGNIS foi escolhido por implementar os algoritmos de treinamento de RNAs e permitir a extração de regras diretamente da RNA treinada. Além disso, o sistema resulta em um conjunto de regras tipo Takagi-Sugeno [6], que apresentam a característica de descrever o comportamento da RNA em uma simples função linear e permitir a extração da influência das entradas nas saídas. Tem-se, como resultado desta abordagem, a extração de menos regras com maior compreensibilidade.

Quarto, uma vez obtido o conhecimento, na forma de regras, o mesmo necessita ser validado. A validação em laboratório seria inviável no que diz respeito ao ponto de vista econômico. Adicionalmente, não haveria tempo hábil para os experimentos serem desenvolvidos. A validação estatística, através do processo de estimação de parâmetros, apresentou-se como uma alternativa adequada e válida cientificamente. Para tanto, o FAGNIS foi alterado para fornecer regras estatisticamente válidas. Adicionalmente, implementações foram feitas visando dotar o software com a capacidade de fornecer mais informações relativas a cada regra. Informações como: o percentual de padrões cobertos por cada regra; o percentual de erro na classificação dos padrões de treinamento e um acompanhamento global das variações no número de regras, erro e cobertura durante o processo de treinamento da RNA.

Por fim, o conhecimento descoberto (*i.e.* as regras), foi disponibilizado com a intenção de fomentar a discussão sobre a validade das regras obtidas bem como do processo pelo qual foram geradas.

Esta dissertação está assim estruturada. No capítulo 2 será apresentada uma compilação dos principais conceitos sobre Biologia Molecular, encontrados na literatura, cuja exposição se

faz necessária para uma leitura satisfatória deste estudo. O mesmo será feito, no capítulo 3, para os principais conceitos sobre Redes Neurais Artificiais e Lógica Difusa. No capítulo 4 serão apresentados os principais trabalhos na área de predição da estrutura secundária de proteínas. As implementações realizadas no FAGNIS, a definição de uma metodologia para o processo de extração de regras e a implementação desta metodologia serão apresentadas no capítulo 5. No capítulo 6 os resultados obtidos serão apresentados. Por fim, no capítulo 7, serão apresentadas as conclusões sobre este estudo e sugestões de implementações futuras que foram detectadas no desenvolver deste trabalho.

Capítulo 2

Conceitos Básicos sobre Biologia Molecular

A Biologia Molecular, como qualquer outra ciência moderna, depende de instrumentos para dissecar a arquitetura e a operação de sistemas inacessíveis aos sentidos humanos.

Além do instrumental físico e químico para separar, quantificar e analisar o material biológico, os biólogos moleculares também se valem de recursos computacionais cada dia mais sofisticados.

Neste capítulo será apresentada uma compilação dos principais conceitos sobre a Biologia Molecular necessários à compreensão deste estudo.

2.1 Definição

A Biologia Molecular compreende as disciplinas de bioquímica e biofísica. Pode ser definida como o estudo da química e da física aplicadas à vida.

Apesar de sobrepor-se a outras disciplinas (*e.g.* biologia celular, genética, farmacologia) é limitada a um número de indagações que incluem, dentre elas:

- Qual a composição química e estrutura tridimensional das moléculas biológicas?
- Como as moléculas biológicas interagem?
- Quais os mecanismos de organização das moléculas biológicas e de coordenação das atividades?

2.2 Aminoácidos

Os aminoácidos são as unidades estruturais básicas das proteínas. Os 20 tipos de aminoácidos presentes nas proteínas são também frequentemente referenciados como *aminoácidos-padrão*. Esta nomenclatura tem a intenção de diferenciá-los dos outros aminoácidos presentes nos organismos vivos mas não nas proteínas.

Estes 20 aminoácidos também são chamados de *α-aminoácidos*. A eles foram associadas abreviações comuns de uma e três letras, conforme pode ser visto na tabela 2.1. Nesta tabela são apresentados os mesmos termos em inglês (nas colunas *Abreviation* e *Identification*) devido a constante referência aos mesmos encontrados na literatura nacional.

TABELA 2.1 – Nomenclatura dos Aminoácidos (em português e inglês)

Letra	Abreviação	Identificação	*Abreviation*	*Identification*
A	ALA	Alanina	ALA	Alanine
C	CIS	Cisteina	CYS	Cysteine
D	ASP	Aspartato	ASP	Aspartate
E	GLU	Glutamato	GLU	Glutamate
F	FEN	Fenilalanina	PHE	Phenylalanine
G	GLI	Glicina	GLY	Glycine
H	HIS	Histidina	HIS	Histidine
I	ILE	Isoleucina	ILE	Isoleucine
K	LIS	Lisina	LYS	Lysine
L	LEU	Leucina	LEU	Leucine
M	MET	Metionina	MET	Methionine
N	ASN	Asparagina	ASN	Aspargine
P	PRO	Prolina	PRO	Proline
Q	GLN	Glutamina	GLN	Glutamine
R	ARG	Arginina	ARG	Arginine
S	SER	Serina	SER	Serine
T	TRE	Treonina	THR	Threonine
V	VAL	Valina	VAL	Valine
W	TRP	Triptofano	TRP	Tryptophan
Y	TIR	Tirosina	TYR	Tyrosine

A título de curiosidade histórica, conforme Lehninger [7], o primeiro aminoácido a ser descoberto foi a Asparagina, em 1806. O último dos 20, em 1938, foi a Treonina. Todos os aminoácidos possuem nomes comuns e, em alguns casos, derivados da fonte onde foram descobertos. A As-

paragina, por exemplo, foi descoberta no aspargo; o Glutamato no glúten; a Tirosina no queijo (do grego, *tyros*) e a Glicina recebeu este nome por ser doce (do grego, *glykos*).

Todas as proteínas, desde as bactérias mais antigas até as mais complexas formas de vida, são constituidas do mesmo conjunto de 20 aminoácidos. O que as difere é a quantidade de aminoácidos, o tipo dos aminoácidos e a seqüência em que os aminoácidos estão unidos. Isto leva a um número imensamente grande de combinações possíveis.

2.2.1 Estrutura dos Aminoácidos

Conforme pode ser visto na figura 2.1 (a), os aminoácidos possuem um grupo carboxila ($COOH$), um grupo amino (NH_2) e um átomo de hidrogênio ligados ao mesmo átomo de carbono denominado de *α-carbono*. A figura 2.1 (a) apresenta a forma não ionizada de um aminoácido e a figura 2.1 (b) apresenta o aminoácido em solução em pH neutro na forma ionte dipolar. Estas duas formas de representação da estrutura geral de um aminoácido se alternam na literatura em geral.

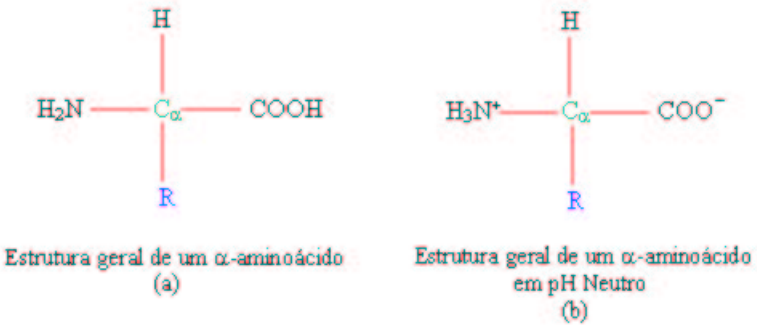

FIGURA 2.1 – Estrutura geral de um aminoácido: (a) estrutura em pH não neutro; (b) estrutura em pH neutro

O que diferencia os 20 *α-aminoácidos* é a sua cadeia lateral (ou grupo R, de Radical) que varia em estrutura, tamanho, carga elétrica, capacidade de formação de pontes de hidrogênio e que influencia a solubilidade do aminoácido na água. Os *α-aminoácidos* são normalmente classificados em 5 classes principais baseadas nas propriedades dos grupos R, em particular, sua polaridade ou tendência a interagir com a água em pH próximo a 7. A figura 2.2 (adaptada a partir de Lehninger [7]) apresenta as formas estruturais dos *α-aminoácidos* divididos nas 5 classes principais.

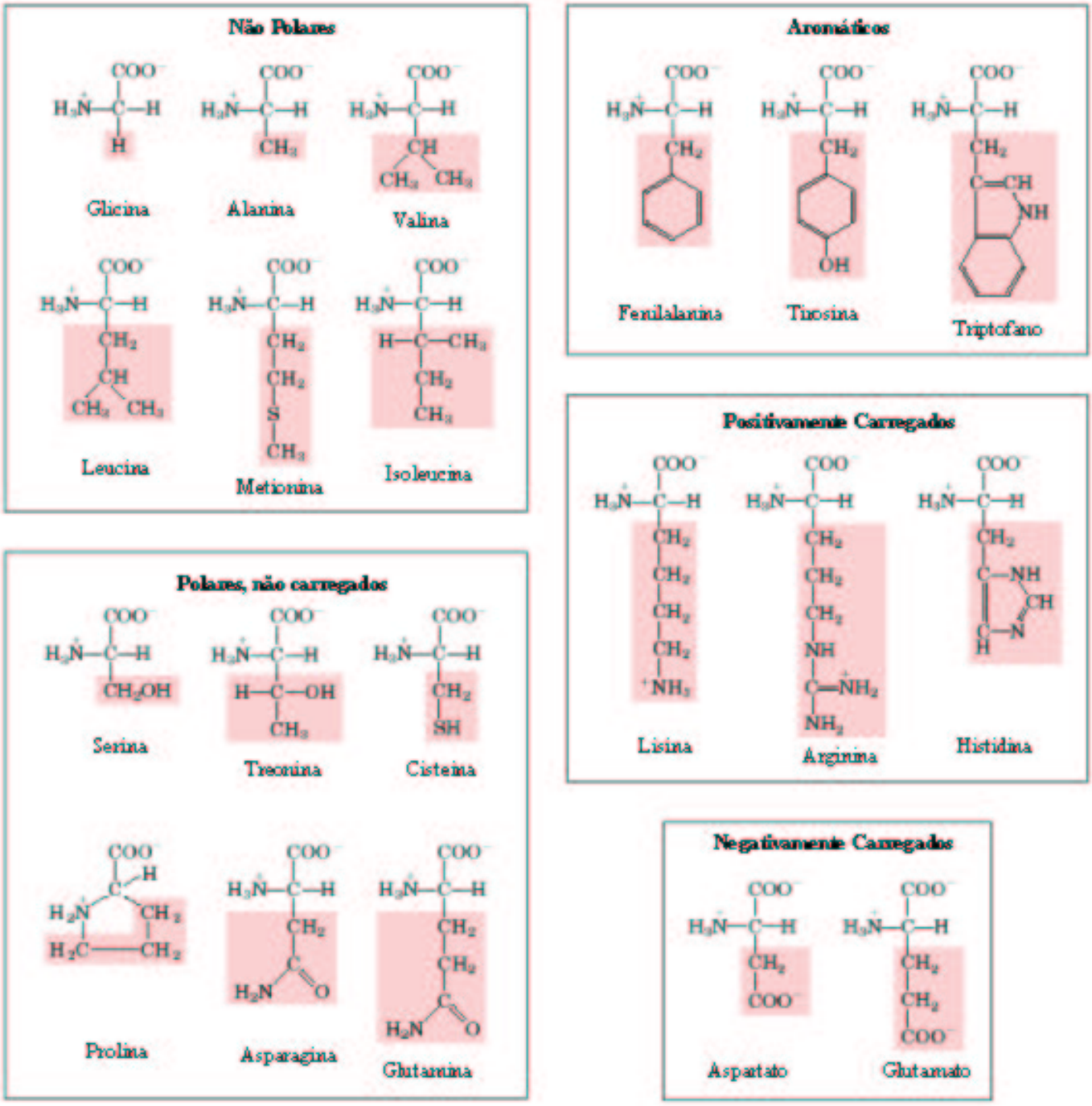

FIGURA 2.2 – Estrutura dos 20 α-*aminoácidos* divididos nas 5 classes principais

2.2.2 Atributos Físico-Químicos dos Aminoácidos

Conforme Schlick [8], cada aminoácido possui uma combinação única de propriedades que afetam, de forma crítica, as interações que formam e estabilizam a estrutura tridimensional das proteínas.

Dentre estas propriedades, podem ser citadas:

- volume encerrado pelo raio de van der Waals;
- massa - também encontrada na literatura como sendo o peso molecular do aminoácido com ou sem a molécula de água;
- área da superfície;
- polaridade;
- refratividade;
- forças eletrostáticas;
- hidrofobicidade;
- hidrofilicidade;
- grau de hidrofobicidade da cadeia lateral (escala HP)
- ponto isoelétrico (pI)
- pK do grupo carboxila;
- pK do grupo amino;
- pK do radical.

A tabela 2.2 (compilada a partir de Wu & McLarty [9], Kyte & Doolittle [10], Lehninger [7] e Voet *et al.* [1]) apresenta os valores de algumas das propriedades acima citadas. Uma lista mais extensa pode ser encontrada em [11].

TABELA 2.2 – Atributos Físico-Químicos dos Aminoácidos

Aminoácido	Volume (A^3)	Massa (*daltons*)	HP	pI	$pK_{Carboxila}$	pK_{Amino}	$pK_{Radical}$
Alanina	67	71,09	1,8	6.01	2,34	9,69	
Arginina	148	156,19	-4,5	10,76	2,17	9,04	12,48
Asparagina	96	114,11	-3,5	5,41	2,02	8,80	
Aspartato	91	115,09	-3,5	2,77	1,88	9,60	3,65
Cisteina	86	103,15	2,5	5,07	1,96	10,28	8,18
Glutamina	114	128,14	-3,5	5,65	2,17	9,13	
Glutamato	109	129,12	-3,5	3,22	2,19	9,67	4,25
Glicina	48	57,05	-0,4	5,97	2,34	9,60	
Histidina	118	137,14	-3,2	7,59	1,82	9,17	6,00
Isoleucina	124	113,16	4,5	6,02	2,36	9,68	
Leucina	124	113,16	3,8	5,98	2,36	9,60	
Lisina	135	128,17	-3,9	9,74	2,18	8,95	10,53
Metionina	124	131,19	1,9	5,74	2,28	9,21	
Fenilalanina	135	147,18	2,8	5,48	1,83	9,13	
Prolina	90	97,12	-1,6	6,48	1,99	10,96	
Serina	73	87,08	-0,8	5,68	2,21	9,15	
Treonina	93	101,11	-0,7	5,87	2,11	9,62	
Triptofano	163	186,12	-0,9	5,89	2,38	9,39	
Tirosina	141	163,18	-1,3	5,66	2,20	9,11	10,07
Valina	105	99,14	4,2	5,97	2,32	9,62	

2.3 Proteínas

Conforme Lehninger [7], proteínas são moléculas formadas pela união de vários aminoácidos. Ocorrem em todas as células e em todas as partes das células.

As moléculas resultantes da união de aminoácidos são genericamente chamadas peptídios. Termo este que tem sua origem no grego *pepsis* que significa digestão. Isto se deve ao fato de os sucos estomacais dos animais quebrarem justamente as ligações entre os aminoácidos. Dois aminoácidos formam um dipeptídeo, três formam um tripeptídeo e assim por diante. O termo polipeptídeo (do grego *poli* = muitos) é comumente utilizado para se referir à moléculas formadas por muitos aminoácidos (classe esta onde a maioria das proteínas se enquadra).

Segundo Stryer [12], as proteínas exercem papéis cruciais em virtualmente todos os processos biológicos. Sua importância e sua notável gama de atividades são exemplificadas em funções como as de: catálise enzimática, transporte e armazenamento, movimento muscular, sustentação

mecânica, proteção imunitária, geração e transmissão de impulsos nervosos, *etc.* A título de curiosidade histórica, o termo proteína é derivado da palavra grega *proteios*, que significa "de primeira ordem". Foi cunhado por Jöns J. Berzelius, em 1838, para salientar a importância desta classe de moléculas.

As tarefas de descrever e compreender a estrutura das proteínas são abordadas, normalmente, em 4 níveis de hierarquia conceitual, conforme pode ser visto na figura 2.3 (extraída de Lehninger [7]). A descrição de todas as ligações covalentes que ligam os resíduos em uma cadeia polipeptídica é chamada de estrutura primária. O elemento mais importante de uma estrutura primária é a seqüência dos resíduos. A estrutura secundária se refere a um arranjo particular estável dos resíduos originando padrões estruturais recorrentes. A estrutura terciária descreve todos os aspectos do dobramento tridimensional de um polipeptídeo. Quando a proteína possui mais do que um polipeptídeo, a sua organização no espaço é referenciada como a estrutura quaternária.

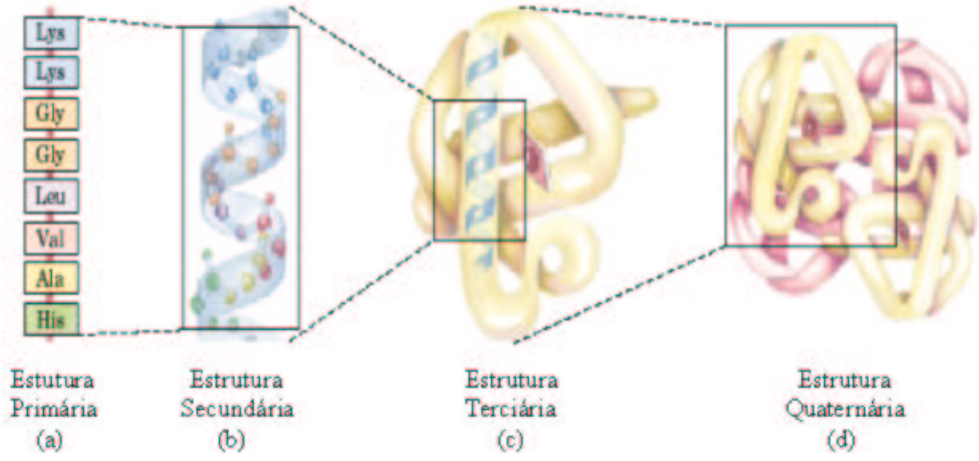

FIGURA 2.3 – Formas de representação estrutural das proteínas.

Destas, nos ateremos às estruturas primária e secundária pois são as pertinentes ao estudo desenvolvido nesta dissertação.

2.3.1 Estrutura Primária das Proteínas

Conforme Voet *et al.* [1], a estrutura primária de uma proteína consiste na seqüência de aminoácidos de sua cadeia polipeptídica ou das suas cadeias polipeptídicas (caso seja constituída por mais de uma cadeia). Constantemente, encontram-se referências ao termo "seqüência de

resíduos" como também utilizado para caracterizar a estrutura primária.

Segundo Stryer [12], o conhecimento da seqüência primária é importante por diversos motivos. Primeiro, para elucidar seu mecanismo de ação (*e.g.* o mecanismo catalítico de uma enzima). Proteínas com novas propriedades podem ser geradas pela alteração de seqüências estruturais conhecidas. Segundo, as análises das relações entre seqüências de aminoácidos e estruturas tridimensionais de proteínas podem vir a revelar as regras que governam o enovelamento de cadeias polipeptídicas. Terceiro, a determinação da seqüência faz parte da patologia molecular. Doenças fatais, como a anemia falciforme e a fibrose cística, podem resultar da alteração de um só aminoácido em uma só proteína. Quarto, a seqüência de uma proteína revela muito acerca de sua história evolutiva.

2.3.2 Estrutura Secundária das Proteínas

Conforme Voet *et al.* [1], a estrutura secundária é o arranjo espacial dos átomos de um esqueleto polipeptídico, sem levar em consideração a conformação de suas cadeias laterais.

Os polipeptídeos podem possuir, conforme Schlick [8], uma variedade de conformações (i.e. estruturas tridimensionais) diferindo apenas pelas orientações rotacionais sobre suas ligações covalentes. Este tipo de flexibilidade rotacional é caracterizada pelos ângulos diedros (também denominados ângulos de torção ou ângulos de rotação). Conforme pode ser visto na figura 2.4, os ângulos diedros ϕ e ψ são usados, respectivamente, para definir rotações em torno da ligação entre os átomos *α-carbono* e nitrogênio do grupo amino (*i.e.* $\phi : C_\alpha - N$) e entre os átomos *α-carbono* e carbono do grupo carboxila (*i.e.* $\psi : C_\alpha - C$).

O ângulo diedro ω define a rotação sobre a ligação peptídica. Devido ao caráter parcial de dupla ligação e as interações estéreis entre cadeias laterais adjacentes, o ângulo ω está tipicamente na configuração *trans* (*i.e.* $\omega = 180°$). Comumente, o tratamento dado na literatura, com relação aos ângulos diedros, apresenta uma visão sem o ângulo ω.

Os valores permitidos dos ângulos ϕ e ψ podem ser calculados. Conformações espaciais proibidas têm os valores de ϕ e ψ que trariam os átomos mais próximos do que as distâncias de van der Walls correspondentes. Tal informação é resumida no diagrama de Ramachandran, assim denominado devido ao seu inventor, G. N. Ramachandran [13]. Conforme pode ser observado na figura 2.5 (extraída de Lehninger [7]) , a maioria das áreas do diagrama de Ramachandran (a maioria das combinações de ϕ e ψ) representa combinações proibidas da cadeia polipeptídica.

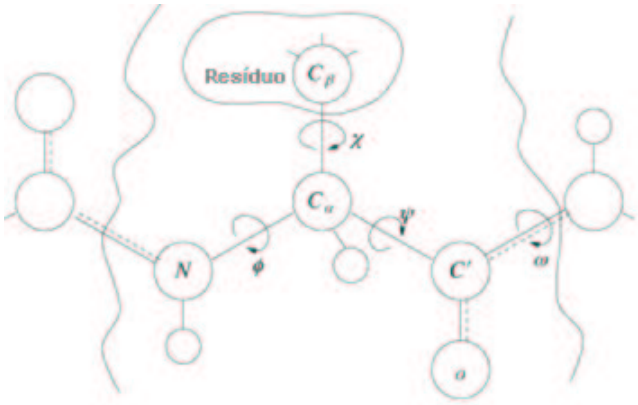

FIGURA 2.4 – Flexibilidade rotacional nos polipeptídeos: definição dos ângulos diedros ϕ, ψ e ω.

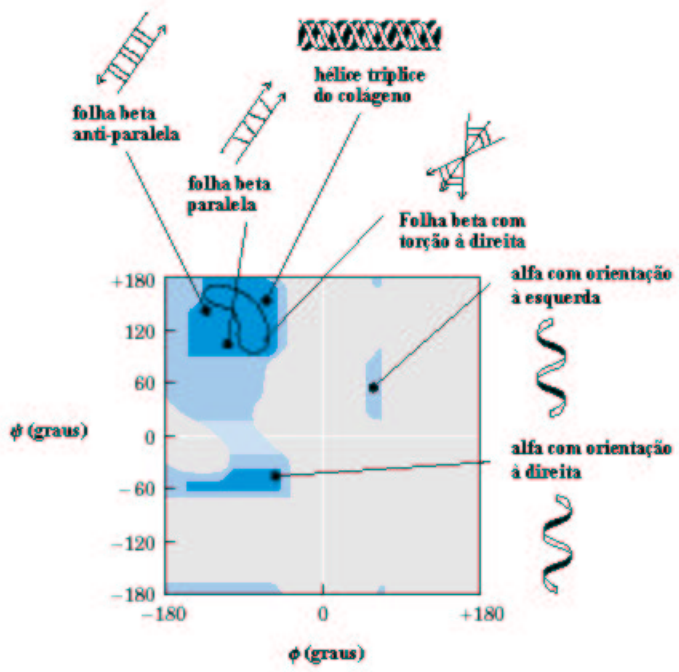

FIGURA 2.5 – Diagrama de Ramachandran.

O arranjo secundário de um polipeptídeo pode ocorrer de forma regular. Isso acontece quando os ângulos das ligações entre carbonos e seus ligantes (os valores de ϕ e ψ) são iguais e se repetem ao longo de um segmento da molécula. São 2 os tipos principais de estruturas

secundárias regulares: *α-hélices* e *folhas β*.

A *α-hélice* (também referenciada como alfa ou hélice) é a forma mais comum de estrutura secundária regular. Conforme pode ser visto na figura 2.6 (extraída de Lehninger [7]), caracteriza-se por uma hélice em espiral formada por 3,6 resíduos de aminoácidos por volta. As cadeias laterais dos aminoácidos se distribuem para fora da hélice, evitando assim o impedimento estérico. A principal força de estabilização da *α-hélice* é a ponte de hidrogênio.

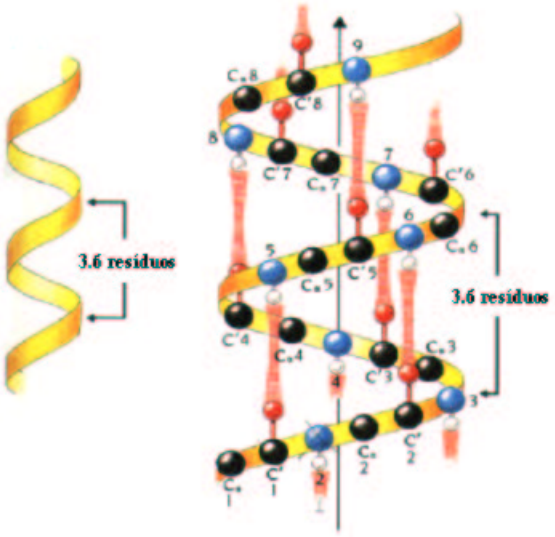

FIGURA 2.6 – Representação de uma *α-hélice* com periodicidade de 3,6 resíduos por volta.

A *folha β* (também referenciada como beta, folha pregueada), ao contrário da *α-hélice*, conforme pode ser visto na figura 2.7 (extraída de Voet *et al.* [1]), envolve 2 ou mais segmentos polipeptídicos da mesma molécula ou de moléculas diferentes, arranjados em paralelo ou no sentido anti-paralelo. Os segmentos em *folha β* da proteína adquirem um aspecto de uma folha de papel dobrada em pregas. As pontes de hidrogênio mais uma vez são a força de estabilização principal desta estrutura.

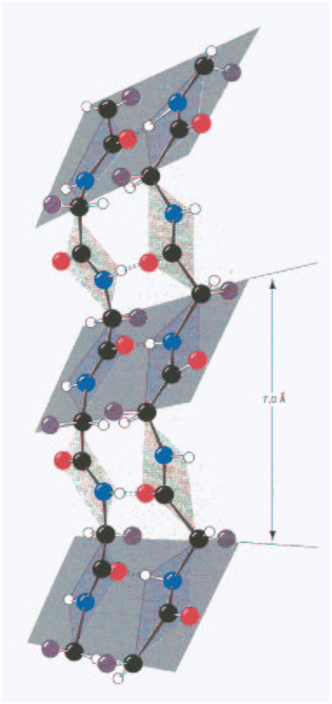

FIGURA 2.7 – Representação de uma *folha* β.

Além das duas estrutura secundárias regulares citadas, conforme pode ser visto na figura 2.8 (feita utilizando-se o software RASMOL), o restante da molécula assume uma estrutura secundária não repetitiva normalmente denominada de *coil*, ou *random coil*.

FIGURA 2.8 – Representação da proteína 1gp1 (*glutathione peroxidase* do boi) onde constam várias α-*hélices* e duas *folha* β, além das várias *coils*.

2.4 Encerramento do Capítulo

Este capítulo apresentou uma compilação dos principais conceitos encontrados na literatura cuja exposição se faz necessária para uma leitura satisfatória desta dissertação.

Foram vistos conceitos pertinentes à Biologia Molecular, especificamente sobre bioquímica. Dentre eles: os aminoácidos, as proteínas e suas conformações estruturais.

No capítulo seguinte será feita uma exposição dos principais conceitos, pertinentes a este trabalho, sobre Redes Neurais Artificiais e Lógica Difusa.

Capítulo 3

Conceitos Básicos sobre Redes Neurais Artificiais e Lógica Difusa

Neste capítulo serão apresentados os principais conceitos sobre RNAs bem como uma introdução à Lógica Difusa. Alguns destes conceitos estão ligados ao entendimento da revisão bibliográfica feita sobre a utilização de RNAs aplicadas ao problema da PESP.

Além da definição de uma RNA propriamente dita, serão apresentados alguns dos seus principais conceitos: neurônios, RNAs multicamadas, o processo de aprendizado, os algoritmos de aprendizado, os tipos de aprendizado, os tipos de problemas tratados pelas RNAs. Adicionalmente, serão vistos alguns dos principais conceitos na intersecção entre RNAs e suas aplicações em Biologia Molecular bem como alguns conceitos genéricos sobre a extração de regras de RNAs.

No que diz respeito à Lógica Difusa, veremos a sua definição, principais características e vantagens.

3.1 Definição de uma RNA

Conforme Haykin [14], uma RNA é um processador maciçamente paralelo e distribuído, constituído de unidades de processamento simples, que têm a função de armazenar conhecimento experimental e torná-lo disponível para o uso.

Assemelha-se ao cérebro em dois aspectos:

- o conhecimento é adquirido pela rede a partir de seu ambiente através de um processo chamado de aprendizagem onde a forma de perceber o ambiente é através de um conjunto de padrões;

- forças de conexão entre as unidades de processamento (neurônios), chamadas pesos sinápticos (ou simplesmente pesos), são utilizadas para armazenar o conhecimento adquirido.

Conforme será visto nas explicações que seguem, o principal atrativo na estrutura de uma RNA reside em sua habilidade de adaptação e aprendizagem. Isto significa que modelos de RNAs podem lidar com dados imprecisos e situações não totalmente definidas. Uma rede treinada de maneira adequada tem a habilidade de generalizar quando é apresentada a entradas que não estão presentes em dados já conhecidos por ela.

3.2 Modelo de Neurônio

Conforme Wu & McLarty [9], o neurônio é a unidade fundamental de processamento de informação de uma RNA. Trata-se, na verdade, de uma representação simplificada de um neurônio biológico.

Os neurônios da RNA, assim como os neurônios do cérebro, possuem conexões de entrada (dendritos) e de saída (axônios). Adicionalmente, também como os neurônios reais, possuem uma forma de processamento interno que gera um sinal de saída em função de um sinal de entrada. Entretanto, enquanto a saída de um neurônio biológico está em constante alteração no tempo, a de um neurônio artificial muda somente em intervalos discretos no tempo, isto é, quando os dados de entrada mudam.

As conexões entre os neurônios possuem pesos sinápticos associados a elas. Também chamados simplesmente de pesos, representam o resultado do aprendizado do processamento da RNA.

A figura 3.1 apresenta o modelo de um neurônio artificial, onde podem ser observados: os sinais de entrada, o conjunto de sinapses, o *bias*, o somador e a função de ativação.

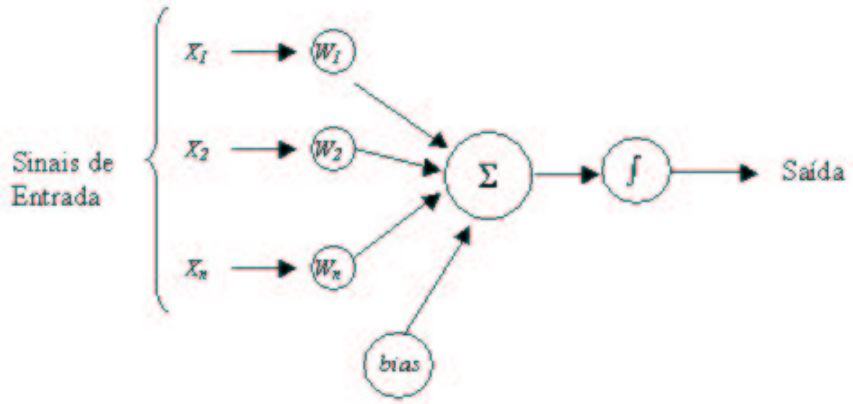

FIGURA 3.1 – Modelo não-linear de um neurônio.

Os *sinais de entrada* representam o mapeamento das entradas para a saída. Cada sinal mapeia um dos atributos contido no arquivo de padrões.

O *conjunto de sinapses* ou *conexões*, cada uma caracterizada por um peso sináptico, tem a função de armazenar o conhecimento sobre os sinais de entrada processados.

O *bias* tem a função de diminuir ou aumentar o valor a ser processado pela função de ativação. Possui o efeito de transladar a função de ativação em torno da origem. Considerando-se duas entradas, um neurônio sem o *bias* é como uma equação da reta sem o termo independente, ou seja, sempre passará pela origem.

O *somador* é um operador com a função de calcular a soma dos sinais de entrada multiplicados pelos respectivos pesos sinápticos.

A *função de ativação* é utilizada para restringir a amplitude do valor de saída do neurônio. É responsável por definir a ativação de saída do neurônio em termos do seu nível de ativação interna. Em outras palavras, o valor de saída em função do somatório das entradas multiplicados pelos seus pesos. Várias são as funções de ativação existentes (e.g. de Limiar, Linear por Partes, Sigmóide, Identidade) sendo que os intervalos de saída, normalmente, são definidos entre -1 e 1, 0 e 1 ou -0.5 e 0.5.

3.3 RNAs Multicamadas

As RNAs multicamadas são arquiteturas onde os neurônios são organizados em duas ou mais camadas de processamento.

As RNAs com apenas duas camadas são constituídas de uma camada de entrada que se conecta a uma camada de neurônios de saída. Os neurônios da camada de entrada são neurônios especiais, cujo papel é exclusivamente distribuir cada uma das entradas da rede (sem modificá-las) a todos os neurônios da camada seguinte. A forma mais simples deste tipo de rede consiste de um único neurônio na camada de saída, sendo conhecido como *perceptron*. Um exemplo de rede com apenas duas camadas é apresentado na Figura 3.2 (a), com 3 neurônios na camada de entrada e 4 neurônios na camada de saída.

Conforme Haykin [14], o *perceptron* foi objeto de intensa pesquisa durante os anos 50 e 60, mas em 1969, M. Minsky e S. Papert provaram matematicamente que este tipo de estrutura de processamento apresenta limitações importantes e só pode ser aplicada com sucesso a uma classe muito restrita de problemas. Mais especificamente foi provado que o perceptron é capaz de resolver apenas problemas linearmente separáveis.

No entanto, com a utilização de redes de múltiplas camadas, com pelo menos uma camada oculta (camada que não é nem entrada, nem saída), muitas das limitações apresentadas pelo *perceptron* deixam de existir. Esta implementação, cujo exemplo pode ser visto na figura 3.2 (b), recebeu o nome de *Multilayer Perceptron* (MLP).

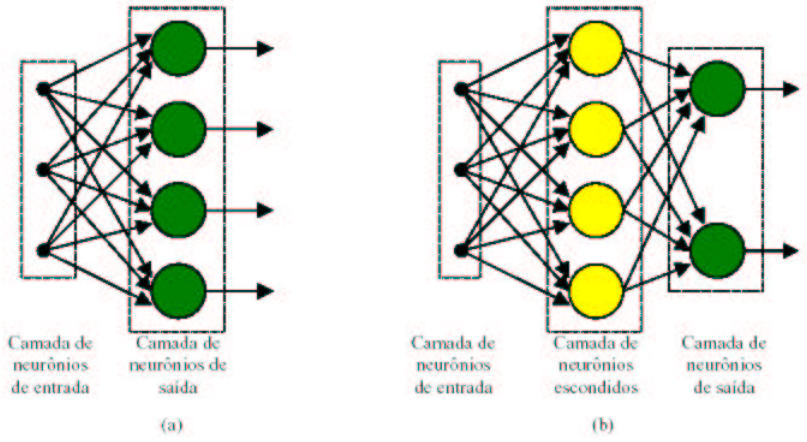

FIGURA 3.2 – RNAs em camadas: (a) rede com duas camadas (uma de entrada e outra de saída); (b) rede com três camadas (uma de entrada, uma oculta e outra de saída.)

3.4 Aprendizado

Conforme Wu & McLarty [9], a idéia fundamental por trás do aprendizado (também chamado de treinamento) de uma RNA é atribuir valores a um conjunto de pesos (inicializado normalmente de forma aleatória), aplicar os padrões à rede e verificar como a mesma responde com estes pesos. Se não responder de forma satisfatória então os pesos devem ser modificados por algum tipo de algoritmo (específico de cada arquitetura) e o processo deve ser repetido. Esta interação deve continuar até que um critério de parada pré-definido seja atingido.

Cada passagem de treinamento sobre todos os padrões é chamada de época. Alterações nos pesos podem ser feitas a cada padrão processado ou após uma época inteira. Normalmente, os pesos são modificados após cada época.

O objetivo do treinamento é gerar uma rede com pesos que melhor atendam aos padrões recebidos de forma a generalizar a solução obtida para dados que venham a ser submetidos posteriormente. Para que isto aconteça, é necessário que seja evitada a situação de *overtraining* onde a rede "memoriza" os padrões recebidos e perde o seu poder de generalização. Para tanto, junto com a base de treinamento, é processada uma base de validação que não altera os pesos da rede mas valida o que foi aprendido até o momento. O ponto no qual a rede (com seus devidos pesos) deve ser salva é quando o erro da validação começa a subir.

Conforme Han & Kamber [15], a preparação inadequada dos conjuntos de arquivos de treinamento e validação pode levar a resultados errôneos devido, principalmente, a situação de *overtraining*. Para evitar este problema, diferentes técnicas são empregadas na preparação destes. Normalmente, o arquivo original de padrões sofre um particionamento dos seus dados o que leva a geração dos arquivos para as etapas de treinamento e validação. Dentre os principais métodos, para a geração destes arquivos, podem ser citados: *holdout, k-fold-cross-validation* e *jackknife*.

O método de *holdout* consiste em separar, de forma aleatória, o arquivo de padrões em dois arquivos. O de treinamento tipicamente conterá dois terços dos dados e o de validação o um terço restante.

O *k-fold-cross-validation* (*k-FCV*) é um método onde o arquivo de padrões é dividido em k arquivos mutuamente exclusivos, todos de igual tamanho. As etapas de treinamento e a validação são repetidas k vezes, sendo utilizados para treinamento *k-1* arquivos e para validação o *k-ésimo* arquivo (não utilizado no treinamento). A cada interação, o arquivo de validação possui um k diferente.

O método de *jackknife*, também conhecido como *leave-one-out*, é semelhante ao *k-FCV*

onde o k é igual ao número de linhas do arquivo de padrões. Com isto, cada arquivo de validação conterá somente uma linha em cada etapa do processo.

Uma variação dos métodos de *holdout* e *k-FCV* é a inclusão da técnica de estratificação. Neste caso, há a preocupação de que cada classe, existente no arquivo de padrões, seja proporcionalmente representada nos arquivos de treinamento e validação. Com isto, conforme Wu & McLarty [9], é eliminada a representação tendenciosa dos dados, um dos principais problemas no treinamento das RNAs.

3.5 Algoritmos de Aprendizado

O algoritmo de aprendizado é um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de RNAs. Estes algoritmos diferem entre si principalmente pelo modo como os pesos da rede são ajustados.

Dentre os principais algoritmos de aprendizado, podem ser citados, entre outros: *Backpropagation, Resilient Propagation (RProp), Cascade Correlation, Kohonen* e *Quickprop.*

Expressões matemáticas, chamadas de regras de aprendizado, descrevem o processo de aprendizado implementado por cada algoritmo. Uma das regras de aprendizado mais conhecidas é a regra *Least Mean Square* (LMS). Sua finalidade é minimizar o erro médio quadrático. Os pesos sinápticos da rede são ajustados de acordo com o erro quadrático para todos os padrões do conjunto de treinamento. O processo de redução gradativa do erro tende à convergência, onde o erro é estável, levando ao encerramento do aprendizado. A evolução do processo de aprendizagem ocorre, até que algum critério seja satisfeito, como um valor mínimo de erro global ou uma diferença sucessiva mínima entre erros.

3.6 Tipos de Aprendizado

A aprendizagem da RNA nada mais é do que o processo de ajustes sucessivos nos pesos das conexões. Ocorre quando, o padrão final corresponde ao que se deseja associar, como resposta ao padrão de entrada. As RNAs possuem dois tipos de aprendizado: supervisionado e não supervisionado.

No aprendizado *supervisionado*, a RNA é treinada através da apresentação de exemplos onde, para os atributos de entrada, a rede deva produzir a saída informada. A resposta calculada pela RNA é comparada com a saída esperada (também fornecida no padrão) e, assim, a rede gera

um valor de erro que corresponde à diferença entre os dois valores. O valor do erro obtido é então utilizado para calcular o ajuste necessário aos pesos sinápticos da rede, os quais serão corrigidos até que a resposta da rede se aproxime da saída desejada. Esse é o processo de minimização do erro. Nesse tipo de aprendizado, os cálculos necessários para minimizar o erro são importantes e estão atrelados ao algoritmo utilizado.

No aprendizado *não supervisionado* a saída desejada não existe. A rede é treinada através de padrões de entrada e então, arbitrariamente, efetua uma separação automática dos padrões de modo a produzir grupos o mais homogêneos possíveis.

3.7 Tipos de Problemas que as RNAs Solucionam

As RNAs atendem a dois tipos de problemas: de classificação e de aproximação de função.

Problemas de classificação, como o nome indica, são aqueles onde a RNA tenta classificar os padrões em classes distintas. Estas classes podem ser conhecidas ou não. Se conhecidas, utiliza-se o método de aprendizado supervisionado. Se não conhecidas, utiliza-se o método não supervisionado. Este, por sua vez, irá gerar as classes e associar os padrões às mesmas.

Em *problemas de aproximação de função*, para o método de aprendizado supervisionado, a RNA executa procedimentos visando gerar uma saída que mais se aproxime do valor esperado.

3.8 Conceitos sobre RNAs Aplicadas à Biologia Molecular

Complementando os conceitos básicos tradicionais sobre RNAs, faz-se necessário uma contextualização quanto à sua aplicação em problemas de biologia molecular. Isto se deve ao fato de que, no decorrer dos anos de pesquisa na área, técnicas e convenções foram adotadas e, hoje, são consideradas de domínio comum. Alguns dos conceitos que serão vistos se confundem entre intrínsecos das RNAs e oriundos do problema a ser resolvido.

O primeiro ponto que merece ser destacado é a constante mescla entre os termos *aminoácido* e *resíduo*. A literatura sobre RNAs aplicadas à problemas de biologia molecular comumente utiliza um destes termos. Muitas vezes, ambos são encontrados em um mesmo texto com a intenção de apresentar o mesmo significado. Este trabalho dará preferência, ao termo *resíduo* pois, como já citado anteriormente, este termo reflete a perda do elemento água quando um aminoácido se junta a outro e, em se tratando determinação da estrutura secundária, a partir da primária, a água não se faz presente.

Dentre os outros ítens que merecem análise, estão: a representação dos resíduos; o jane-

lamento; o resultado produzido pela RNA; a métrica sobre os resultados. Conforme poderá ser notado, expressões como "tipicamente", "normalmente" e "comumente" serão empregadas com uma certa freqüência nos parágrafos que seguem. Isto se deve ao fato de que as explicações dadas são baseadas na maioria dos casos observados e não na sua totalidade. Para cada possibilidade apresentada, várias alternativas oriundas da criatividade humana podem ser encontradas na literatura.

3.8.1 A Representação dos Resíduos

Cada resíduo de uma seqüência normalmente é representado por um código ou um conjunto de atributos físico-químicos que o caracterizam.

Com relação ao *código*, a representação mais utilizada é composta por 20 entradas binárias onde cada uma representa um resíduo. Por exemplo, a Alanina poderia ser representada por 00000000000000000001, a Arginina por 00000000000000000010 e assim por diante, terminando pela Valina representada por 10000000000000000000. Esta representação também é encontrada na literatura com o nome de codificação ortogonal.

Com relação aos *atributos físico-químicos*, vários são utilizados. Sua codificação oscila entre entradas binárias, entradas reais não normalizadas e entradas reais normalizadas. Conforme Wu & McLarty [9], alguns dos atributos encontrados com mais freqüência em pesquisas que utilizam RNAs são: hidrofobicidade, volume, massa, área, propensidade a pertencer a uma determinada estrutura secundária, refratividade. Conforme Baldi & Brunak [16], ainda podem ser acrescentados a esta lista: carga, família, distâncias do resíduo com relação as extremidades da proteína, *etc.*

3.8.2 O Janelamento

Conforme Holbrook *et al.* [4], é bastante provável a hipótese de que, na determinação da estrutura secundária, os resíduos intrinsicamente possuem certas preferências conformacionais e, estas preferências, podem ser influenciadas pelos resíduos próximos. Valendo-se desta informação, as RNAs têm sido desenhadas para prever a estrutura secundária a partir do contexto em que o resíduo se encontra. Com a intenção de capturar estas informações de curta distância, o conceito de janelamento é utilizado.

Com esta técnica, é criada uma "janela" virtual que vai "deslizando" sobre a seqüência de resíduos da proteína, percorrendo toda a sua extensão. Cada padrão contém informações sobre todos os resíduos "vistos" pela janela em uma determinada posição. A janela tem sempre

comprimento ímpar e o resíduo central é o que está tendo a sua estrutura determinada. A intenção é que a informação dos resíduos laterais ajudem nesta determinação.

A figura 3.3 apresenta um exemplo da técnica de janelamento. Como pode ser visto, a janela possui um tamanho de 5 resíduos. O processamento de uma proteína com tamanho de 8 resíduos resultou na geração de 4 padrões a serem processados pela RNA.

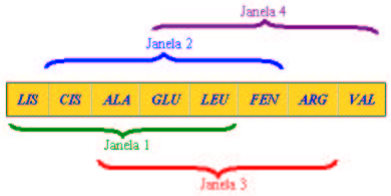

FIGURA 3.3 – Deslizamento de uma janela de 5 resíduos sobre uma proteína de 8 resíduos.

3.8.3 O Resultado Produzido pela RNA

O resultado gerado pela RNA está diretamente ligado a arquitetura implementada e esta, por sua vez, ao problema que se deseja resolver.

A primeira grande divisão que ocorre nas implementações de RNAs para o problema da PESP é quanto ao número de classes que a arquitetura tenta prever. As mais comuns são as de 3 classes: *alfa*, *beta* e *coil*. Algumas implementações fazem uma classificação mais detalhada como, por exemplo, a divisão entre alfa direita e esquerda e/ou entre folhas beta paralelas e anti-paralelas.

Uma vez definido o número de classes, que normalmente corresponde ao número de saídas da rede, ocorre agora a definição do tipo de resultado. Tipicamente, são 2 os tipos de resultados reportados: binário e proporcional. O resultado binário apresenta 1 na saída que caracteriza a estrutura e 0 nas demais. O resultado proporcional apresenta um valor em cada saída. A que possuir o maior valor caracteriza a estrutura vencedora. É a chamada técnica de *winner-take-all*. Em algumas implementações, o valor proporcional é igual ao conceito valor percentual.

3.8.4 A Métrica dos Resultados

Uma etapa importante no processo de avaliação dos resultados obtidos por uma RNA é a definição de uma métrica sobre a qualidade da predição feita. Conforme Baldi & Brunak [16], uma variedade de abordagens para calcular a exatidão da RNA tem sido sugeridas em diferentes

contextos e isto acabou por gerar alguma confusão na comparação entre os diferentes estudos.

Conforme Wu & McLarty [9], a performance pode, normalmente, ser medida por: sensibilidade (*sensitivity*), singularidade (*specificity*), predição positiva (*positive predictive value*), predição negativa (*negative predictive value*), exatidão (*accuracy*) e coeficiente de correlação (*correlation coefficient*).

As equações que seguem apresentam a implementação de cada métrica:

$$Sensibilidade = \frac{V_p}{V_p + F_n} \tag{3.1}$$

$$Singularidade = \frac{V_n}{V_n + F_p} \tag{3.2}$$

$$Predição\ Positiva = \frac{V_p}{V_p + F_p} \tag{3.3}$$

$$Predição\ Negativa = \frac{V_n}{V_n + F_n} \tag{3.4}$$

$$Exatidão = \frac{V_p + V_n}{V_p + V_n + F_p + F_n} \tag{3.5}$$

$$Coeficiente\ de\ Correlação = \frac{(V_p V_n) - (F_p F_n)}{\sqrt{(V_p + F_p)\,(F_p + V_n)\,(V_n + F_n)\,(F_n + V_p)}} \tag{3.6}$$

onde: *Vp* são os *Verdadeiros Positivos* (classificados como positivos e que são realmente positivos); *Fp* são os *Falsos Positivos* (classificados como positivos e que são, na verdade, negativos); *Vn* são os *Verdadeiros Negativos* (classificados como negativos e que são realmente negativos); *Fn* são os *Falsos Negativos* (classificados como negativos e que são, na verdade, positivos).

A *sensibilidade* (equação 3.1) é a proporção de todos os padrões corretamente classificados como verdadeiros. A *singularidade* (equação 3.2) é a proporção de todos os padrões corretamente

classificados como falsos. Eles podem ser considerados como uma métrica de quão bem os *falsos negativos* e *falsos positivos* são eliminados.

A *predição positiva* (equação 3.3) é a probabilidade de um padrão identificado como verdadeiro ser realmente verdadeiro. Já a *predição negativa* (equação 3.4) é a probabilidade de um padrão identificado como falso ser realmente falso.

A *exatidão* (equação 3.5) é o percentual de todas as predições corretas.

O *coeficiente de correlação* (equação 3.6) foi introduzido por Matthews [17]. Os valores entre 1 e -1 correspondem, respectivamente, a quão correta e quão completamente errada é a predição.

Outra forma de se apresentar os resultados encontrados é através de uma *matriz de exatidão* (conforme tabela 3.1).

TABELA 3.1 – *Matriz de exatidão 3x3* para a predição da estrutura secundária

A	α	β	L	Esperado
α	$A_{\alpha\alpha}$	$A_{\alpha\beta}$	$A_{\alpha L}$	b_α
β	$A_{\beta\alpha}$	$A_{\beta\beta}$	$A_{\beta L}$	b_β
L	$A_{L\alpha}$	$A_{L\beta}$	A_{LL}	b_L
Encontrado	a_α	a_β	a_L	N

$$a_i = \sum_{j=1}^{3} A_{ji}, \qquad para\, i = \alpha, \beta, L \tag{3.7}$$

$$b_i = \sum_{j=1}^{3} A_{ij}, \qquad para\, i = \alpha, \beta, L \tag{3.8}$$

$$N = \sum_{j=1}^{3} a_j = \sum_{j=1}^{3} b_j \tag{3.9}$$

$$Q_i = Q_i^{\%obs} = 100\frac{A_{ii}}{b_i} \qquad para\, i = \alpha, \beta, L \tag{3.10}$$

$$Q_i^{\%pred} = 100\frac{A_{ii}}{a_i} \qquad para\, i = \alpha, \beta, L \tag{3.11}$$

$$Q_3 = \frac{100}{N}\sum_{i=1}^{3} A_{ii} \tag{3.12}$$

onde A_{ij} é o número de resíduos preditos na estrutura i e observados na estrutura j; a_i e b_i são as freqüências preditas e observadas dos resíduos em uma dada estrutura; N é o número total de resíduos analisados; Q_i é a percentagem de resíduos corretamente preditos, em relação aos observados, de uma dada estrutura; $Q_i^{\%pred}$ é o percentual de resíduos corretamente preditos, dentre todos os resíduos preditos, de uma dada estrutura; e Q_3 é a exatidão percentual de todos os resíduos corretamente preditos.

Da mesma forma que a matriz as equações acima foram utilizadas para apresentar os resultados da classificação em 3 classes (*alfa, beta* e *loop/coil*), as mesmas são utilizadas quando a classificação se faz em um número diferente de classes.

De todas as métricas acima, as três mais utilizadas são: a *exatidão* (equação 3.5), o *coeficiente de correlação* (equação 3.6) e a *matriz de exatidão* (tabela 3.1).

3.9 Extração de Regras de RNAs

Uma das características mais importantes das RNAs é a sua capacidade de adaptação e aprendizagem, a partir de uma base de treinamento, levando à generalização da solução encontrada. Isto abriu novas perspectivas para a automatização da aquisição de conhecimento.

Durante muito tempo, as RNAs foram tidas como um tipo de "enigma numérico". Muitas vezes denominadas de "caixa-preta", em particular por fornecer ao usuário nenhuma informação sobre o conhecimento adquirido.

Visando reparar esta situação, um esforço considerável tem sido empregado no problema de suprir as RNAs com uma capacidade de explanação. Em particular, uma parte substancial deste esforço tem sido direcionada em uma linha de investigação que gira em torno do desenvolvimento de técnicas para a extração do conhecimento adquirido através de regras expressas em uma linguagem compreensível ao usuário. Estas regras, utilizadas por vários métodos, incluem: regras de inferência (*if-then-else*), árvores de decisão, regras difusas (utilizadas neste estudo) e outras.

Para a extração destas regras, a granularidade na análise do conhecimento adquirido pelas RNAs, segundo Andrews *et al.* [18] e Tickle *et al.* [19], possui dois extremos. De um lado, a visão *decomposicional* onde a rede é analisada em um nível menor de granularidade, isto é, no nível dos neurônios das camadas oculta e de saída e nos pesos sinápticos. Do outro lado, a visão *pedagógica* onde a rede é vista em um nível máximo de granularidade. Onde a análise se dá simplesmente pela influência das entradas nas saídas. Desta forma, mantendo um certo *status* de "caixa-preta".

Conforme Andrews *et al.* [18] e Tickle *et al.* [19], independente do tipo de regra gerada, algumas características devem ser buscadas em todas as soluções propostas. Dentre elas: exatidão, fidelidade, consistência e compreensibilidade. Por exatidão entenda-se o grau com o qual o conjunto de regras extraídos é capaz de classificar exemplos "não-vistos" de forma correta. A fidelidade indica o grau de similaridade entre as regras extraídas e a RNA a partir da qual se originaram. A consistência denota o grau com que, sob diferentes treinamentos, a RNA gere regras que produzam a mesma classificação para os casos "não-vistos". Por fim, a compreensibilidade é dada por uma medida do tamanho do conjunto de regras extraídas em termos do número de regras e do número de antecedentes por regra.

A partir da extração do conhecimento interno das RNAs, e valendo-se das características que as regras devem apresentar, alguns benefícios podem ser facilmente enumerados:

- podem ser descobertos novos relacionamentos e/ou características importantes a partir das regras extraídas;
- formalismo para expressar o conhecimento;
- a capacidade de gerar explicações das decisões tomadas, em nível interno pela RNA, pode facilitar a aceitação do uso da rede pelos usuários;
- as regras, na forma *if... then...* ou árvores de decisão, facilitam a integração com sistemas simbólicos baseados em conhecimento;
- por expressarem o conhecimento adquirido pela rede, as regras simbólicas podem ser utilizadas para descobrir em que situações a rede pode cometer erros de generalização;
- podem possibilitar a identificação de regiões no espaço de entrada que não se fizeram representar no conjunto de treinamento.

Para uma melhor compreensão sobre o tipo de regras extraídas das RNAs, que este estudo utiliza, faz-se necessária uma introdução ao assunto Lógica Difusa.

3.10 Lógica Difusa

A Lógica Difusa pode ser definida como sendo uma ferramenta capaz de capturar informações vagas, em geral descritas em uma linguagem natural e convertê-las para um formato numérico de fácil manipulação. Com base na teoria dos Conjuntos Nebulosos (*Fuzzy Sets*), a lógica difusa tem se mostrado mais adequada para tratar imperfeições da informação do que a teoria das probabilidades.

A Lógica Difusa, ou Lógica Nebulosa, também pode ser definida como a lógica que suporta os modos de raciocínio que são aproximados, ao invés de exatos - como estamos naturalmente acostumados a trabalhar. Está baseada na teoria dos conjuntos nebulosos, introduzido por Lofti A. Zadeh [20], em 1965.

Na teoria clássica dos conjuntos, os conjuntos são ditos *crisp*, de tal forma que um dado elemento do universo em discurso (domínio) pertence ou não pertence ao referido conjunto. Na teoria dos conjuntos nebulosos existe um grau de pertinência de cada elemento a um determinado conjunto.

Este conceito pode ser melhor compreendido através do exemplo que segue. Conforme pode ser visto na figura 3.4, suponha dois conjuntos: a) conjunto dos homens que são passiveis de participar da seleção de basquete (dado por uma altura superior a 1.75m); b) conjuntos dos homens altos. Existe uma diferença fundamental entre ambos os conjuntos. Por exemplo, apresentado um homem com 1.74m, podemos afirmar sem nenhuma discussão ou dúvida se ele pode vir a se candidatar à seleção de basquete ou não. Esta questão não é tão simples quando lidamos, por exemplo, com o conjunto dos homens altos. Uma pessoa que tenha 1,75m de altura, seria considerada como pertencente a esse conjunto? E a de 1,74m ou 1.76m? Não há claramente uma fronteira bem definida que separe os elementos do conjunto dos homens altos dos elementos do conjunto dos homens não altos.

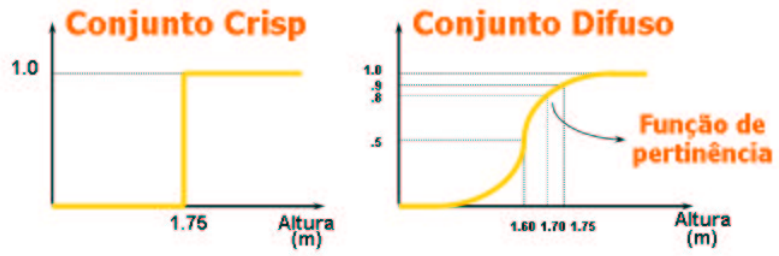

FIGURA 3.4 – Representação de um conjunto crisp e um conjunto difuso.

3.10.1 Conjunto Difuso

Um conjunto difuso A definido no universo de discurso U é caracterizado por uma função de pertinência μ_A, a qual mapeia os elementos de U para o intervalo [0, 1]. Ou seja, μ_A:U $\Rightarrow$[0, 1].

Desta forma, conforme pode ser acompanhado na figura 3.5, a função de pertinência associa com cada elemento x pertencente a U um número real $\mu_A(x)$ no intervalo [0, 1], que representa o grau de possibilidade de que o elemento x venha a pertencer ao conjunto A, isto é, o quanto é possível para o elemento x pertencer ao conjunto A.

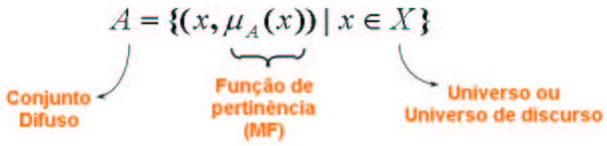

FIGURA 3.5 – Definição de um conjunto nebuloso A.

Desta forma, uma representação difusa da variável lingüística "Idade", formada pelos valores "jovem", "maduro" e "idoso" pode ser expressa como na figura 3.6.

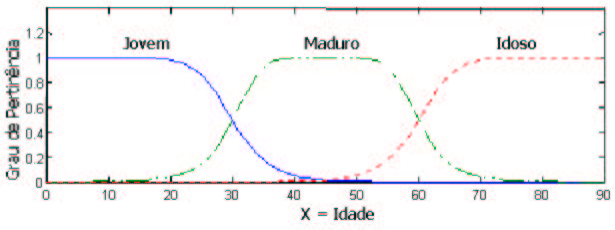

FIGURA 3.6 – Uma possível representação difusa da idade.

3.10.2 Sistemas Difusos

Sucintamente, os sistemas difusos são aqueles que têm, em seu sistema de inferência, o emprego de conjuntos difusos.

Dentre as suas vantagens, comparativamente a sistemas clássicos, podem ser citadas: a habilidade para modelar problemas extremamente complexos; o aumento da modelagem cognitiva

dos sistemas especialistas; a habilidade para modelar sistemas envolvendo vários especialistas; a redução da complexidade do modelo; uma melhora na manipulação de incerteza e possibilidades.

Para um melhor entendimento, este sistema pode ser dividido em 4 módulos básicos, conforme demonstrado na figura 3.7.

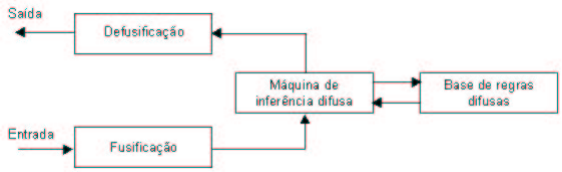

FIGURA 3.7 – Representação de um sistema difuso.

Estes módulos operam a cada ciclo com a seguinte seqüência: primeiro a medição das variáveis envolvidas no controle do processo é feita, servindo como entrada ao processo de fusificação. Os dados, já na forma difusa, são enviados à máquina de inferência difusa, que avalia a ação de controle a ser tomada, partindo da base de regras difusas. O resultado passa para a fase de defusificação remetendo um valor numérico ao processo. De forma sucinta, cada módulo pode ser assim descrito:

- Fusificação: processo que torna qualquer quantidade numérica em quantidade difusa. Isto é necessário para que a entrada do processo se torne compatível com a representação difusa adotada na base de regras.

- Base de regras: tem como objetivo representar de forma sistemática a maneira como o controlador gerenciará o sistema sob sua supervisão. Adotando valores lingüísticos iguais aos utilizados por nós quando efetuamos um controle sobre determinado processo, as regras envolvidas apresentam a forma sintática *if-then*, onde a parcela *if* é relacionada a um estado de entrada (antecedente) enquanto que a parcela *then* indica uma ação de controle (conseqüente).

- Máquina de inferência: é a responsável pela combinação do dado de entrada, já no formato de número difuso, com as regras difusas existentes, as quais, trabalhando em cima de regras de produção, descrevem o processo de tal forma que se obtenha, através de inferência, o desejado valor de saída.

- Defusificação: é a fase final do raciocínio difuso. Definida como a operação inversa da fusificação, ou seja, tem como objetivo converter cada conclusão difusa do sistema em uma variável numérica.

3.10.3 Características da Lógica Difusa

De forma resumida, a Lógica Difusa possui as seguintes características:

- está baseada em palavras e não em números, ou seja, os valores verdade são expressos lingüisticamente. Por exemplo: quente, muito frio, verdade, longe, perto, rápido, vagaroso, médio;
- possui vários modificadores de predicado como por exemplo: muito, mais ou menos, pouco, bastante, médio;
- possui também um amplo conjunto de quantificadores, como por exemplo : poucos, vários, em torno de, usualmente;
- faz uso das probabilidades lingüísticas, como por exemplo : provável, improvável, que são interpretados como números fuzzy e manipulados pela sua aritmética;
- manuseia todos os valores entre 0 e 1, tomando estes, como um limite apenas.

3.10.4 Vantagens da Lógica Difusa

Dentre as principais vantagens compiladas na literatura podem ser citadas:

- requer poucas regras, valores e decisões;
- o uso de variáveis lingüísticas nos deixa mais perto do pensamento humano;
- simplifica a solução de problemas;
- proporciona um rápido protótipo dos sistemas;
- simplifica a aquisição da base do conhecimento.

3.11 Encerramento do Capítulo

Este capítulo apresentou uma compilação dos principais conceitos encontrados na literatura, sobre RNAs e Lógica Difusa, cuja exposição se faz necessária para uma leitura satisfatória desta dissertação.

Foi apresentada uma visão geral sobre Redes Neurais Artificiais, onde foram vistos conceitos sobre neurônios, tipos de RNAs, o processo de treinamento.

Por fim, foram apresentados conceitos genéricos sobre extração de regras de RNAs e, necessário para a compreensão destas regras, uma introdução sobre Lógica Difusa.

No capítulo seguinte será feita uma análise sobre alguns dos principais trabalhos encontrados no meio científico referentes ao problema da PESP que possam, de forma direta ou indireta, auxiliar este estudo.

Capítulo 4

Revisão Bibliográfica

Neste capítulo serão apresentados os principais trabalhos na área de PESP. Inicialmente, serão vistas algumas das várias técnicas aplicadas. Após, uma análise específica dos principais trabalhos que utilizam RNAs.

4.1 Técnicas Computacionais Aplicadas ao Problema da PESP

Nos paragráfos que seguem, serão aparesentados alguns dos principais trabalhos relativos ao problema da PESP desenvolvidos com técnicas computacionais que não a de RNA.

Utilizando uma abordagem estatística, os trabalhos de Chou & Fasman [21, 22] são os primeiros a apresentarem resultados animadores no problema da PESP com base na seqüência de resíduos. A partir da análise de 15 seqüências (2.473 resíduos) foram tabuladas as freqüências relativas de cada uma das 3 classes para cada tipo de resíduo [21]. A partir destes parâmetros, foram formuladas uma série de regras para a predição da estrutura secundária [22]. O resultado obtido foi de 70 a 80% de exatidão. Mais tarde, Chou & Fasman [23] extenderam a sua análise incluido mais 29 seqüências (2.268 resíduos) e reportaram mudanças significativas nos percentuais da Metionina e mudanças menos significativas em outros resíduos levando a necessidade de aprofundamento nas pesquisas.

Outro método utilizado no problema da PESP é o *Nearest Neighbor*. Trata-se de um método onde os dados de teste são classificados de acordo com a sua proximidade aos dados de treinamento. Para tanto, é necessária uma base de dados já conhecida no processo de treinamento. No contexto da PESP, os dados de teste podem ser janelas de n resíduos consecutivos e a classificação (normalmente em 3 classes) é feita sobre o resíduo central da janela, como apresentado no trabalho de Yi & Lander [24]. Como exemplos, podem ser citados os trabalhos de

Nishikawa & Ooi [25], Levin *et al* [26], Zhang *et al.* [27], Salzberg & Cost [28], Salamov & Solovyev [29, 30] e Levin [31]. Destes, implementações funcionais dos trabalhos de Salamov & Solovyev [29, 30], aprensentando percentuais de exatidão de cerca de 72%, podem ser encontradas em um serviço na *web* [32].

Redes Bayesianas também foram exploradas no problema da PESP. Trata-se de uma forma de descrição concisa de distribuições de probabilidade conjunta. Podem ser consideradas como diagramas que organizam o conhecimento através de um mapeamento entre causas e efeitos. Como exemplos de trabalhos na área podem ser citados Gibrat *et al.* [33], Stolorz *et al.* [34] e Raymer *et al.* [35].

Outro método utilizado no problema da PESP é o *Hidden Markov Model* (HMM). Trata-se de um autômato finito com uma probabilidade de emissão de símbolo em cada estado e com probabilidade de transição entre estados. No contexto da PESP, seqüência de resíduos é o símbolo e a estrutura secundária a seqüência (oculta) de transições de estados. Como exemplos, podem ser citados os trabalhos de Asai *et al.* [36], Stultz *et al.* [37], Di Francesco *et al* [38, 39], Hargbo & Elofsson [40] e Lin *et al.* [41].

Por último, um método mais recente que está sendo aplicado ao problema da PESP é o *Support Vector Machine* (SVM). Trata-se de um método muito eficaz para problemas de classificação e regressão linear. Em seu trabalho, Hua & Sun [42] introduzem o uso de SVMs na previsão em 3 classes. O percentual de exatidão (73,5%) é comparável aos contemporâneos obtidos por RNAs. Para tanto, foi utilizada uma base com 513 seqüências não homólogas e um treinamento que utilizou o método de 7-FCV.

4.2 RNAs Aplicadas ao Problema da PESP

O problema da PESP é uma das aplicações mais antigas das RNAs em Biologia Molecular e tem sido extensamente revisitado nos últimos anos. Nos parágrafos que seguem, serão vistos alguns dos principais trabalhos desenvolvidos na área. Desde os clássicos, comumente citados na literatura, até os de vanguarda que utilizam, dentre outros recursos, RNAs recorrentes bidirecionais.

Qian & Sejnowski [43] desenvolveram um dos primeiros trabalhos significativos na utilização de RNAs aplicadas ao problema da PESP. Conforme pode ser visto na figura 4.1, foi utilizada uma rede MLP totalmente conectada e com uma única camada oculta. Para processamento dos resíduos foi utilizada a técnica de janelamento onde vários tamanhos de janela (*i.e.* 1,3,5,7,9,11,13,15,17,21) foram testados, sendo 13 o tamanho ótimo encontrado. Janelas menores

apresentaram resultados inferiores. Janelas maiores, que poderaim ter apresentado resultados superiores, mas não significativos a ponto do aumento de complexidade não justificar seu uso, também apresentaram resultados inferiores. Cada posição da janela foi representada por uma codificação ortogonal de 21 entradas sendo as 20 primeiras correspondentes aos 20 *α-aminoácidos* e a 21ª codificando o símbolo terminal da seqüência. Com relação à camada oculta, testes com vários números de neurônios (*i.e.* 0,3,5,7,10,15,20,30,40,60) determinaram 40 como sendo o de melhor percentual de exatidão. A camada de saída, dada por 3 neurônios (um para cada classe), teve a classe ganhadora dada pela de maior ativação conforme a técnica de *winner-take-all*. O percetual de exatidão Q_3 reportado foi de 63,7%, sendo $Q_\alpha = 0{,}35$, $Q_\beta = 0{,}29$ e $Q_c = 0{,}38$.

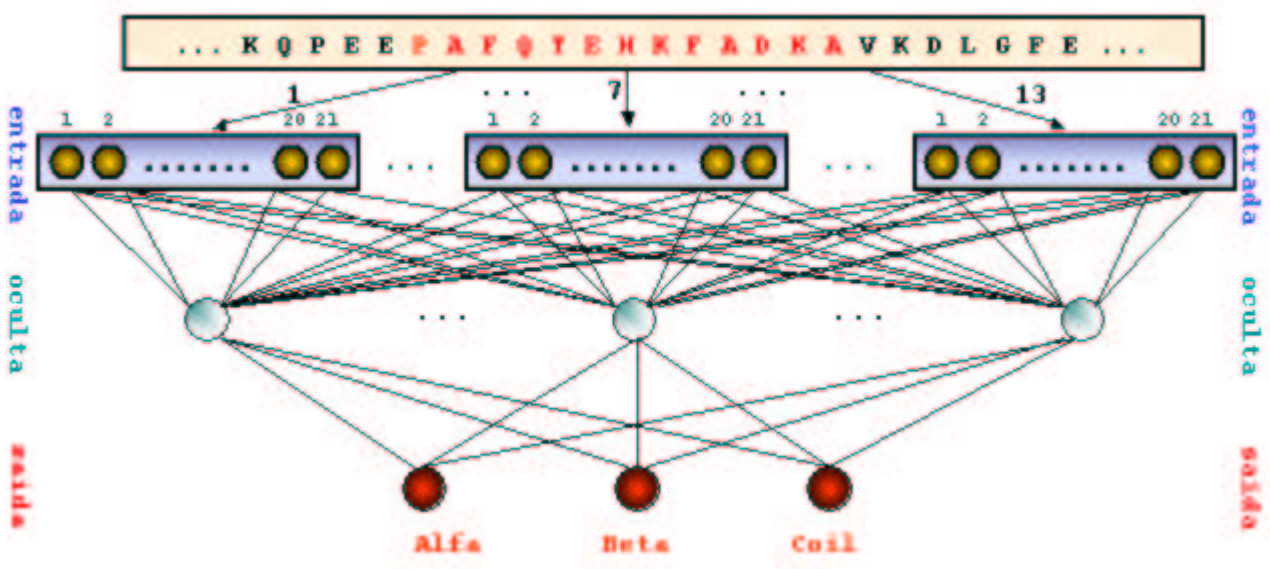

FIGURA 4.1 – RNA utilizada por Qian & Sejnowski

Maclin & Shavlik [44] apresentam um estudo onde implementaram uma RNA que incorpora, na sua arquitetura inicial, o conhecimento disponibilizado no trabalho de Chou & Fasman [23]. Isto foi feito via uma extensão ao método Knowledge-Based Artificial Network (KBANN) proposto por Towell *et al.* [45]. O KBANN traduz regras simples de um domínio em uma RNA com arquitetura cujos neurônios e pesos de ativação tentam representar este conhecimento prévio. Através do Finite-State KBANN (FSKBANN), Maclin & Shavlik implementaram o conceito de recorrência na RNA criada com a intenção de capturar as interações de curta distância entre os resíduos. Utilizando a mesma base de Qian & Sejnowski [43], obtiveram um percentual de exatidão Q_3 de 63,4%.

Rost & Sander [46], em seu trabalho, treinaram uma RNA tipo *feed-forward* sobre uma base de 130 proteínas não homólogas - sendo 126 globulares e 4 da membrana. A inovação deste estudo foi a utilização de duas camadas de RNAs. A primeira RNA processa uma janela de 13 resíduos e seu correspondente vetor de atributos bem como distâncias do resíduo ao final da proteína. Como saída, e entrada à segunda RNA, são gerados os percentuais de classificação das classes *alfa*, *beta* e *outros*. A segunda RNA, por sua vez, processa uma janela de 17 resultados percentuais e apresenta a estrutura secundária resultante através da técnica de *winner-take-all*. Os autores salientam que a inclusão da informação da família a qual pertence o aminoácido aumentou de 6 a 8 pontos o percentual de exatidão. O resultado foi um percentual de exatidão de 70,.8% para proteínas globulares e 70,2% com a inclusão de 4 proteínas da membrana. O treinamento utilizou a técnica de 7-FCV e o teste foi feito sobre um novo conjunto de 26 proteínas não homólogas. Deste trabalho resultou uma base de dados comumente utilizada pela comunidade científica (e.g. [47–49]) para testes de exatidão das soluções propostas. A base RS126 possui as 126 proteínas globulares não homólogas anteriormente citadas.

Riis & Krogh [50, 51] utilizaram uma abordagem composta por quatro componentes principais. Primeiro, foi feita uma redução no número de parâmetros, comumente utilizados nos trabalhos até então, através de uma abordagem hierárquica. Na primeira camada da rede, uma codificação ortogonal de 20 unidades, em uma janela de tamanho W, foi processada e o resultado, então, apresentado a outra camada para um novo processamento. Este método é chamado de *weight sharing* na literatura de RNAs. Segundo, foi utilizado uma arquitetura de rede distinta para cada classe. Com isto, particularidades dos dados pertinentes à cada classe puderam ser melhor trabalhados por cada rede. Terceiro, conjuntos (*ensembles*) de redes foram utilizados com a intenção de melhoria na predição. Cinco diferentes redes foram utilizadas, para cada tipo de estrutura secundária (*i.e.* classe) em cada posição da seqüência. E, por fim, uma rede final que combina todos os conjuntos foi utilizada para a determinação da classe resultante. O estudo apresentou a sua validade também pela visão de "dividir e conquistar" aplicada ao problema de classificação. Dividir em RNAs especializadas em cada classe e conquistar em uma RNA agregadora. O percetual de exatidão Q_3 reportado foi de 71,3%, sendo $Q_\alpha = 0{,}59$, $Q_\beta = 0{,}50$ e $Q_c = 0{,}41$.

No trabalho de Baldi *et al.* [47] é introduzida uma arquitetura que extende as RNAs através de uma recorrência bi-direcional que tenta capturar as informações "passadas" e "futuras" da seqüência de resíduos. Esta implementação pode ser vista na figura 4.2 (extraído de Baldi *et al.*[47]) onde, por simplicidade, todas as RNAs apresentam uma única camada oculta. Trata-se de uma nova abordagem à implementação tradicional da janela de resíduos, já vista em trabalhos

anteriores. Porém, com o diferencial importante de tentar capturar interações de longa distância pois, conforme Rost [52], cerca de 35% da formação da estrutura secundária depende das interações de longa distância. Este novo algoritmo conseguiu, já de início, percentuais de exatidão (76%) iguais aos vigentes disponibilizados por outras arquiteturas de RNAs.

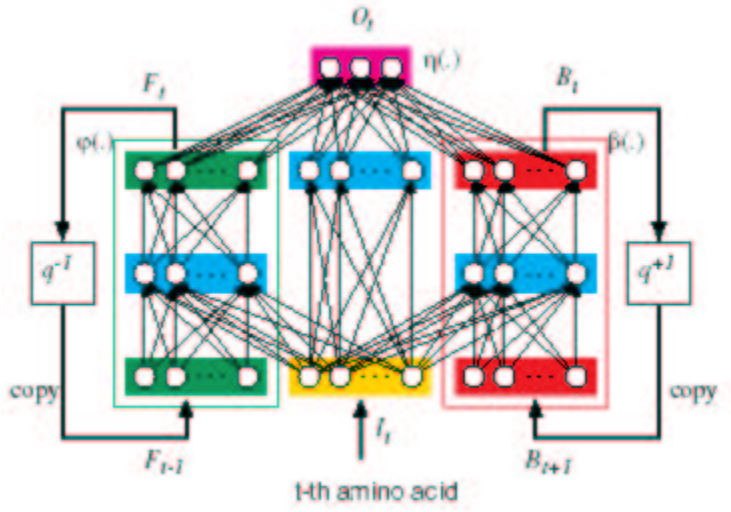

FIGURA 4.2 – RNA recorrente bi-direcional

Zhu *et al.* [53] apresentaram uma RNA Multi-Modal (RNAMM). A idéia central da RNAMM é a predição através de múltiplas RNAs, em paralelo. O resultado final é determinado pela soma proporcional das saídas de cada uma das RNAs. Sendo que 3 são as saídas de cada RNA, uma para cada estrutura (*alfa, beta* e outros). A que obtiver maior valor será a vencedora. O resultado obtido, de 66%, foi comparado a uma RNA simples, com a mesma codificação de entrada/saída, e mostrou-se 6.9% superior.

Lamont et al [54] ressaltaram a importância de identificar a correta representação dos atributos a serem submetidos à RNA. Em seu estudo, investiga o efeito da tradicional codificação binária dos aminoácidos frente a uma nova representação que leva em consideração os ácidos nucleicos originais de cada códon (*i.e.* A, T, C e G). Para isto, dois conjuntos de experimentos foram realizando, sendo que em ambos foram realizadas variações no número de neurônios da camada oculta. No primeiro, a tradicional codificação ortogonal de 20 entradas binárias foi comparada contra uma codificação onde os ácidos nucleicos receberam os valores de: A = 1000, C = 0100, G = 0010 e T = 0001. Combinações de dois ou mais ácidos nucleicos foram implementadas através do operador OU (*e.g.* A ou T = 1001). Estes experimentos demonstraram resultados

estatisticamente semelhantes. Para a codificação ortogonal de 20 entradas os percentuais de exatidão reportados foram de: $Q_3 = 62{,}24\%$, sendo $Q_\alpha = 0{,}65$, $Q_\beta = 0{,}34$ e $Q_c = 0{,}73$. Para a codificação do códon, os percentuais de exatidão reportados foram de: $Q_3 = 61{,}35\%$, sendo $Q_\alpha = 0{,}62$, $Q_\beta = 0{,}40$ e $Q_c = 0{,}70$. No segundo experimento, a codificação ortogonal foi comparada contra uma codificação ortogonal de 10 entradas que representaram os ácidos nucleicos. Os percentuais de exatidão reportados foram de: $Q_3 = 61{,}18\%$, sendo $Q_\alpha = 0{,}61$, $Q_\beta = 0{,}38$ e $Q_c = 0{,}74$. Para a codificação do códon, os percentuais de exatidão reportados foram de: $Q_3 = 61{,}68\%$, sendo $Q_\alpha = 0{,}64$, $Q_\beta = 0{,}37$ e $Q_c = 0{,}72$. Como base de treinamento, foi utilizada a disponibilizada por Cuff & Barton [48].

Pollastri *et al.* [49], dando continuidade ao trabalho iniciado por Baldi *et al.* [47], apresentam uma solução para o problema da PESP baseada em um conjunto de RNAs recorrentes bi-direcionais. Como base de treinamento, foram utilizados os mesmos dados submetidos ao servidor de previsão de estruturas secundárias SSPro v1 [55] . Para testes, duas bases foram utilizadas, a RS126 [46] e um conjunto de 223 proteínas não homólogas (com um total de 47.370 resíduos) obtidas a partir do projeto EVA [56]. Como resultado, foram liberados o SSPro v2 [55] e o SSpro8 v1 [55] (que classifica os resíduos em 8 classes, ao invés das 3 tradicionais). O percentual de exatidão obtido no SSPro v2 foi de 78%.

4.3 Encerramento do Capítulo

Neste capítulo foram apresentados alguns dos principais trabalhos na área de PESP.

Analisando-se os métodos vistos no item 4.1 com os trabalhos que utilizam RNAs, reportados no item 4.2, uma questão surge naturalmente: Porque utilizar RNAs e não métodos estatísticos tradicionais? Segundo Wu & McLarty [9], esta não é a questão correta. A questão deveria ser: Como ambas as escolas podem trabalhar em conjunto visando obter melhores resultados na solução dos difíceis problemas desta área? Conforme será visto no próximo capítulo, este estudo, através do processo de validação das regras extraídas das RNAs, tem a intenção de demonstrar uma das possibilidades de como isto pode ser feito.

Dos trabalhos analisados que empregam RNAs ao problema da PESP, algumas observações podem ser feitas. Primeira, quer pelo processo de janelamento, quer pelo de redes recorrentes, as informações do contexto no qual o resíduo se encontra parecem auxiliar a identificação da estrutura secundária. Segunda, a preocupação com a utilização de bases que não apresentem seqüências homólogas é tida como muito importante. Terceira, todos os estudos giram em torno do item "percentual de exatidão". Exceto pelo trabalho pioneiro de Chou & Fasmam [21, 22] (com

abordagem estatística), não foi encontrado trabalho algum com a preocupação de entender o que está sendo aprendido pelas RNAs no que diz respeito ao problema da PESP.

No capítulo seguinte será apresentada a metodologia proposta para a extração de regras de RNAs aplicadas ao problema da PESP. Serão cobertas as etapas tradicionais de preparação dos dados, definição da arquitetura da rede, treinamento. Adicionalmente, serão apresentados os processos de extração de regras e a validação estatística das regras extraídas.

Capítulo 5

Extração de Regras de RNAs Aplicadas ao Problema da PESP

Neste capítulo serão apresentados os principais pontos considerados na abordagem sobre como executar a extração de regras de RNAs aplicadas ao problema da PESP.

Inicialmente serão expostos a motivação, os objetivos a serem atingidos e a contribuição pessoal dada. A seguir, será visto o sistema de extração de regras que será utilizado neste estudo. Logo em seguida, será apresentada a implementação realizada para a validação das regras obtidas. No item seguinte, uma descrição, de forma genérica, da metodologia empregada no desenvolvimento de estudos que envolvam a extração de regras de RNAs. Por fim, será apresentada a aplicação desta metodologia no problema da PESP.

5.1 Motivação e Objetivos

Através de uma análise na produção científica disponível sobre o uso de RNAs aplicadas ao problema da PESP observam-se dois aspectos interessantes. Primeiro, que a meta dos pesquisadores é por melhores percentuais de exatidão na predição realizada. Segundo, que o conhecimento, aprendido pelas RNAs, não tem sido explorado, mesmo que recursos já existam para tal.

Alguns questionamentos já deveriam ter sido feitos pelos pesquisadores de forma autocrítica. Por exemplo: Será que, mesmo obtendo resultados percentuais distintos, as RNAs não assimilaram informações semelhantes? Será que, mesmo RNAs distintas em arquitetura ou método de aprendizado, mas que utilizam dados semelhantes no seu treinamento, não estão

assimilando o mesmo conhecimento? E, mais importante, será que o conhecimento assimilado pelas RNAs já não poderia ter sido utilizado de forma satisfatória pelos biólogos moleculares em benefício da ciência?

O rótulo de "caixa-preta", atribuído às RNAs vem sendo, ao longo dos últimos anos, derrubado através da disponibilização de métodos de extração do conhecimento assimilado pelas mesmas. Conforme Cechin [5], alguns exemplos específicos de extração de regras difusas de RNAs são: FNES (Fuzzy Neural Expert System), proposto por Hayashi & Nakai [57], que executa uma extração automática de regras difusas a partir de uma RNA treinada com o algoritmo *Backpropagation*; NNDFR (Neural Network Driven Fuzzy Reasoning), desenvolvido por Takagi & Hayashi [58, 59], que também extrai regras difusas a partir de uma RNA treinada com o algoritmo *Backpropagation*; Poechmueller *et al.* [60] trataram o problema da extração de regras difusas a partir de redes RBFs (Radial Basis Function). Mesmo redes mais complexas, como as recorrentes, já possuem técnicas de extração de conhecimento.

Dentro deste contexto, o tema de pesquisa proposto neste estudo trata o problema da PESP como uma oportunidade para a disponibilização do conhecimento aquirido pelas RNAs. Na busca por este conhecimento, será utilizado o software de extração de regras de RNAs denominado *Fuzzy Automatically Generated Neural Inferred System* (FAGNIS) [5]. Sua escolha se deu, principalmente, devido a sua característica de trabalhar com regras do tipo Takagi-Sugeno [6] pois o conjunto de regras gerado por esta técnica é mais conciso do que os de outras técnicas. Com uma redução no número de regras, atendo-se as mais significativas, o processo torna-se mais produtivo.

Uma vez obtido o conhecimento, na forma de regras, o mesmo necessita ser validado. A validação em laboratório é inviável no que diz respeito ao ponto de vista econômico. Adicionalmente, não haveria tempo hábil para os experimentos serem desenvolvidos. Dentro deste contexto, a validação estatística apresentou-se como uma alternativa adequada e cientificamente válida.

Este estudo propõe uma métrica para a validação da solução encontrada pela RNA baseada no número de parâmetros realmente utilizados pela rede. Ou seja, no número de regiões lineares e no número de parâmetros, de cada região linear, onde realmente existam padrões de treinamento. Embora a RNA seja uma solução tipicamente não linear, a mesma pode ser vista como uma solução linear a partir do momento que passamos a trabalhar com o resultado gerado pelo FAGNIS. Isto é, com as equações lineares que descrevem o comportamento da RNA e às quais estão associadas funções de pertinência que caracterizam qual a equação a ser utilizada para cada padrão processado.

Uma vez obtidas as regras estatisticamente válidas, as mesmas serão disponibilizadas como contribuição pessoal à comunidade científica com o intuito de fomentar novas pesquisas e de serem comprovadas, ou refutadas, em testes de laboratório.

Cabe salientar que este estudo, diferente dos demais encontrados na literatura, não visa atingir altos graus de exatidão na predição da estrutura secundária. O objetivo maior é a descoberta de conhecimento e não participar da corrida na busca por melhores resultados percentuais.

Outro ponto que merece ser comentado é o de que estudo tenta descobrir o conhecimento em interações de curta distância (*i.e.* uma janela de 13 resíduos), conforme será visto nos itens que seguem. Embora seja de conhecimento comum que o processo de dobramento de proteínas provavelmente sofra a influência de interações de longa distância, esta foi uma escolha baseada no histórico de trabalhos anteriores encontrados na literatura de RNAs aplicadas ao problema da PESP.

De forma resumida, os objetivos deste estudo podem ser enumerados em:

- dominar o processo de extração de conhecimento de RNAs;
- implementar um processo de validação estatística das regras obtidas;
- disponibilizar o conhecimento obtido, na forma de regras estatisticamente válidas, à comunidade científica.

5.2 O Sistema de Extração de Regras

O sistema de extração de regras escolhido para ser aplicado no problema da PESP foi o FAGNIS, proposto por Cechin [5].

Dentre os principais motivos para a escolha do FAGNIS, podem ser citados:

- permite a extração de regras diretamente da RNA treinada;
- atende ao tipo de rede utilizado (*i.e. MLP feedforward*);
- atende parcialmente a tarefa de classificação (*i.e.* determinação da estrutura secundária);
- garante uma equivalência de funções entre a RNA e o sistema de inferência difuso extraído;
- apresenta uma alta compreensibilidade das regras extraídas (*i.e.* regras mais simples e em menor quantidade);

- dispõe-se do seu código-fonte o que possibilita a implementação dos procedimentos de validação estatística das regras obtidas (a ser exposto em detalhes no item 5.3).

Além disso, o sistema resulta em um conjunto de regras do tipo Takagi-Sugeno [6], que apresentam a característica de descrever o comportamento da RNA em uma simples função linear e permitir a extração da influência das entradas nas saídas.

De forma simplifica, o processamento executado pelo FAGNIS pode ser descrito como ato de particionar a região de trabalho da RNA em regiões lineares (uma clusterização do espaço de entrada). Isto é feito de forma semelhante a regressão linear por partes, à medida que a rede vai sendo treinada. Cada região linear encontrada pelo FAGNIS corresponde a uma regra, para a qual são reportadas:

- o protótipo: que descreve o comportamento típico dos padrões pertencentes à regra;
- a equação linear: que determina a saída calculada para os padrões correpondentes à regra a partir de uma análise dos pesos e ativações da rede;
- a função de pertinência: que indica o grau de pertinência do padrão para com a regra.

Apesar das vantagens citadas, o FAGNIS apresenta a característica de consumir recursos de forma excessiva, especialmente no que diz respeito a memória. As estruturas para mapeamento do espaço neural e extração das regras exigem um volume considerável. Para as simulações feitas neste estudo, foi necessário uma máquina com 1Gb de memória principal.

5.2.1 Regras do Tipo Takagi-Sugeno

Takagi & Sugeno [6] apresentaram uma ferramenta matemática para construir um modelo difuso utilizado para modelar um sistema em que a premissa é a descrição do subespaço de entradas e a sua conseqüência é uma relação linear entre a entrada e a saída. Isto foi possível através de um sistema de raciocínio difuso multidimensional onde o número de implicações difusas foi drasticamente reduzido. Como conseqüência, com esta melhoria, o raciocínio foi simplificado.

A implicação difusa, apresentada por Takagi & Sugeno [6], está baseada na partição difusa do espaço de entrada. Em cada subespaço, uma relação linear de entrada e saída é formada. A saída de um raciocínio difuso é dada pela agregação dos valores inferidos por implicações que foram aplicadas para uma entrada.

O sistema de inferência difuso Takagi-Sugeno, com variáveis de entrada x_1, ... , x_n e a variável de saída y, contém um sistema de regras com o seguinte formato:

$$R\text{: } If\ \ f\,(x_1\ is\ F_1,\ ...,\ x_n\ is\ F_n)\ \ Then\ y = g\,(x_1\ ...,\ x_n) \tag{5.1}$$

onde:

y é a variável da conseqüência cujo valor será inferido;

x_1, ... , x_n são as variáveis da premissa que aparecem também na conseqüência;

F_1, ... , F_n são os conjuntos difusos, com suas funções de pertinência lineares, representando o subespaço difuso no qual a implicação R pode ser aplicada para raciocínio;

f é a função lógica que conecta as proposições na premissa;

g é a função que implica o valor de y quando x_1, ... , x_n satisfaz a premissa.

Para cada conjunto difuso F há uma relação única com uma função de pertinência. Motivo pelo qual, algumas vezes, os dois termos são empregados para o mesmo símbolo F.

Embora a função g seja genericamente definida, Takagi & Sugeno reportaram apenas o uso de uma função linear $g(x_1, ..., x_n) = p_0 + p_1x_1 + ... + p_ix_i + ... + p_nx_n$ onde p são constantes. A inferência (função f) é implementada com a operação de multiplicação (*i.e.* prod) e a composição com a soma (*i.e.* sum), portanto, este algoritmo pode ser chamado de regra *sum-prod.*

5.2.2 Implementação em uma Unidade Sigmóide

Conforme Cechin [5], a implementação feita no FAGNIS, para representar um neurônio de ativação (*i.e.* uma unidade sigmóide), é dada pelas 3 regras que seguem:

$$\left.\begin{array}{lll} If\ x\,is\,F_1 & Then & y = -1 \\ If\ x\,is\,F_2 & Then & y = \frac{1}{2}x \\ If\ x\,is\,F_3 & Then & y = +1 \end{array}\right\} \tag{5.2}$$

Isto significa que a função sigmóide equivale à soma:

$$sigmoide(x) = \frac{2}{1+e^{-x}} - 1 = F_1(x)(-1) + F_2(x)(\frac{1}{2}x) + F_3(x)(+1) \tag{5.3}$$

onde a função de pertinência F_2 representa a derivada de $sigmoide(x)$:

$$sigmoide'(x) = \frac{1}{2}F_2(x) \tag{5.4}$$

enquanto que as funções de pertinência F_1 e F_3 são determinadas por:

$$\left.\begin{array}{l} F_1(x) = If\ \ x < 0\ \ Then\ \ -sigmoide(x) + sigmoide'(x)x\ \ Else\ \ 0 \\ F_3(x) = If\ \ x > 0\ \ Then\ \ sigmoide(x) - sigmoide'(x)x\ \ Else\ \ 0 \end{array}\right\} \tag{5.5}$$

Os resultados obtidos com a variação do valor de x podem ser visualizados na figura 5.1. As equações lineares que mapeiam este valores podem ser observados em 5.1(a) e as funções de pertinência correspondentes no gráfico reportado em 5.1(b).

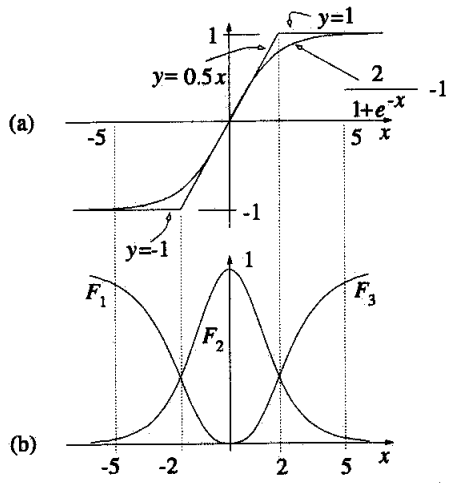

FIGURA 5.1 – Função *sigmoid* e a parte conseqüente de cada regra difusa (a); número ideal de funções de pertinência para cada região da unidade de ativação sigmóide (b).

5.3 Implementação da Validação Estatística das Regras Extraídas

Uma vez que o FAGNIS implementa o treinamento da RNA e o gradual particionamento da região de trabalho da mesma, em regiões lineares à medida que a rede é treinada, torna-se

possível medir a precisão da solução encontrada para cada regra bem com a precisão da solução global.

Partindo-se do princípio de que estamos trabalhando com uma distribuição normal nos dados, a implementação feita realiza uma regressão linear sobre os padrões de cada regra. Isto é feito via regressão linear matricial, conforme a equação 5.6, cuja demonstração pode ser vista em Gujarati [61]. Através deste procedimento, uma nova equação linear Eq_R, além da originalmente fornecida pelo FAGNIS, pode ser utilizada.

$$Eq_R : \left(X_R^T X_R\right)^{-1} X_R^T Y_R \tag{5.6}$$

Onde:
X_R é a matriz contendo os padrões pertencentes à regra R;
X_R^T é a matriz transposta de X_R;
Y_R é a matriz contendo os resultados esperados dos padrões existentes em X_R.

Apesar do resultado obtido com a equação linear original, disponibilizada pelo FAGNIS, ser muito semelhante ao da equação obtida a partir da regressão linear optou-se pela última. O motivo para esta escolha foi devido à mesma ser estatisticamente válida, ao invés de uma aproximação inicial do subespaço neural (como a original).

Isto é feito para todas as regiões onde o número de padrões atendidos por uma regra seja superior ao número de restrições independentes do modelo (*i.e.* o número de parâmetros estimados). Para regiões onde o número de padrões é inferior ao número de parâmetros estimados, as regras geradas não são utilizadas. Os padrões que foram atendidos pelas regras não utilizadas são contabilizados e considerados como uma *perda* natural que ocorre quando modelos mais complexos, como os não-lineares, são reduzidos à modelos mais simples, como os lineares.

Uma vez executada a regressão linear sobre os padrões das regras passíveis de tal procedimento, informações significativas sobre as mesmas podem ser descobertas. Além da *cobertura da regra*, implementada no processo de partição do espaço neural, as informações de *erro RMS da regra*, *inexatidão da regra* e a medida de validação do *erro global* foram implementadas a partir da equação linear e serão apresentadas nos tópicos que seguem.

Cabe salientar que estas métricas atuam sobre as equações lineares identificadas a partir do espaço neural da RNA. As métricas de erro específicas da RNA, como os erros *Root Mean Square* (RMS), *Mean Square Error* (MSE), *Sum of Square Errors* (SSE) e outras, continuam

válidas e disponíveis. Porém, sua área de atuação é o espaço não linear original onde a rede atua.

5.3.1 Cobertura da Regra

A cobertura C_R de uma regra, dada pela equação 5.7, representa o percentual de padrões que foram atendidos pela regra, no momento da parada do treinamento. Seu valor varia entre 0 e 1.

$$C_R = \frac{P_R}{T} \tag{5.7}$$

Onde :
P_R é a quantidade de padrões P atendidos pela regra R;
T é a quantidade total de padrões submetidos no treinamento da RNA.

Além da cobertura individual de cada regra, também é disponibilizada a cobertura total. A cobertura total Ct, dada pela equação 5.8, indica a cobertura global das regras extraídas, ou seja, o quanto as regras abrangem do espaço original de entrada.

$$Ct = \sum_{R=1}^{N} C_R \tag{5.8}$$

Onde :
C_R é a cobertura individual da regra R;
N é o número de regras geradas.

Apesar de ser o somatório da cobertura de cada regra R, seu valor pode ser inferior a 100% pois, conforme exposto anteriormente, somente são geradas as regras com um número de padrões superior ao número de parâmetros estimados do modelo. A equação 5.9 representa o percentual de padrões não atendidos pelas regras extraídas.

$$Cn = 100 - Ct \tag{5.9}$$

5.3.2 Erro RMS da Regra

O erro $Erms_R$ da regra, definido como o erro padrão da estimativa, representa o desvio-padrão dos valores de saída do modelo em relação à reta da regressão estimada (obtida na equação 5.6).

$$Erms_R = \sqrt{\frac{\sum_{R=1}^{N} (Y_e - Y_c)^2}{P_R - p}} \tag{5.10}$$

Onde :

N é o número de regras geradas;

Y_e é o valor esperado da saída;

Y_c é o valor calculado da saída;

P_R é a quantidade de padrões P atendidos pela regra R;

p é o número de parâmetros estimados do modelo.

5.3.3 Inexatidão da Regra

Embora o erro RMS seja um indicativo válido, ele não apresenta uma leitura significativa, em termos de compreensão para o usuário, do real significado do erro de classificação.

Suponha um problema de classificação onde as saídas esperadas sejam: 1 (caso pertença à classe) e -1 (caso não pertença). Suponha, também, uma rede que possa responder com um valor real entre -1 e 1. Onde, por convenção, se adota que valores maiores ou iguais a zero sejam considerados, lógicamente, como 1; e menores do que zero como -1. Agora suponha que, para todos os padrões submetidos, a rede tenha respondido com o valor 0.4 e a saída esperada, de todos, fosse 1. O erro RMS seria de 0.6. Mas, pela convenção lógica adotada (característica de problemas de classificação), a rede teria acertado em 100% dos casos.

Para lidar com o exemplo descrito, uma abordagem típica para problemas de classificação é a utilização, como métrica para o cálculo da performance, da *Exatidão* (conforme equação 3.5). A *Exatidão* apresenta o percentual de acerto da solução proposta. Como complemento, ou seja, como uma medida de erro, este estudo implementou a Inexatidão I_R de uma regra (conforme equação 5.11).

$$I_R = 1 - Exatidao_R \tag{5.11}$$

5.3.4 Erro Global

Dada a topologia de uma RNA não-linear, deve-se garantir que a RNA não seja um modelo tendencioso dos dados aprendidos. Os dois métodos mais utilizados para resolver este tipo de problema são o *weight-decay* e *early-stopping*. O primeiro reduz gradualmente os pesos (parâmetros) da RNA de forma uniforme, enquanto o segundo procura parar o treinamento (ajuste) em um ponto ótimo. Tanto o decréscimo dos pesos, quanto a parada, exigem dados de validação também chamados dados "não vistos" ou, simplesmente, dados não apresentados à RNA.

Outras métricas baseadas no número de parâmetros do modelo derivam de métodos estatísticos para o erro quadrático médio dos modelos de regressão. Nestes, divide-se o erro pelo número de padrões menos o número de parâmetros do modelo. RNAs, porém são modelos não-lineares onde os pesos do modelo interagem, resultando em menos graus de liberdade do que o número total de parâmetros do modelo.

Suponha uma RNA como apresentada na figura 5.2. Se todos os 40 neurônios ocultos, implementando cada um a função não-linear sigmóide, trabalharem em sua região central linear, a RNA será equivalente a um modelo linear, ou seja, o número de pesos será muito maior do que o número real de graus de liberdade do modelo, portanto, medidas de erro de generalização utilizando o número de parâmetros do modelo serão erroneamente geradas.

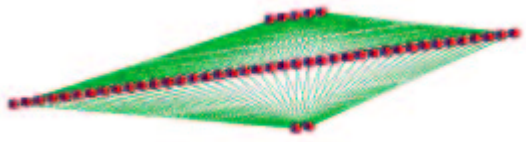

FIGURA 5.2 – RNA com 5 neurônios na camada de entrada, 40 neurônios não-lineares na camada oculta e 2 neurônios na camada de saída

É interessante notar que, neste contexto, o método *weight-decay* tenta diminuir os pesos da RNA o que significa reduzir a região de trabalho dos neurônios na camada oculta à região central linear. Isto aumenta a capacidade de generalização da RNA pela redução do número de graus de liberdade da mesma, já que a aproxima de um modelo linear equivalente. O caso extremo seriam RNAs sem neurônios na camada oculta, que são equivalentes a modelos lineares generalizados.

Assim, para esta subclasse pode-se usar métodos clássicos para o erro de generalização através do conceito de graus de liberdade do modelo.

Este estudo propõe uma nova métrica (*i.e.* erro global) baseada, não no número de parâmetros (pesos) do modelo, mas no número de parâmetros realmente utilizados pela RNA, ou seja, no número de regiões lineares e no número de parâmetros de cada região linear realmente utilizada pela RNA (onde existam padrões de treinamento). Com isto, torna-se possível determinar o ponto aproximado de parada do treinamento sem a necessidade de dados de validação.

O *erro global* apresenta uma visão do que vai acontecendo com a solução global a medida que o treinamento vai sendo executado. Duas formas de cálculo do *erro global* foram implementadas. A primeira, conforme pode ser visto na equação 5.12, utiliza o erro *RMS* de cada regra como base de cálculo. Pode ser utilizada tanto para problemas de aproximação de função quanto de classificação. A segunda, conforme a equação 5.13, utiliza a *Inexatidão* da regra. Desta forma, para problemas de classificação, uma leitura mais próxima da realidade pode ser obtida. Este segundo erro é disponibilizado somente quando o problema for de classificação.

$$Egr = \frac{\sum_{R=1}^{N} (C_R \, Erms_R)}{T} \tag{5.12}$$

Onde :
N é o número de regras geradas;
C_R é a cobertura individual da regra R;
$Erms_R$ é o erro RMS da regra R;
T é a quantidade total de padrões submetidos no treinamento da RNA.

$$Egi = \frac{\sum_{R=1}^{N} (C_R \, I_R)}{T} \tag{5.13}$$

Onde :
N é o número de regras geradas;
C_R é a cobertura individual da regra R;
I_R é a Inexatidão regra R;

T é a quantidade total de padrões submetidos no treinamento da RNA.

5.4 Metodologia para Extração e Validação das Regras

Conforme já visto no capítulo 3, uma série de decisões devem ser tomadas visando uma utilização adequada dos recursos disponibilizados pelas RNAs. A metodologia aqui definida tem por objetivo apresentar os principais passos a serem observados no processo extração de regras de RNAs que envolvam o sistema de extração utilizado neste estudo.

Como pode ser acompanhado no fluxo apresentado na figura 5.3, as fases que devem ser observadas incluem: preparação da base de dados; definição da arquitetura da RNA; definição do tipo do algoritmo de aprendizado; treinamento e extração de regras estatisticamente válidas; e, por fim, uma eventual validação experimental das regras extraídas.

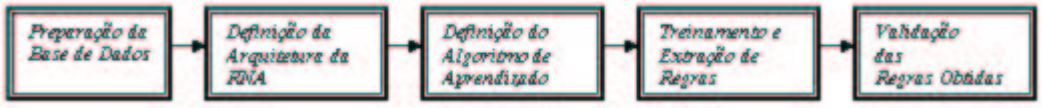

FIGURA 5.3 – Fluxo da metodologia para extração e validação de regras

Na fase de *preparação da base de dados*, devem ser definidos os dados que serão utilizados como entrada à RNA e que tipo de resultado se deseja obter. Uma vez definidos, os dados devem ser obtidos e disponibilizados em um formato adequado ao processamento da RNA. A obtenção dos dados pode envolver várias fontes distintas. Nem sempre todos os dados necessários estão disponíveis em um mesmo repositório. Adicionalmente, raras são as situações em que os dados obtidos apresentam-se em um formato adequado ao processamento das RNAs. Um processamento adicional de formatação dos dados pode ser necessário e envolver normalizações, discretizações e outras operações pertinentes à característica da técnica aplicada.

A fase de *definição da arquitetura da RNA* envolve a topologia da rede, isto é, o número de neurônios das camadas de entrada, oculta e de saída bem como os tipos de ligações entre as mesmas.

Conforme explanado no item 3.5, a fase de *definição do algoritmo de aprendizado* está relacionada com o tipo de problema e com a forma como os dados se apresentam.

As fases de *treinamento* e *extração de regras* são repetidas até que o número de regras extraídas seja o desejado ou que um outro critério de parada qualquer seja estabelecido (*e.g.*

tempo, número de épocas, erro global).

Por fim, a fase de *validação das regras obtidas* pode ser eventualmente realizada se houverem condições para tal. A comprovação experimental se dá, como o próprio nome indica, através da comprovação por experimentos reais que aceitem, ou refutem, as afirmações feitas pelas regras obtidas. Sua viabilidade, normalmente, está ligada a questões de custo financeiro, técnico, fatores humanos e disponibilidade de tempo.

5.5 Aplicação da Metodologia ao Problema da PESP

Nos itens que seguem serão apresentados os procedimentos adotados para abordar o problema da PESP. Estes procedimentos foram implementados seguindo-se a metodologia descrita na seção anterior.

5.5.1 Preparação da Base de Dados

No que diz respeito aos dados a serem utilizados no processo de aprendizado de RNAs, Cuff & Barton [48], bem como Rost [62], discutem os efeitos das seqüências homólogas e não homólogas no desenvolvimento e avaliação de novos algoritmos de previsão. Visando minimizar os efeitos de uma escolha inadequada, optou-se pela utilização de uma base de dados já validada e consolidada.

A análise de vários trabalhos na área levou à identificação de 4 bases de dados comumente utilizadas pela comunidade científica: QS106 [43], RS126 [46], CB396 [48] e a base do projeto EVA [56] da universidade de Columbia. A QS106, uma das primeiras bases de teste que se tem notícia nesta área, contém um conjunto de 106 proteínas globulares não-homólogas com estrutura conhecida. A RS126 contém um conjunto de 126 seqüências de proteínas globulares não homólogas. A CB396 apresenta um conjunto de 396 seqüências não homológas e distintas das reportadas em RS126. A base do projeto EVA, na data do acesso feito ao site, apresentava um conjunto de 2.980 seqüências não homólogas cadastradas.

Devido a qualidade extremamente apurada da análise feita por Qian & Sejnowski [43], optou-se pela utilização da base QS106. Com isto, os resultados obtidos poderão ser mais facilmente comparados e analisados.

A partir da relação das 106 proteínas [43], os arquivos correspondentes foram obtidos no site Protein Data Back (PDB) [63]. Destas 106, conforme pode ser visto no Apêndice A, duas não se encontravam mais presentes e 29 tiveram seus códigos alterados desde a realização do estudo original de Qian & Sejnowski em 1988.

Das 104 proteínas efetivamente recuperadas do PDB, geraram-se 123 cadeias visto que, para algumas proteínas, mais de uma cadeia (*chainid*) foi utilizada no estudo original. Destas 123 cadeias, 20 foram descartadas por apresentarem códigos de aminoácidos diferente dos 20 α-aminoácidos. O resultado final foi o aproveitamento de 91 proteínas, com 103 cadeias e 17.857 resíduos.

Definição dos Atributos

Conforme já exposto anteriormente, a intenção deste trabalho é dominar o processo de extração de conhecimento e trazer à tona informações sobre como os atributos físico-químicos dos resíduos influenciam na formação da estrutura secundária das proteínas. Para tanto, uma abordagem reducionista foi utilizada. Minimizando, desta forma, a complexidade computacional do processo de extração de regras.

Para cada um dos resíduos pertencentes às cadeias obtidas, um conjunto preliminar de atributos foi escolhido para ser o alvo de estudo sobre a sua influência na determinação da estrutura secundária. Destes, visando a utilização de atributos que apresentem um grau de associação linear baixo (conforme tabela 5.1), além da redução da complexidade computacional e o envolvimento em uma posterior análise, optou-se por trabalhar com o índice hidropático e o ponto isoelétrico de cada resíduo. Respectivamente encontrados na literatura como *HP* (*Hidropathic Index*) e *pI* (*Isoeletric point*). Seus valores podem ser vistos na tabela 2.2. Adicionalmente, a escolha destes dois atributos também se deu devido às referências encontradas na literatura especializada que apresentam um grau maior de importância da hidrofobicidade e da carga dos resíduos no processo de dobramento da proteína.

O índice hidropático, introduzido por Kyte & Doolittle [10], apresenta uma combinação entre os valores de hidrofobicidade e hidrofilicidade das cadeias laterais dos aminoácidos. Pode ser utilizado como uma tendência do aminoácido de busca por um ambiente aquoso (valores menores) ou fuga dos mesmos (valores maiores).

O ponto isoelétrico, conforme Lehninger [7], é definido como o pH no qual a carga do aminoácido é zero. Calcula-se em função do pH dos grupos carboxil, amino e da cadeia lateral (dependendo do tipo de aminoácido).

TABELA 5.1 – Grau de correlação Pearson entre os atributos preliminares escolhidos

	Ponto Isoelétrico	Índice Hidropático	Volume	Massa	Superfície
Ponto Isoelétrico	1	-.205	.363	.196	-.181
Índice Hidropático	-.205	1	-.080	-.271	.841
Volume	.363	-.080	1	.934	.182
Massa	.196	-.271	.934	1	.109
Superfície	-.181	.841	.182	.109	1

Formatação dos Atributos

Após selecionados, os atributos passaram por uma fase de formatação visando serem adequados às caracteristicas do processo de treinamento das RNAs e facilitar a análise das regras extraídas. Neste processo, os atributos índice hidropático e ponto isoelétrico foram normalizados, ficando com seus valores como reportados na tabela 5.2. Por "normalização" entenda-se o processo de transformar os dados para terem média 0 e desvio-padrão 1.

TABELA 5.2 – Atributos físico-químicos utilizados

Aminoácido	HP Original	HP Normalizado	pI Original	pI Normalizado
Alanina	1,8	0,766663	6.01	-0,021484
Arginina	-4,5	-1,342497	10,76	2,664004
Asparagina	-3,5	-1.00771	5,41	-0,360703
Aspartato	-3,5	-1.00771	2,77	-1,85327
Cisteina	2,5	1.001014	5,07	-0,552928
Glutamina	-3,5	-1.00771	5,65	-0,225016
Glutamato	-3,5	-1.00771	3,22	-1,598855
Glicina	-0,4	0,030131	5,97	-0,044099
Histidina	-3,2	-0,907273	7,59	0,871794
Isoleucina	4,5	1,670588	6,02	-0,01583
Leucina	3,8	1,436237	5,98	-0,038445
Lisina	-3,9	-1,141624	9,74	2,087331
Metionina	1,9	0,800141	5,74	-0,174133
Fenilalanina	2,8	1,10145	5,48	-0,321128
Prolina	-1,6	-0,371614	6,48	0,244238
Serina	-0,8	-0,103784	5,68	-0,208055
Treonina	-0,7	-0,070305	5,87	-0,100635
Triptofano	-0,9	-,137263	5,89	-0,089328
Tirosina	-1,3	-0,271178	5,66	-0,219362
Valina	4,2	1,570152	5,97	-0,044099

Embora este estudo se proponha a calcular as estruturas secundárias, elas necessitam ser fornecidas à RNA na etapa de treinamento. Isto deve ser feito, exatamente, para que a mesma aprenda a determiná-las. Para a geração da saída esperada, ou seja, do valor da estrutura secundária de cada resíduo, foi utilizado o software *Define Secondary Structure of Proteins* (DSSP) [64], criado por Kabsch & Sander [65].

O DSSP determina as estruturas secundárias a partir das coordenadas dos átomos e das pontes de hidrogênio que se formam entre os mesmos. As coordenadas atômicas são obtidas através do processamento do arquivo PDB correspondente à proteína.

Para cada resíduo, é fornecida uma saída como na tabela 5.3. Como este estudo trabalha com a determinação de 3 classes, sem detalhamento de suas subclasses, as "especializações" nas classes foram convertidas sendo que os valores utilizados são os reportados na coluna *Classe utilizada*.

TABELA 5.3 – Classes reportadas pelo DSSP

Classe DSSP	Descrição DSSP	Classe utilizada
B	*residue in isolated beta-bridge*	Beta
E	*extended strand, participates in beta ladder*	Beta
G	*3-helix (3/10 helix)*	Alfa
H	*alpha helix*	Alfa
I	*5 helix (pi helix)*	Alfa
S	*bend*	Coil
T	*hydrogen bonded turn*	Coil

Geração dos Arquivos

Como já exposto, a intenção deste trabalho é a aquisição do conhecimento interno da RNA. Conforme será visto no item 5.5.2, optou-se pela utilização de 3 redes; cada uma com o foco em uma estrutura secundária a ser determinada; processando padrões adaptados para a classe alvo da extração do conhecimento. Trata-se de uma técnica comum na utilização de RNAs em problemas de classificação.

Devido a isto, 3 arquivos de dados foram gerados para o treinamento das RNAs. No primeiro, os padrões cuja classe era *alfa*, tiveram seus valores de saída setados para 1. Caso contrário, classes *beta* e *coil*, foram setados para -1. No segundo, os padrões cuja classe era *beta*, tiveram seus valores de saída setados para 1; caso contrário, para -1. No terceiro, os padrões cuja classe era *coil*, tiveram seus valores de saída setados para 1; caso contrário, para -1.

Outro procedimento adotado foi o de executar um balanceamento das bases. RNAs são sensíveis à bases desbalanceadas. Podendo vir a adotarem soluções triviais como a de dar uma saída constante onde classifiquem todos os padrões como pertencentes, ou não, à classe. Em outros termos, *respondam* sempre *sim* ou sempre *não*. Nestes casos, alguns neurônios se especializariam em sempre fornecer a mesma resposta. Com isto, as regras extraídas não seriam úteis.

Para resolver este problema, as bases foram balanceadas utilizando-se o critério de aproveitar 100% dos padrões originais da classe a ser determinada. Para as outras duas classes, aproveita-se um valor que seja a metade do que já foi utilizado para a principal. Sendo que os padrões destas duas são escolhidos de forma aleatória. Por exemplo, supondo uma base com 6.000 *alfas*, 7.000 *betas* e 10.000 *coils*. Para uma base direcionada à RNA que irá extrair o conhecimento que leva a formação de *alfas*, teriamos uma base com 6.000 *alfas*, 3.000 *betas* escolhidas

aleatóriamente e 3.000 *coils* também escolhidas de forma aleatória.

5.5.2 Definição da Arquitetura da RNA

Conforme adiantado no item 5.5.1, optou-se pela utilização de uma técnica comum na utilização de RNAs em problemas de classificação. Ou seja, a utilização de uma rede para cada classe a ser determinada, com o foco no problema a ser atendido. Desta forma, evita-se a interferência entre-classes no treinamento. Este problema ocorre quando, na tentativa de ajustar os pesos da rede ao reconhecimento dos padrões de uma classe, pesos já ajustados para outra classe são alterados com conseqüências negativas. As interferências intra-classe, ao contrário, são desejadas sendo o principal fator que propicía a generalização das redes.

Com isto, as explicações que seguem são válidas para as 3 redes. Uma vez que todas apresentam as mesmas características.

Quanto ao tamanho da janela de resíduos, a análise de trabalhos [43, 46, 66], bem como observações feitas por Wu & McLarty [9] e Holbrook *et al.* [4], levaram a escolha de um tamanho igual a 13. Conforme observado nestes estudos, tamanhos de janela superiores não apresentaram ganhos na predição ou, quando apresentaram, não foram significativos.

Para tanto, como entrada à RNA foi definida a utilização de 26 neurônios devido ao processamento dos atributos pI e HP em uma janela de 13 resíduos.

Com relação à camada oculta, testes foram realizados visando identificar um número adequado de neurônios com a intenção de garantir a generalização da solução e facilitar o processo de extração de regras. Foram testadas redes com 4, 8 e 16 neurônios na camada oculta, sendo a escolhida a de 4 neurônios.

Por fim, a definição da camada de saída está intimamente ligada a informação que será calculada. Foi utilizado 1 neurônio para dizer se o padrão pertence (valor igual a 1), ou não (valor igual a -1), à classe.

A camada de entrada foi definida como completamente conectada à oculta e esta, por sua vez, à de saída. Todos os neurônios da camada oculta e de saída foram definidos com *bias*.

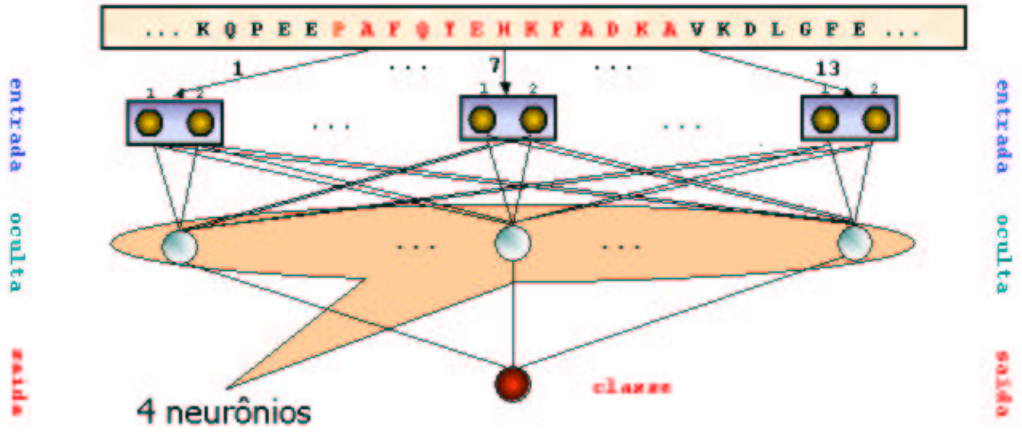

FIGURA 5.4 – Representação simplificada da rede implementada e utilizada

5.5.3 Definição do Algoritmo de Aprendizado

Como ferramenta para treinamento das RNAs foi utilizado o próprio FAGNIS [5], que já possui implementado o algoritmo de aprendizado *Resilient Propagation (RProp)*. A escolha deste algoritmo foi devido aos ótimos resultados reportados por Riedmiller & Braun [67]. As redes geradas seguiram as seguintes características:

- os pesos das conexões inicializados de forma aleatória;
- a ordem de atualização dos pesos determinada pela ordem topológica;
- a função de ativação da camada de entrada definida como *Identity*;
- a função de ativação da camada oculta definida como sigmóide;
- a função de ativação da camada de saída definida como sigmóide;
- demais parâmetros deixados com seus valores padrões.

5.6 Encerramento do Capítulo

Neste capítulo foram apresentados os principais pontos considerados na abordagem sobre como executar a extração de regras de RNAs. Foram analisados o software de extração de regras, as implementações feitas e a metodologia definida.

No próximo capítulo serão apresentados os resultados obtidos a partir da metodologia aqui empregada.

Capítulo 6

Resultados Obtidos

Neste capítulo serão apresentados os resultados obtidos a partir da aplicação da metodologia definida no capítulo anterior.

Isto será feito através da análise das regras extraídas para uma rede treinada com os atributos físico-químicos propostos (*i.e.* ponto isoelétrico e índice hidropático). Neste procedimento serão reportados os resultados das métricas implementadas. Adicionalmente, embora não seja do escopo deste trabalho apresentar um conhecimento bioquímico a ponto de permitir uma análise no nível da biologia molecular, considerações básicas sobre o conhecimento extraído serão apresentadas.

Salienta-se que o conhecimento extraído por esta, ou qualquer outra metodologia, deve ser analisado por profissionais que detenham a competência para tal. Especificamente, no contexto deste estudo, por biólogos moleculares. Estes, por sua vez, devem fornecer um retorno para que a técnica seja aprimorada nos pontos que se façam necessários. Desta forma, fechando um ciclo de produção de informação, consumo e retorno para uma produção mais exata.

Antes desta análise, porém, visando demonstrar que a técnica utilizada é capaz de chegar a um conhecimento já existente, será apresentado o resultado de uma extração de regras a partir de uma codificação ortogonal. Desta forma, será possível fazer uma comparação qualitativa do método aqui proposto com o conhecimento disponibilizado por Qian & Sejnowski [43].

6.1 Comparação com o Estudo de Qian & Sejnowski

Conforme apresentado no capítulo 4, o estudo de Qian & Sejnowski [43] utilizou uma codificação ortogonal para o processamento da seqüência de resíduos e definição da estrutura secundária. Além do que foi apresentado, o estudo também expôs a influência dos resíduos através

do processamento destes por uma RNA sem neurônios na camada oculta, ou seja, uma solução linear. Como resultado, foi gerado um diagrama de Hinton [68] para cada classe determinada.

O diagrama de Hinton pode ser interpretado como uma forma de representar a influência das entradas na determinação da saída. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

O resultado que irá determinar se o padrão pertence, ou não, à classe é dado pelo somatório do valor da multiplicação entre cada atributo de entrada e o coeficiente correspondente da equação linear. E, ao final, acrescentado o termo independente k. No caso da rede implementada, um valor maior ou igual a zero foi convencionado como pertencente à classe. E, menor do que zero, como não pertencente.

Em uma codificação ortogonal, como a aqui utilizada, a tendência pode ser vista diretamente pela representação do quadrado. Visto que as entradas são apenas de 0s e 1s, basta analisar diretamente o quadrado referente ao resíduo na posição onde o mesmo aparece na janela. Quadrados brancos são positivos e com o caráter excitatório, que "levam" à classe. Quadrados pretos são negativos, inibitórios, que "afastam" da classe.

Visando comprovar que a técnica aqui utilizada é capaz de chegar a um conhecimento já existente, foi realizado um processamento semelhante ao de Qian & Sejnowski onde os resultados podem ser vistos nas figuras 6.1, 6.2 e 6.3. Para cada figura, o primeiro diagrama (no canto superior esquerdo) representa o resultado original obtido por Qian & Sejnowski. Os demais diagramas representam os resultados das regras obtidas a partir da aplicação da metodologia aqui descrita.

A partir da análise dos diagramas apresentados, observações podem ser feitas a respeito das tendências de cada aminoácido na formação da estrutura secundária. Conforme pode ser observado nos itens a seguir, a maioria do conhecimento extraído corroborou o já existente. Porém, com o diferencial do resultado ter sido obtido através de uma técnica que é uma generalização para uma solução não-linear.

Para alguns casos foram ressaltadas as situações onde as regras extraídas apresentaram uma informação significativamente diferente da original reportada por Qian & Sejnowski. Nestes pontos, há a necessidade de uma investigação mais detalhada pois pode indicar um novo conhecimento válido ou uma deficiência do método utilizado. Segue a relação, das principais observações pertinentes:

- Alanina (ALA): tendência em formar *alfa*;

- Arginina (ARG): tendência em formar *alfa* quando aparece no centro ou no lado direito da janela; merece investigação a forte tendência em formar *beta* quando aparece na nona posição da janela bem como a forte tendência em inibir a formação de *coil* quando aparece na sexta posição da janela;
- Asparagina (ASN): tendência em formar *coil*;
- Aspartato (ASP): tendência em formar *coil*; merece investigação a tendência em formar *alfa* quando aparece no lado esquerdo da janela;
- Glutamato (GLU): tendência em formar *alfa*;
- Glicina (GLY): tendência em formar *coil*;
- Histidina (HIS): merece investigação a tendência em formar *alfa* quando aparece no lado esquerdo da janela bem como a forte tendência em inibir a formação de *beta* quando aparece no lado esquerdo da janela;
- Isoleucina (ILE): tendência em formar *beta*; merece investigação a leve tendência em formar *alfa* quando nas extremidades da janela, ao invés de *coil*;
- Leucina (LEU): tendência em formar *alfa*;
- Lisina (LYS): leve tendência em formar *alfa*;
- Metionina (MET): forte tendência em formar *alfa*;
- Prolina (PRO): merece investigação a tendência na formação de *coil* ter sido detectada mais fortemente no oitavo resíduo da janela, e vizinhos; também merece investigação a baixa tendência em inibir a formação de *alfa*, quando na quinta posição da janela;
- Triptofano (TRP): merece investigação a tendência em formar *alfa*, reportada para algumas regras, quando nas posições 4 a 7 da janela; também merece investigação a alta tendência em inibir a formação de *beta*, quando na extremidade esquerda da janela;
- Tirosina (TYR): tendência em formar *beta*; merece investigação a forte tendência em inibir a formação de *alfa*, reportada para algumas regras, quando nas posições 5, 6 e 11 da janela;
- Valina (VAL): tendência em formar *beta*;

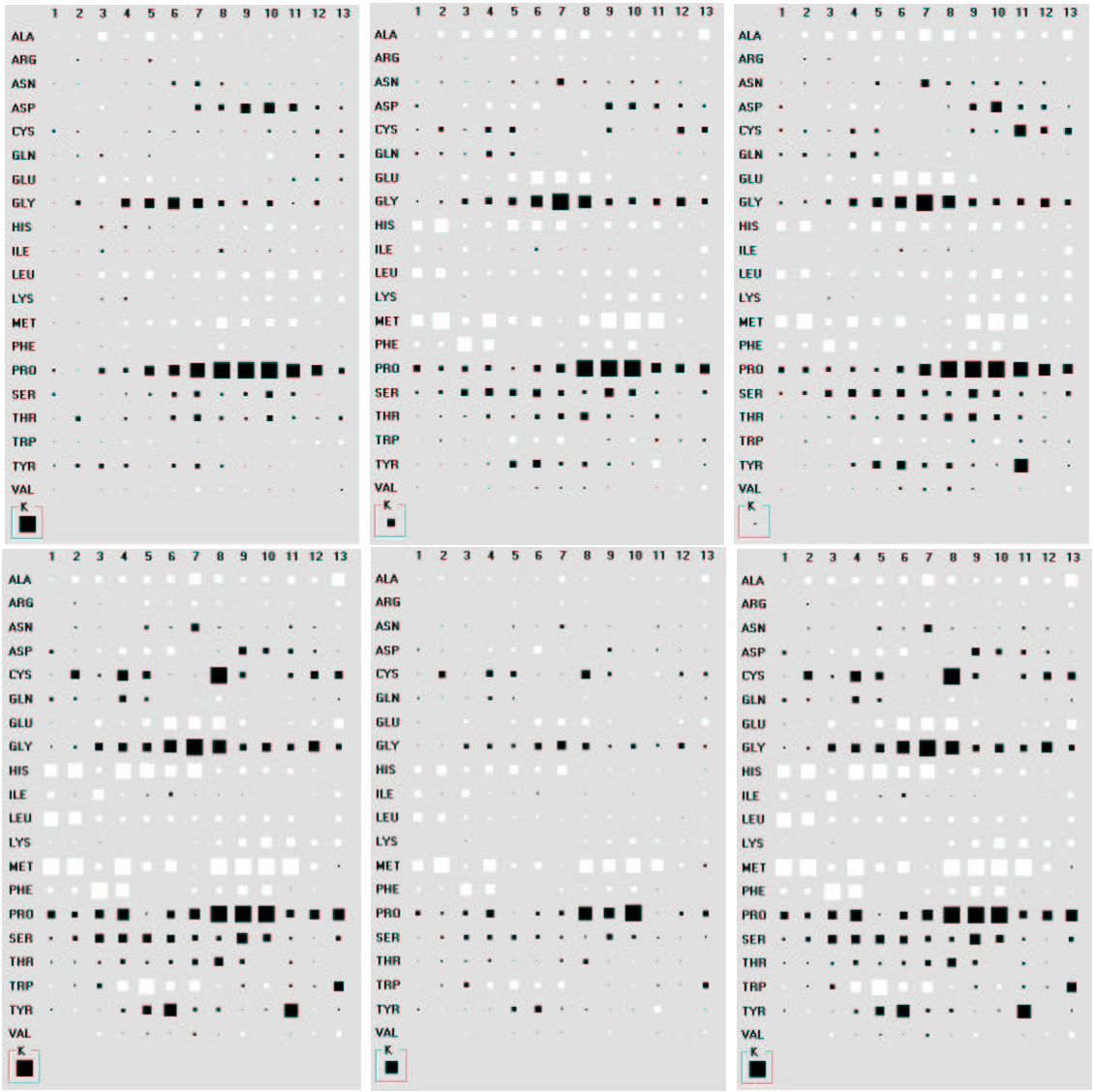

FIGURA 6.1 – Diagramas de Hinton para a classe *alfa*. Diagrama original de Qian & Sejnowski (canto superior esquerdo) e 5 diagramas (regras 10, 1, 20, 2 e 16) obtidos a partir da aplicação da metodologia aqui descrita. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

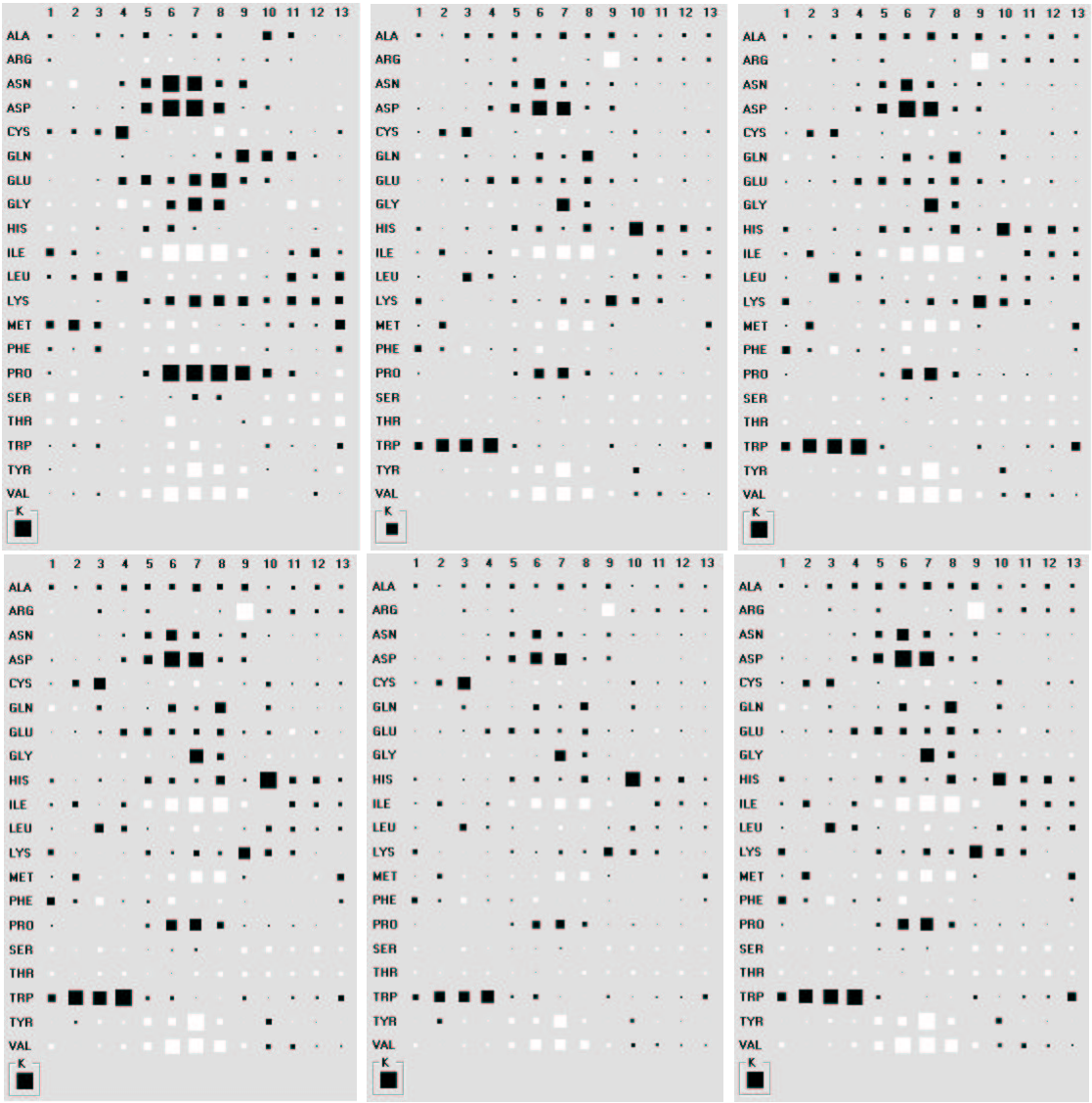

FIGURA 6.2 – Diagramas de Hinton para a classe *beta*. Diagrama original de Qian & Sejnowski (canto superior esquerdo) e 5 diagramas (regras 1, 24, 31, 42 e 51) obtidos a partir da aplicação da metodologia aqui descrita. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

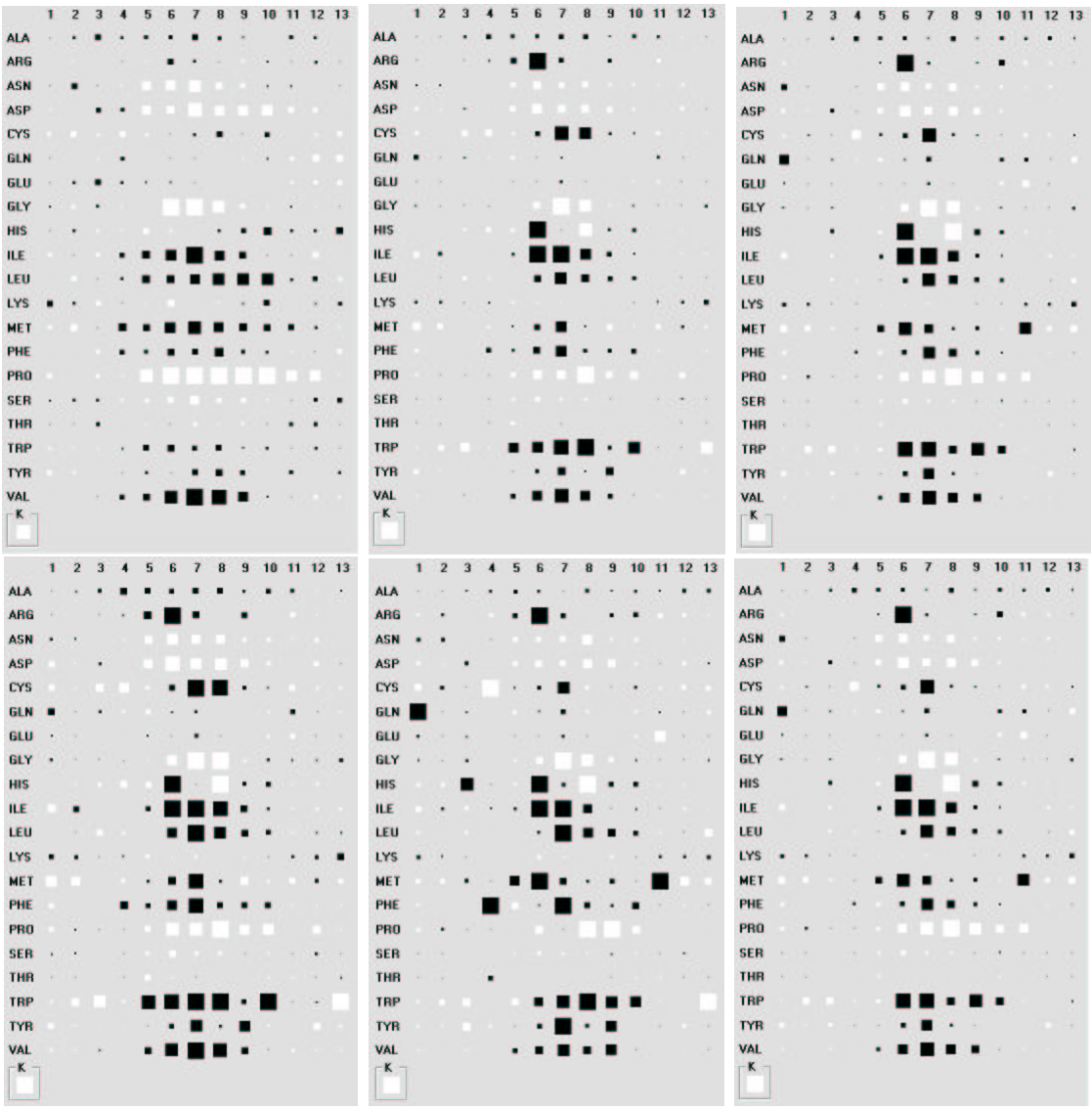

FIGURA 6.3 – Diagramas de Hinton para a classe *coil*. Diagrama original de Qian & Sejnowski (canto superior esquerdo) e 5 diagramas (regras 4, 6, 10, 13 e 32) obtidos a partir da aplicação da metodologia aqui descrita. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

6.2 Análise do Comportamento das Regras Extraídas ao Longo do Treinamento

Durante o processo de treinamento e extração de regras, para cada época, informações são geradas e armazenadas para posterior análise. Dentre estas informações estão: o número de regras geradas, o número de regras estatisticamente válidas, a cobertura total (conforme equação 5.8) e o erro de inexatidão global (conforme equação 5.13). Adicionalmente, quanto a RNA, são armazenados: o erro de inexatidão da rede e o seu erro RMS. Por "erro de inexatidão da rede" entenda-se o resultado do processamento do cálculo de inexatidão, conforme equação 5.11 porém feito a partir do resultado da RNA. Esta medida de erro foi implementada por dois motivos: para a comparação com a inexatidão da solução linear proposta; e pelo problema tratado ser de classificação.

Para cada classe a ser determinada, a rede correspondente foi treinada e os procedimentos de extração de regras executados em conjunto. Inicialmente, o treinamento foi realizado por 20.000 épocas, visando-se obter uma noção aproximada do problema. Os testes seguintes foram restritos a 3.000 épocas embora, como será visto mais adiante, bem menos épocas seriam suficientes. Este valor foi utilizado apenas para a produção dos gráficos e análise do comportamento do erro global no decorrer do treinamento.

A tabela 6.1 apresenta informações pertinentes a solução encontrada para cada classe. Em *Regras geradas* é reportado o número de regras que foram extraídas para cada classe. Destas, o número de regras válidas (vide item 5.3) é apresentado em *Regras usadas*. O percentual de padrões atendidos por estas regras é dado por *Cobertura*. Por fim, o percentual de erros de classificação e o Erro RMS são reportados, respectivamente, nas colunas *Inexatidão* e *Erro RMS*.

TABELA 6.1 – Número de regras extraídas para cada classe

Classe	Regras geradas	Regras usadas	Cobertura (%)	Inexatidão (%)	Erro RMS
alfa	67	45	98.45	31.79	0.929386
beta	62	23	97.66	31.98	1.120927
coil	78	41	96.76	29.95	0.888196

A figura 6.4 apresenta estas informações para as 3 classes determinadas. Os valores foram recalculados para serem apresentados em uma escala onde todas as informações pudessem ser visualizadas. Desta forma, ficando com valores entre 0 e 1. O número de regras total e o de regras válidas foi dividido pelo maior valor existente entre ambos. Os percentuais de cobertura

total, inexatidão da solução linear e inexatidão da RNA foram divididos por 100. E o erro RMS da RNA foi dividio pelo seu maior valor.

Com isto, ao analisar os gráficos, devem ser buscadas tendências de comportamento. Comparações com valores absolutos não podem ser feitas. A menos que seja entre as inexatidões.

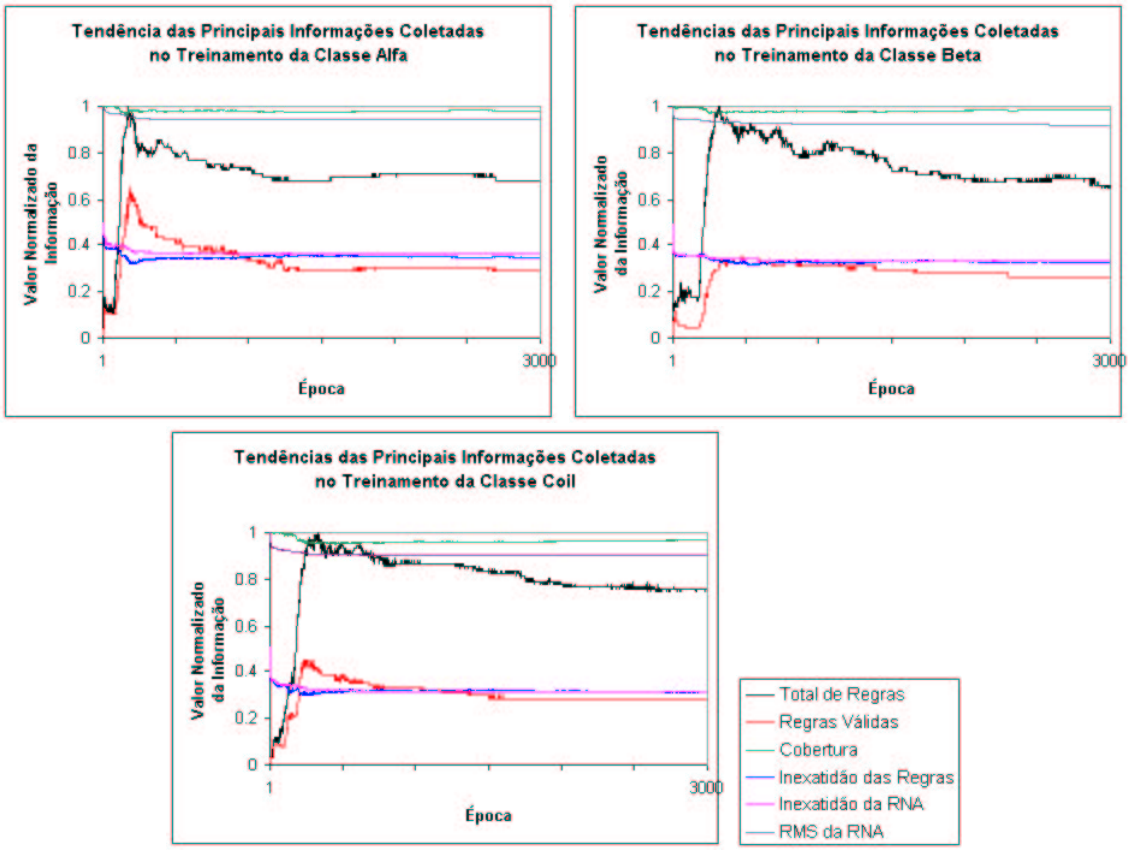

FIGURA 6.4 – Tendência das principais informações coletadas durante o processo de treinamento das redes. O eixo x apresenta a época do treinamento. O eixo y apresenta as informações coletadas onde os valores foram normalizados para serem apresentados em uma escala de 0 a 1.

Uma análise das tendências reportadas na figura 6.4, nos leva a alumas constatações interessantes.

Observando-se a curva do número total de regras, pode-se constatar o processo de aprendizado da rede onde, aproximadamente até a época 300, os neurônios estão sendo sub-utilizados, isto é, apenas pequenas faixas de suas regiões estão ativas. Com isto, o número de regras é pequeno. A partir do momento em que a rede começa a se especializar, novas regiões dos neurônios

são utilizadas. Neste ponto, nota-se um aumento rápido no número de regras total. Porém, como estas regiões estão recebendo poucos padrões, o número de regras estatísticamente válidas não cresce na mesma medida que o total, conforme pode ser melhor observado na figura 6.5. Como conseqüência, a cobertura total apresenta um descréscimo nesta fase.

Outro ponto que pode ser observado é que, após o número de regras total atingir um valor máximo, mantem-se uma proporção entre o número de regra total e o número de regras válidas. Uma explicação possível seria devido ao fato de, após atingir o ápice no número de regras geradas, as regras que deixam de existir, pela mundaça das ativações nos neurônios da rede, são as com poucos padrões. Padrões estes que, como o aprendizado, vão migrando para outras regiões do espaço neural.

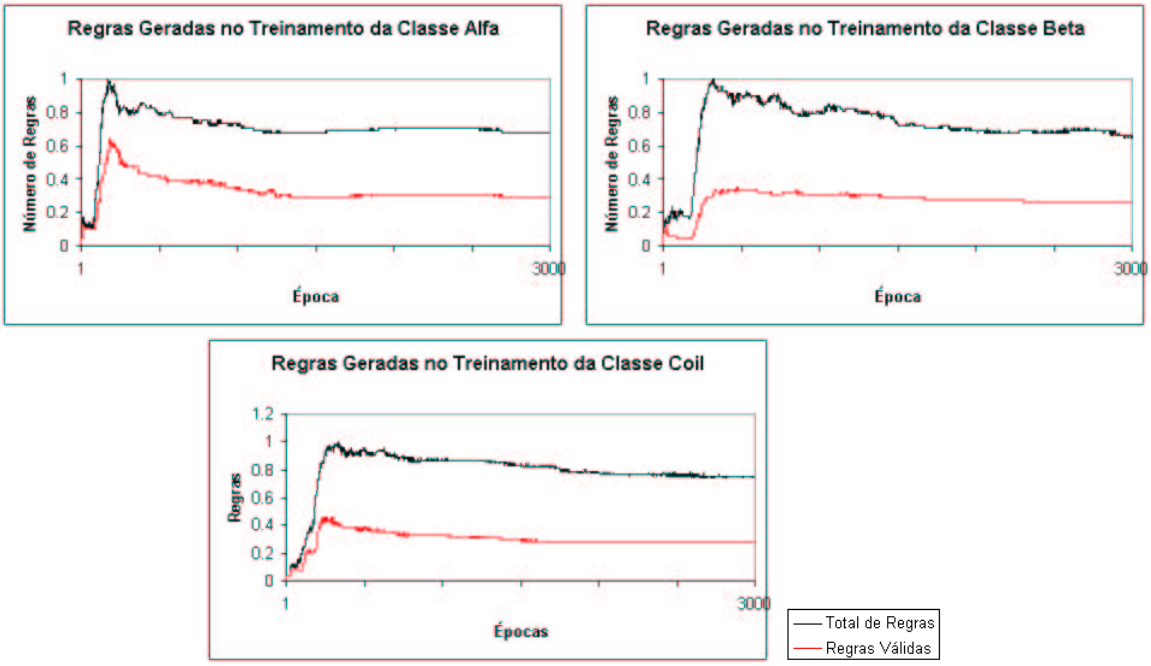

FIGURA 6.5 – Total de regras geradas e regras válidas. O eixo x apresenta a época do treinamento. O eixo y apresenta o número de regras onde os valores foram normalizados para serem apresentados em uma escala de 0 a 1.

Também pode ser visto, sendo melhor observado na figura 6.6, é o fato da inexatidão linear ser menor do que a não-linear. Ou seja, a soma das soluções lineares apresenta um resultado melhor do que a RNA original. Um explicação possível seria pela especialização das equações lineares, melhor adapatadas aos padrões específicos da sua região. Deve ser lembrado que a solução linear apresenta uma perda na cobertura de padrões.

Outra observação que pode ser feita analisando-se a figura 6.6 é que, com o decorrer do treinamento das classes *beta* e *coil*, a inexatidão da RNA se igualou a da solução linear. Isto pode ter acontecido pois a RNA vai se especializando em regiões onde poucos padrões são atendidos. Fato este que a diferencia da solução linear pois estas regiões, ao serem linearizadas, podem vir a ser desconsideradas se o número de padrões for inferior ao número de parâmetros do modelo.

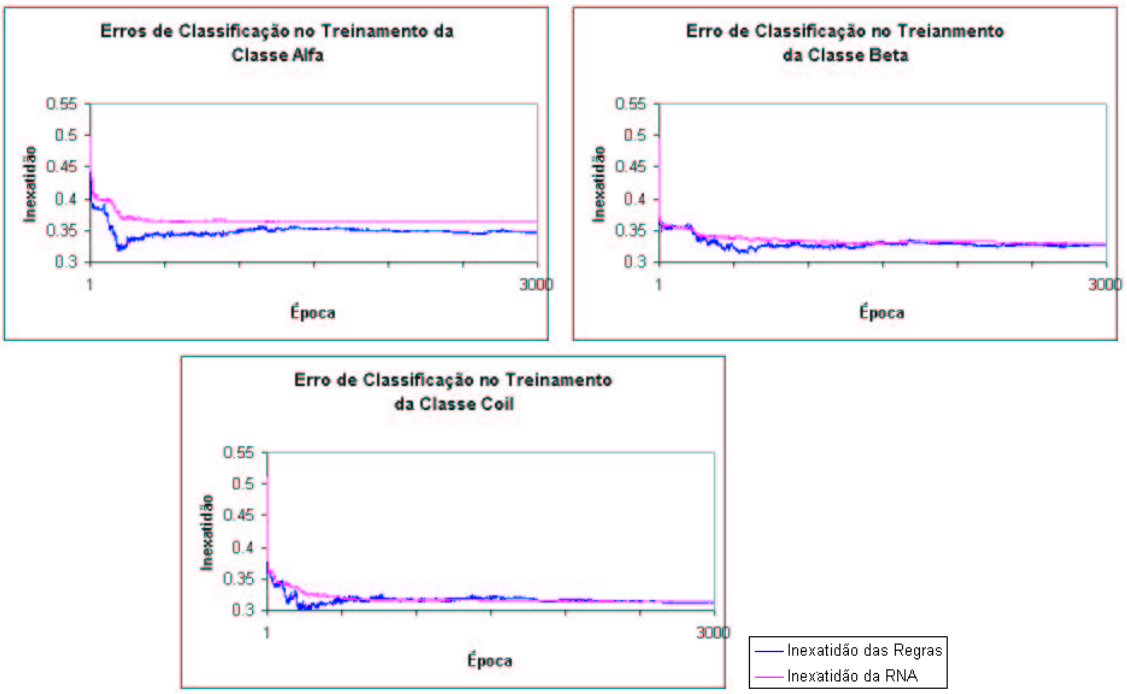

FIGURA 6.6 – Inexatidão global da solução linear e inexatidão da RNA. O eixo x apresenta a época do treinamento. O eixo y apresenta as informações percentuais coletadas para as taxas de inexatidão.

Tratam-se de constatações qualitativas baseadas no treinamento de poucas inicializações randomicas dos pesos da rede. Uma análise mais detalhada merece ser feita no sentido de ser executado um número de experimentos considerado válido para cálculo de médias e desvio-padrão. Desta forma, pode-se ter uma visão mais embasada para generalizações.

6.3 Análise das Regras Extraídas

Nesta seção serão apresentados os resultados obtidos a partir da análise feita sobre algumas regras extraídas das redes treinadas com os atributos físico-químicos propostos (*i.e.* ponto

isoelétrico e índice hidropático).

Duas abordagens foram utilizadas para interpretar estas regras. Primeiro, as regras foram ordenadas pela sua *cobertura*. Sendo analisadas as que apresentaram os maiores valores, isto é, com maior quantidade de padrões atendidos. Segundo, as regras foram ordenadas pela sua *inexatidão*. Sendo analisadas as com menores valores, isto é, com menor erro de classificação. Para cada caso, serão apresentadas as primeiras 3 regras, ao invés de uma análise exaustiva de todas.

Nesta exposição, dois conjuntos de informações se repetirão. Primeiro, uma tabela contendo informações sobre as regras a serem analisadas. A seguir, para cada regra, o gráfico do seu protótipo e o diagrama de Hinton para os coeficientes da equação linear. No gráfico do protótipo, que descreve o comportamento típico dos padrões pertencentes à regra (conforme item 5.2), poderá ser observada a variação do ponto isoelétrico e do índice hidropático ao longo da janela de resíduos. Onde, no eixo x, é apresentada a janela de 13 resíduos e, no eixo y, o valor do coeficiente.

Quanto ao diagrama de Hinton (apresentado no item 6.1), poderão ser observadas as influências do pI e HP ao longo da janela. Cabe ressaltar que as entradas não mais são binárias.

Com a abordagem utilizada, serão analisadas cerca de 25% das regras válidas da classe *beta* e 15% das classes *alfa* e *coil*. Regras significativas, no que diz respeito a informações contidas, podem estar sendo deixadas de lado. Porém, devido à limitação de tempo, esta foi a alternativa encontrada.

Conforme poderá ser notado, pela leitura das subseções que seguem, a imediatamente a seguir (item 6.3.1) apresenta um conteúdo mais extenso do que as demais. Especificamente, os itens existentes a mais são: uma análise mais estruturada de um protótipo e de um diagrama de Hinton; a corroboração de um protótipo; e a equação linear gerada para uma regra extraída. A intenção de detalhar um pouco mais estes itens foi a de apresentar um embasamento para a compreensão das explicações. Nas seções que seguem, estes itens foram omitidos visando não deixar a estrutura muito repetitiva.

Adicionalmente, visando apresentar algumas das variações na interpretação do conhecimento contido nos gráficos dos protótipos, as explicações sobre os mesmos foram escritas de forma diversificada. Deste modo, espera-se demonstrar algumas das possibilidades de como interpretar e expressar o conhecimento bem como tornar a leitura menos cansativa. Isto pode ser visto através da alternância de tipos de análise (*e.g.* hora se preocupando com periodicidades, hora com valores, hora com busca de relações) bem como no emprego de sinônimos (*e.g.* ponto isoelétrico/pI/carga, índice hidropático/HP, positivo/negativo/hidrofóbico/hidrofílico).

6.3.1 Regras da Classe Alfa por Cobertura

Neste tópico, serão apresentadas as regras da classe *alfa* com maior percentual de cobertura, isto é, que abrangem o maior número de padrões atendidos.

A tabela 6.2 apresenta as 3 regras com maior percentual de cobertura. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton, para os coeficientes da equação linear, podem ser vistos na figura 6.7.

TABELA 6.2 – Primeiras 3 regras com maior percentual de cobertura da classe *alfa*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
9	18.92	37.34	0.957974
6	16.48	33.68	0.940372
12	9.67	39.38	0.991682

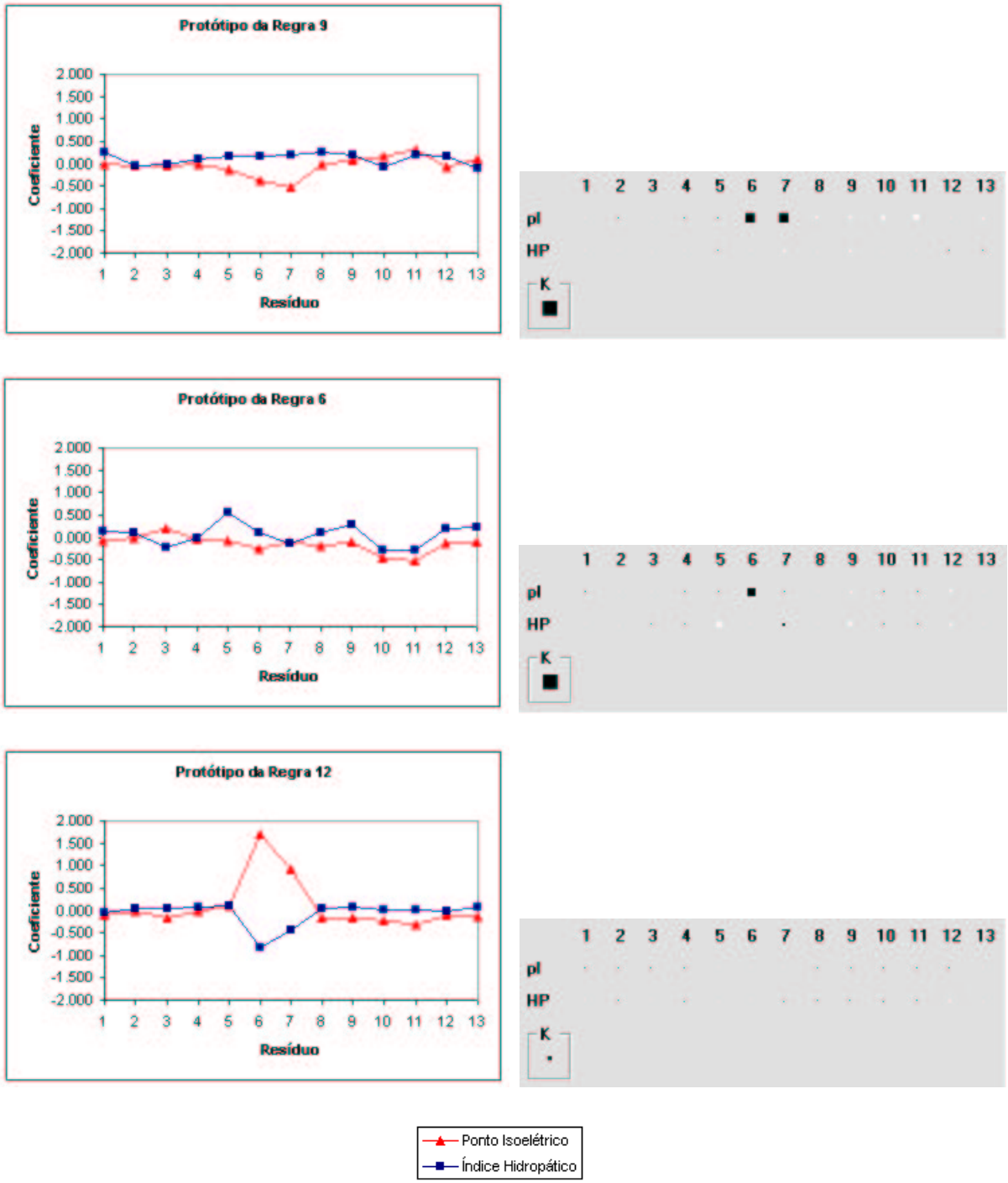

FIGURA 6.7 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 9, 6 e 12), por cobertura, da classe *alfa*. O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação ao protótipo da regra 6, as seguintes observações podem ser feitas:

- analisando o ponto isoelétrico:
  - o resíduo central apresenta uma carga muito próxima de zero;
  - os dois primeiros resíduos, vizinhos ao central, apresentam um comportamento de se tornarem negativos e, logo após, voltarem a ser neutros;
  - a partir do 3 resíduo, após o central, nota-se uma simetria inversa nas tendências, ou seja, enquanto uma direção apresenta decréscimo de carga, a outra apresenta aumento, e vice-versa;
- analisando o índice hidropático:
  - o resíduo central apresenta-se em um equilíbrio entre hidrofobicidade e hidrofilicidade;
  - tomando o ponto central como partida, nota-se uma periodicidade, em ambas as direções, onde ocorre um aumento da hidrofobicidade por 2 resíduos, passando para hidrofílicos por 2 e voltam a ser hidrofóbicos por mais 2.

Com relação ao diagrama de Hinton da regra 6, as seguintes observações podem ser feitas:

- analisando o ponto isoelétrico:
  - o sexto resíduo apresenta uma tendência alta, comparado contra os demais, em indicar a formação, ou não, de *alfas*; quanto mais positivo for o pI do sexto resíduo, menor é a tendência de formar uma *alfa* e quanto mais negativo, maior a tendência;
- analisando o índice hidropático:
  - o quinto e nono resíduos apresentam as maiores tendências, comparados contra os demais, em indicar a formação, ou não, de *alfas*; quanto mais positivos forem os HPs destes resíduos, maior é a tendência de formar uma *alfa* e quanto mais negativo, menor a tendência;

Com relação à regra 9, não se observa periodicidade alguma no pI, exceto uma divisão, a partir do centro, entre valores negativos e positivos. Quanto ao HP, o resíduo central apresenta-se

como hidrofóbico, assim como praticamente toda a cadeia. Oscilações mais fortes são detectas nas extremidades.

No que diz respeito às tendências da regra 9, o pI do sexto e sétimo resíduos apresenta uma grande contribuição na determinação, ou não, de *alfas*; quanto mais positivo forem os pIs, menor é a tendência de formar uma *alfa* e, quanto mais negativo, maior a tendência.

Com relação à regra 12, picos de carga e índice hidropático são observados no resíduo central e no imeditamente à sua esquerda. No resto da janela, ambos se comportam com valores muito próximos de 0.

No que diz respeito às tendências da regra 12, tanto pI quanto o HP apresentam, na maioria dos resíduos da janela, um equilíbrio nas tendências. Com isto, qualquer valor mais alto ou baixo, no valor do atributo pertencente ao resíduo, pode contribuir de forma mais direta na determinação estrutura secundária.

De forma geral, nota-se que as tendências, indicadas pelos diagramas de Hinton, não possuem grandes especializações com relação a um ou mais resíduos da janela. Como se tratam de regras com as maiores quantidades de padrões, isto pode indicar que os dados cobertos ainda sejam uma mescla de várias tendências distintas. E, na média, os valores não se sobressaem. Eventualmente, caso a rede conseguisse separar melhor os padrões, provavelmente regras mais especializadas seriam geradas e, por conseguinte, tendências mais distintas seriam notadas. Até as 20.000 épocas inicialmente utilizadas para o treinamento, isto não foi detectado.

Corroborando o conhecimento fornecido pelos protótipos, podem ser apresentados os padrões, usados no treinamento, que possuam o maior grau de pertinência para com a sua regra. Tomando como base o protótipo da regra 6, podemos observar na figura 6.8 a distribuição dos pontos no espaço de entrada, para o ponto isoelétrico e para o índice hidropático.

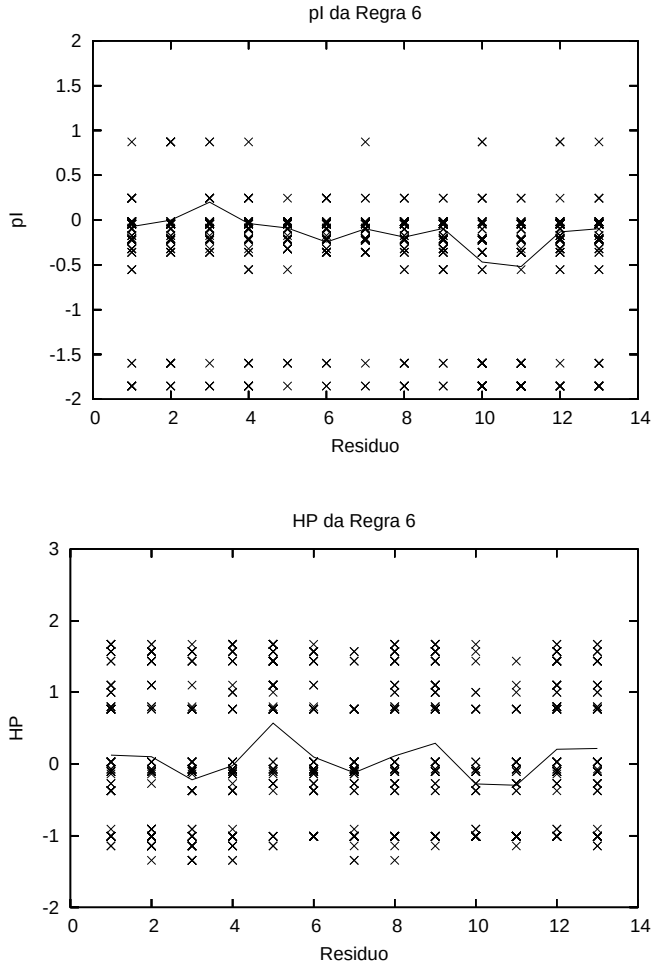

FIGURA 6.8 – Corroboração da regra 6, da classe *alfa*

Contudo, a visualização da quantidade de padrões com, por exemplo, pI igual a -0,044099 (da Glicina) ficou dificultada pois todos são apresentados no mesmo ponto. Para contornar este problema, optou-se pela representação através de uma alteração mínima no valor de entrada, causando um pequeno deslocamento para ter-se a representação de intensidade. A corroboração efetiva do padrão pode ser vista na figura 6.9.

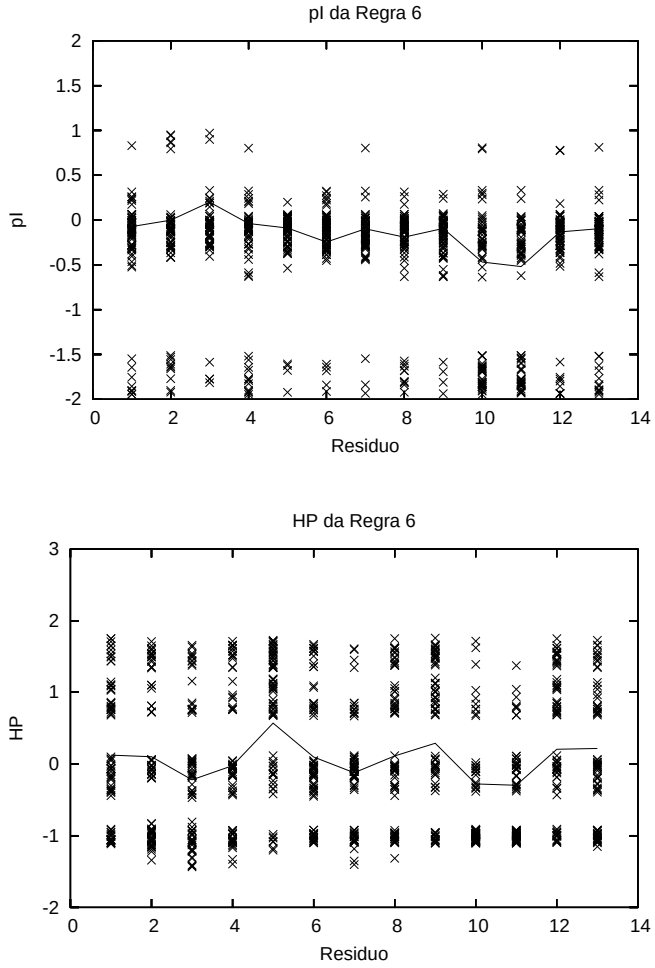

FIGURA 6.9 – Corroboração da regra 6, da classe *alfa*, com representação de intensidade dos pontos

Conforme apresentado no item 5.2, para cada regra, além do protótipo que caracteriza os padrões por ela atendidos, também é disponibilizada uma equação linear que irá determinar a saída encontrada. Neste caso, o resultado da classificação. Na equação 6.1, pode ser vista a estrutura que as equações lineares geradas apresentam. Os 13 resíduos da janela, seus respectivos coeficientes, resultado do algoritmo de extração de regras, e o termo independente.

$$
\begin{aligned}
R_6: \quad & -0.034093Pi_1 + 0.081456Hp_1 + 0.004571Pi_2 + 0.049922Hp_2 \\
& +0.021662Pi_3 - 0.084841Hp_3 - 0.044895Pi_4 - 0.023076Hp_4 \\
& -0.056758Pi_5 + 0.214972Hp_5 - 0.341631Pi_6 + 0.025795Hp_6 \\
& -0.079390Pi_7 - 0.149484Hp_7 + 0.063821Pi_8 + 0.084386Hp_8 \\
& +0.092090Pi_9 + 0.209021Hp_9 - 0.060223Pi_{10} - 0.018637Hp_{10} \\
& -0.004262Pi_{11} - 0.072064Hp_{11} + 0.128643Pi_{12} + 0.098707Hp_{12} \\
& +0.063558Pi_{13} + 0.065529Hp_{13} - 0.656367
\end{aligned} \tag{6.1}
$$

6.3.2 Regras da Classe Alfa por Inexatidão

Neste tópico, serão apresentadas as regras da classe *alfa* com menor inexatidão, isto é, que possuem o menor erro de classificação.

A tabela 6.3 apresenta as 3 regras com menor erro. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton para os coeficientes da equação linear podem ser vistos na figura 6.10.

TABELA 6.3 – Primeiras 3 regras com menor erro da classe *alfa*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
32	0.37	0	0.591280
24	0.36	0	0.820214
53	0.34	0	0.218695

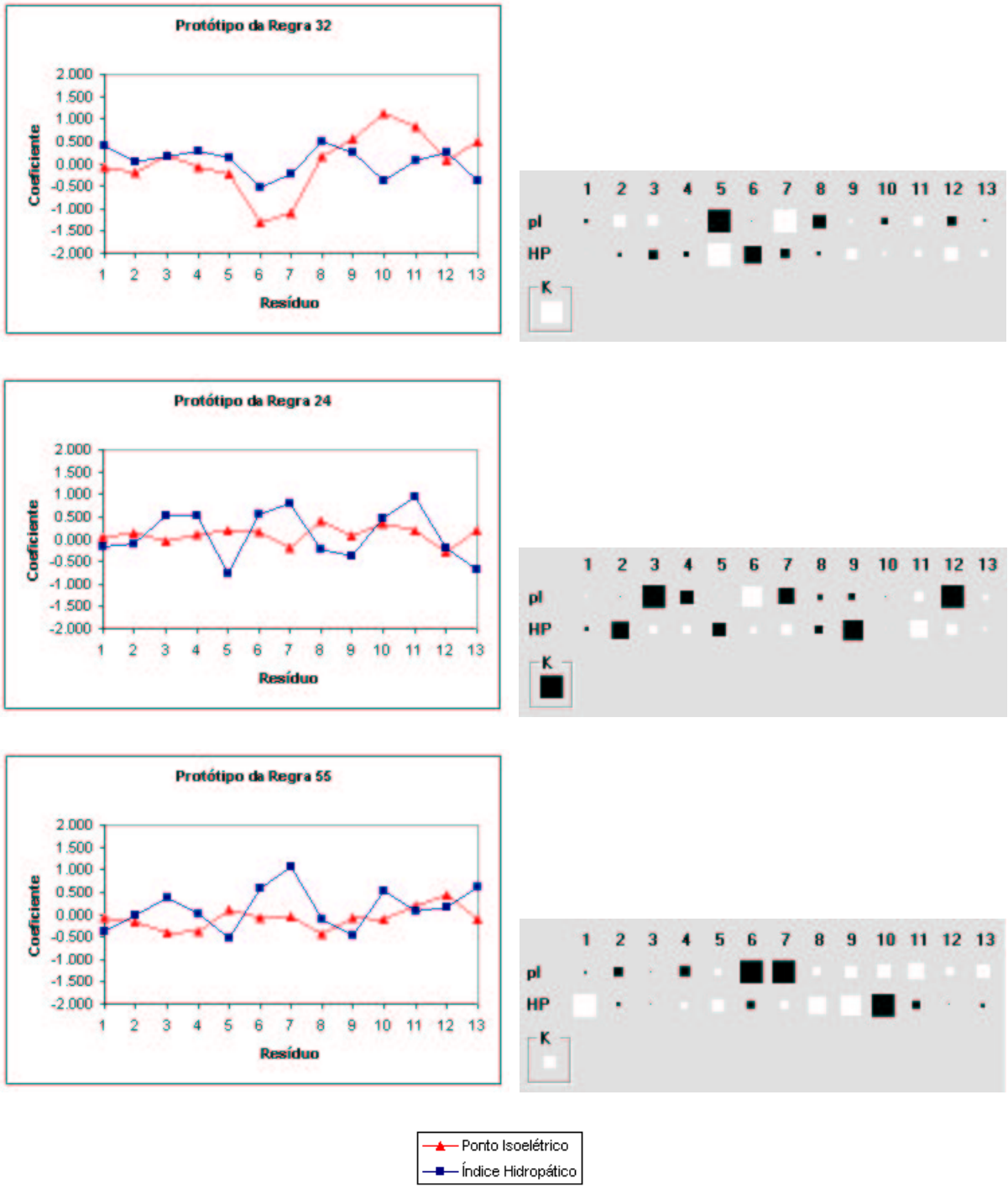

FIGURA 6.10 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 32, 24 e 55), por inexatidão, da classe *alfa* O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação à regra 32, picos no pI podem ser observados nos resíduos 6 e 10. Sendo que, a oscilação entre ambos pode ser considerada como grande, dada a distância entre os mesmos. Tomando-se o sexto resíduo como base, pode ser observada uma periodicidade com relação ao HP onde, de 2 em 2, os resíduos aumentam e diminuem de valor.

No que diz respeito às tendências da regra 32, o pI do quinto e sétimo resíduos apresenta uma grande contribuição na determinação, ou não, de *alfas*. Quanto mais positivo for o valor do pI do quinto resíduo, menor é a tendência de formar uma *alfa* e quanto mais negativo, maior a tendência. O inverso vale para o sétimo resíduo. Quanto ao HP, o quinto e o sexto resíduo são os que mais contribuem. Se todos os resíduos tiverem pI e HP próximos de 0, a tendência é pela formação de *alfa* visto que o termo independente k possui um alto valor positivo.

Com relação à regra 24, tomando-se o resíduo central como base, pode ser observada uma periodicidade com relação ao HP onde, de 2 em 2, os resíduos diminuem e aumentam de valor. Embora o resíduo central seja levemente negativo, quase que toda a janela apresenta-se como levemente positiva.

No que diz respeito às tendências da regra 24, o pI de 5 dos resíduos foi considerado importante para determinar, ou não, a formação de *alfas*. O HP teve uma importância relativa menor mas espalhada por quase toda a extensão da janela.

Com relação à regra 55, nota-se o pico de pI no resíduo 12, ou seja, distante do resíduo central. O HP apresenta uma oscilação, grande parte de 2 em 2 resíduos ao longo da janela. Nesta oscilação, nota-se a passagem de hidrofilicidade e hidrofobicidade.

No que diz respeito às tendências da regra 55, nota-se que o do resíduo central, e do imeditamenta a sua esquerda, foram considerados os mais importantes para determinar, ou não, a formação de *alfas*. O HP teve uma importância maior na extremidade esquerda janela e no lado direito da mesma.

De forma geral, nota-se que as tendências, indicadas pelos diagramas, possuem grandes especializações com relação a um ou mais resíduos da janela. Isto, provavelmente, indica que a rede conseguiu separar melhor os padrões e, por conseguinte, tendências mais distintas foram detectadas.

Visando corroborar o conhecimento apresentado pelo protótipo da regra 32, o padrão de mais alto grau de pertinência foi analisado na proteína correspondente. De toda a a seqüência primária, selecionou-se a parte que contém a janela de 13 resíduos, conforme pode ser observado na figura 6.11. A seqüência foi marcada com a cor magenta e o resíduo predito como *alfa* marcado em verde.

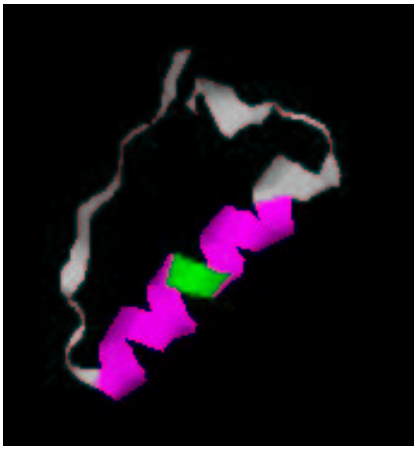

FIGURA 6.11 – Janela de 13 resíduos da proteína 1abe (*L-ARABINOSE da Escherichia coli*) predita como tendo uma *alfa* em sua seqüência

6.3.3 Regras da Classe Beta por Cobertura

Neste tópico, serão apresentadas as regras da classe *beta* com maior percentual de cobertura, isto é, que abrangem o maior número de padrões atendidos.

A tabela 6.4 apresenta as 3 regras com maior percentual de cobertura. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton para os coeficientes da equação linear podem ser vistos na figura 6.12.

TABELA 6.4 – Primeiras 3 regras com maior percentual de cobertura da classe *beta*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
8	24.53	35.26	0.946301
12	10.61	33.02	0.938120
7	7.99	31.2	0.912490

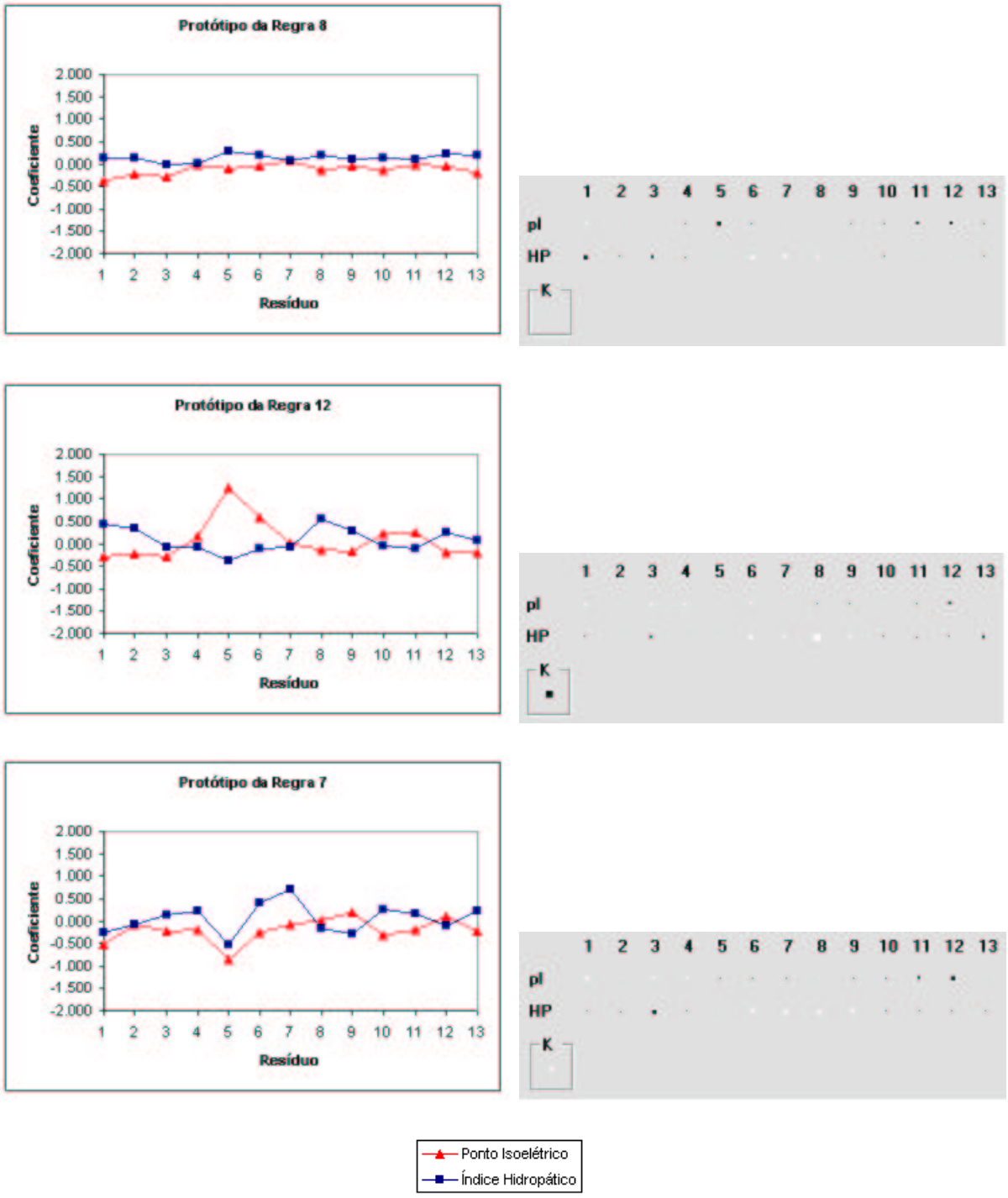

FIGURA 6.12 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 8, 12 e 7), por cobertura, da classe *beta* O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação à regra 8, nota-se um comportamento muito semelhante entre os atributos pI e HP, quase um espelhamento de tendências. Quando um se aproxima de 0 o outro também se aproxima. Quando um se afasta, o outro também se afasta. Outro ponto é a não existência de valores "altos" em ambos os atributos.

Com relação à regra 12, observa-se uma periodicidade no índice hidropático. O ponto isoelétrico apresenta um pico no quinto resíduo e, do lado direito da janela, uma periodicidade que se alterna de 2 em 2 resíduos.

Com relação à regra 7, observa-se uma simetria do HP, com relação às tendências, e não a valores, os resíduos ao redor do central. Quanto ao pI, nota-se uma simetria inversa de tendências também quando se toma como base o resíduo central.

De forma geral, a exemplo do que já havia sido detectado na análise da determinação da classe alfa, nota-se que as tendências não possuem grandes especializações com relação a um ou mais resíduos da janela.

6.3.4 Regras da Classe Beta por Inexatidão

Neste tópico, serão apresentadas as regras da classe *beta* com menor inexatidão, isto é, que possuem o menor erro de classificação.

A tabela 6.5 apresenta as 3 regras com menor erro. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton para os coeficientes da equação linear podem ser vistos na figura 6.13.

TABELA 6.5 – Primeiras 3 regras com menor erro da classe *beta*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
36	0.42	0	0.640554
35	0.54	4.65	0.815124
9	0.45	5.56	1.083139

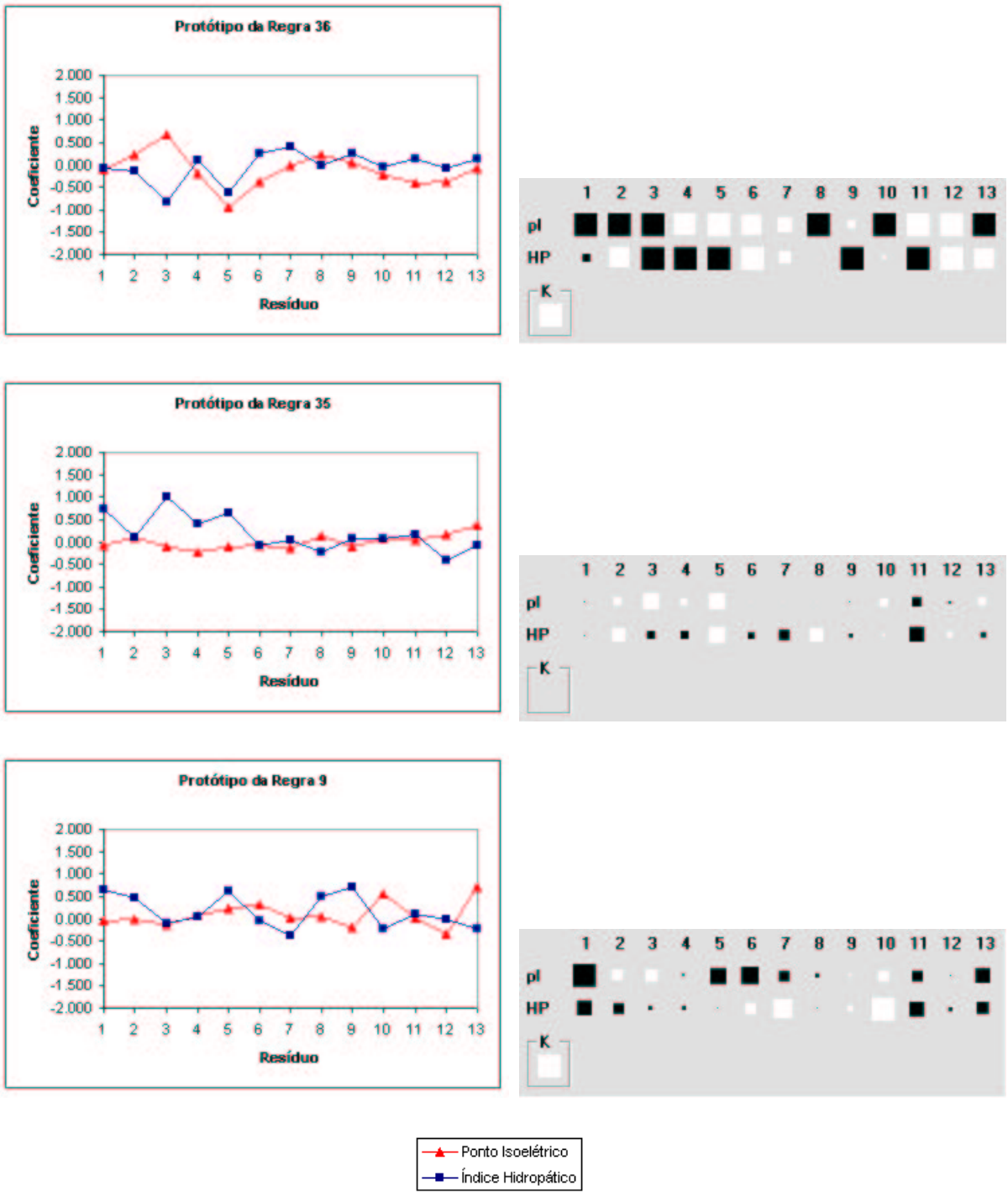

FIGURA 6.13 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 36, 35 e 9), por inexatidão, da classe *beta* O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação à regra 36, nota-se uma periodicidade do índice hidropático do lado direito do resído central. Quanto ao lado esquerdo, os valores alternam-se com variações significativas entre hidrofóbicos e hidrofílicos. Quanto ao pI, nota-se uma oscilação "suave" do lado direito da janela e uma mais "brusca" do lado esquerdo.

Com relação à regra 35, o pI apresenta-se com pequenas oscilações em torno do 0, na extensão da janela. Quanto ao HP, valores altamente hidrofóbicos são apresentados do lado esquerdo do resíduo central.

Com relação à regra 9, apresenta quase que uma totalidade de resíduos hidrofóbicos. O pI, por sua vez, também apresenta-se todo acima do 0, com picos de carga positiva na extremidade direita da janela.

De forma geral, a exemplo da classe alfa, nota-se que as tendências, possuem grandes especializações com relação a um ou mais resíduos da janela. Isto, provavelmente, indica que a rede conseguiu separar melhor os padrões e, por conseguinte, tendências mais distintas foram detectadas.

Visando corroborar o conhecimento apresentado pelo protótipo da regra 36, o padrão de mais alto grau de pertinência foi analisado na proteína correspondente. De toda a a seqüência primária, selecionou-se a parte que comtém a janela de 13 resíduos, conforme pode ser observado na figura 6.14. A seqüência foi marcada com a cor magenta e o resíduo predito como *beta* marcado em verde.

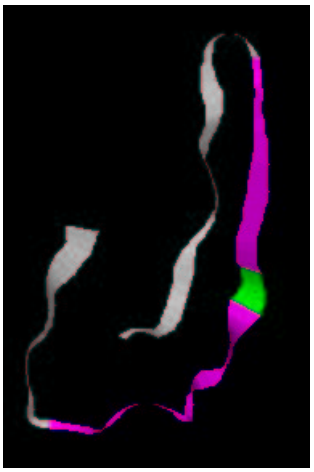

FIGURA 6.14 – Janela de 13 resíduos da proteína 1azu (*Azurina da Pseudomonas aeruginosa*) predita como tendo uma *beta* em sua seqüência

6.3.5 Regras da Classe Coil por Cobertura

Neste tópico, serão apresentadas as regras da classe *coil* com maior percentual de cobertura, isto é, que abrangem o maior número de padrões atendidos.

A tabela 6.6 apresenta as 3 regras com maior percentual de cobertura. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton para os coeficientes da equação linear podem ser vistos na figura 6.15.

TABELA 6.6 – Primeiras 3 regras com maior percentual de cobertura da classe *coil*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
6	22.48	35.97	0.950997
4	12.95	25.05	0.858929
8	6.51	24.34	0.859074

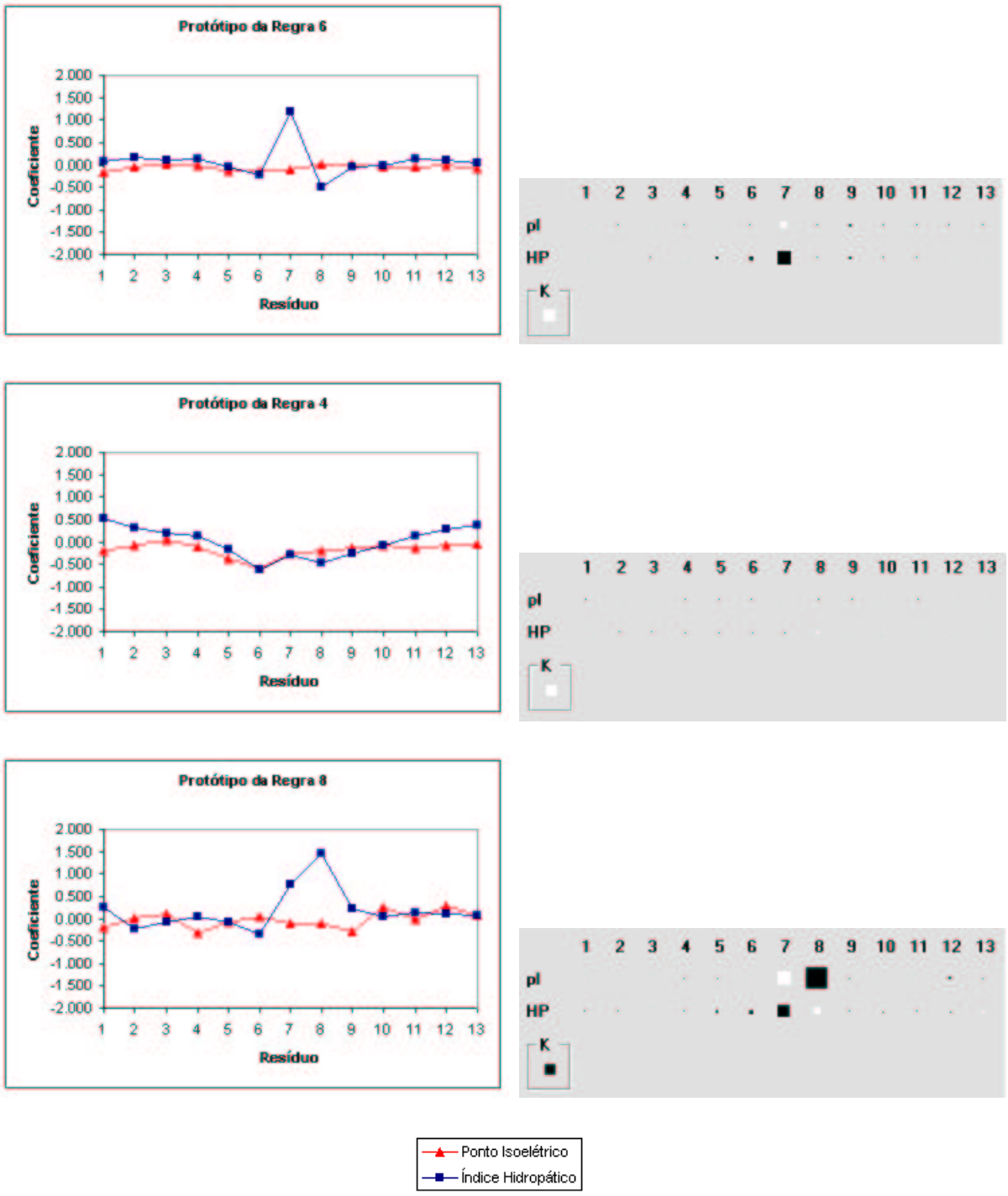

FIGURA 6.15 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 6, 8 e 4), por cobertura, da classe *coil* O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação à regra 6, observa-se um pico do índice hidropático no resíduo central. No resto da janela, em ambas as direções, após uma brusca queda para um valor levemente hidrofílico, seu comportamento se estabiliza perto do 0. Com relação ao pI, apresenta um comportamento de carga próximo de 0 durante toda a extensão da janela.

Com relação à regra 8, nota-se uma grande hidrofobicidade no resíduo central, com variações distintas nos resíduos imediatamente anterior e posterior. Após isto, o lado direito da janela assume um caráter levemente hidrofóbico e o esquerdo levemente hidrofílico. Quanto ao pI, nota-se uma periodicidade do lado esquerdo da janela.

Com relação à regra 4, apresenta um comportamento muito semelhante entre o pI e o HP. Especificament quanto ao HP, nota-se um desenho no formato de um W, a partir do resíduo central que tem um valor próximo de 0, seguido, de cada lado, por um valor negativo e numa ascenção constante em ambas as direções.

De forma geral, a exemplo do que já havia sido detectado na análise da determinação das classes anteriores, nota-se que as tendências não possuem grandes especializações com relação a um ou mais resíduos da janela.

6.3.6 Regras da Classe Coil por Inexatidão

Neste tópico, serão apresentadas as regras da classe *coil* com menor inexatidão, isto é, que possuem o menor erro de classificação.

A tabela 6.7 apresenta as 3 regras com menor erro. Para cada regra, o gráfico do protótipo correspondente e o diagrama de Hinton para os coeficientes da equação linear podem ser vistos na figura 6.16.

TABELA 6.7 – Primeiras 3 regras com menor erro da classe *coil*

Regra	Cobertura (%)	Inexatidão (%)	Erro RMS
51	0.43	0	0.870050
44	0.40	0	0.765612
45	0.39	0	0.187438

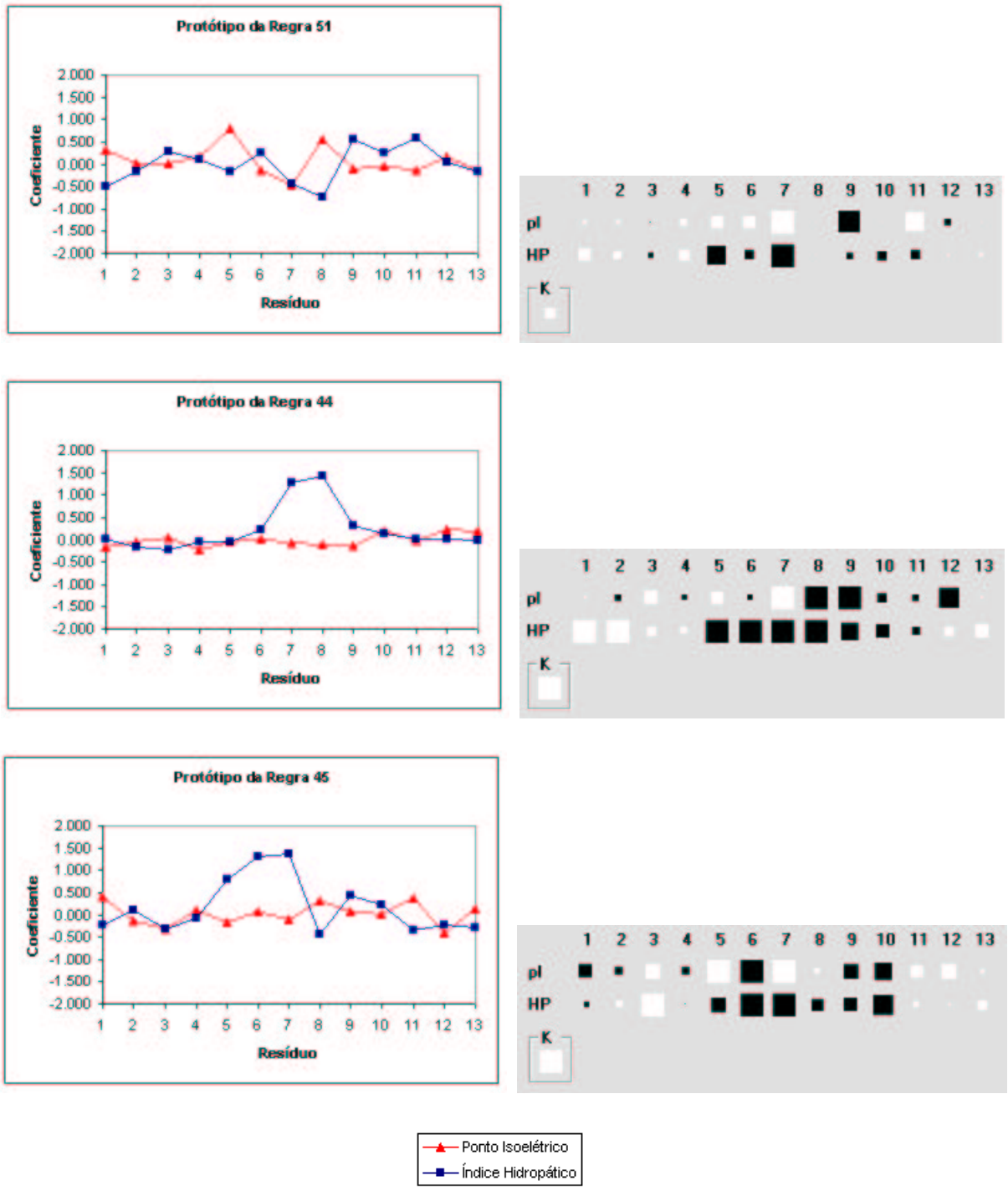

FIGURA 6.16 – Protótipo e diagrama de Hinton das primeiras 3 regras (*i.e.* 51, 44 e 45), por inexatidão, da classe *coil* O gráfico do protótipo apresenta, no eixo x, as posições do resíduo na janela e, no eixo y, o valor normalizado do ponto isoelétrico e do índice hidropático. O diagrama de Hinton apresenta as tendências na formação da estrutura secundária. Cada quadrado representa um coeficiente da equação linear que irá processar os padrões. A cor representa o sinal do coeficiente (brancos são positivos e pretos são negativos) e o tamanho representa o valor. Ao final, em separado, é feita a representação do termo independente k da equação.

Com relação à regra 51, observa-se um aumento no valor do pI, nos resíduos imediatamente ao lado do central. Após isto, um gradual descréscimo com estabilização perto do 0. Quanto ao HP, nota-se uma constante oscilação entre hidrofóbicos e hidrofílicos mas sem que se possa observar um padrão de periodicididade.

Com relação à regra 44, nota-se um comportamento praticamente estável do pI em torno do valor 0. O HP, por sua vez, apresenta altos valores de hidrofobicidade no resíduo central e a sua direita. Após isto, em ambas as direções, estabiliza-se como o pI.

Com relação à regra 45, nota-se um oscilação constante entre pIs negativos e positivos ao longo da janela. O HP apresenta, no resíduo central, um pico de hidrofobicidade. A esquerda da janela, um comportamento altamente hidrofóbico que vai decaindo até um nível que oscila entre levemente hidrofóbico e levemente hidrofílico.

De forma geral, a exemplo das classes anteriores, nota-se que as tendências, possuem grandes especializações com relação a um ou mais resíduos da janela. Isto, provavelmente, indica que a rede conseguiu separar melhor os padrões e, por conseguinte, tendências mais distintas foram detectadas.

Visando corroborar o conhecimento apresentado pelo protótipo da regra 51, o padrão de mais alto grau de pertinência foi analisado na proteína correspondente. De toda a a seqüência primária, selecionou-se a parte que comtém a janela de 13 resíduos, conforme pode ser observado na figura 6.17. A seqüência foi marcada com a cor magenta e o resíduo predito como *coil* marcado em amarelo.

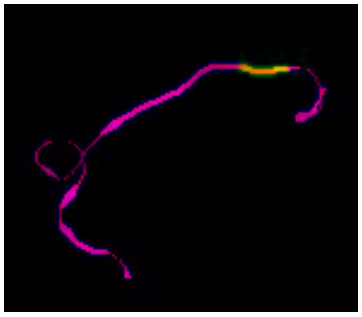

FIGURA 6.17 – Janela de 13 resíduos da proteína 1tgs (*inibidor da tripsina do pâncreas do boi*) predita como tendo uma *coil* em sua seqüência

Capítulo 7

Conclusão

Este estudo utilizou o problema da predição da estrutura secundária de proteínas como uma oportunidade para o conhecimento, prática e domínio de uma técnica de extração de conhecimento de RNAs.

Para viabilizar esta tarefa, uma série de procedimentos foram executados. Iniciando com a compreensão e o domínio da técnica de extração de regras de RNAs e do software de treinamento e extração. Passando por implementações de novas funcionalidades no software e pela definição de uma metodologia para extração de regras. Culminando com a implementação desta metodologia, aplicada ao problema em estudo e, por fim, com a apresentação e análise dos resultados obtidos.

Tendências no comportamento do processo de treinamento e extração de regras foram detectadas e apresentadas. Um maior número de experimentos necessita ser feito para um embasamento quantitativo destas tendências mas, indicativos sobre "o que" pesquisar foram fornecidos.

O conhecimento apresentado, na forma de regras, foi obtido e disponibilizado. Conforme pôde ser observado, as regras apresentaram comportamentos distintos, de acordo com o tipo de problema que a RNA tentou resolver. Fato este que era esperado, uma vez que representam diferentes regiões do espaço neural.

Devido às limitações de tempo intrínsecas a uma dissertação de mestrado, optou-se pela utilização de duas abordagens na interpretação das regras extraídas. Foram analisadas as regras com maior percentual de cobertura e as com menor erro na classificação. Consequentemente, o escopo da análise foi reduzido. Todas as regras deveriam ter sido analisadas pois o conhecimento pode se manifestar em qualquer uma das saídas geradas. Inclusive nas regras que estabelecam um equilíbrio entre os dois critérios adotados.

Concluiu-se que a implementação da metodologia de extração de regras atingiu o seu objetivo ao produzir resultados qualitativamente comparáveis aos já existentes e ao demonstrar um

conhecimento novo a ser investigado.

O objetivo de extrair conhecimento a respeito do comportamento dos atributos físico-químicos ao longo da cadeia de resíduos foi atingido. Nota-se que estes atributos tiveram um percentual de exatidão na classificação semelhante ao obtido com codificações ortogonais. O que nos leva a hipótese de que os mesmos podem ser utilizados como forma de representar o conhecimento adquirido pela rede mas que não possuem a capacidade de melhorar a exatidão da solução. Isto ficou claro quando analisados os atributos da Leucina e Isoleucina. A Leucina possui forte tendência na formação de *alfa*, enquanto que a Isoleucina induz a formação de *beta*. Porém, ao analisar a lista dos possíveis atributos físico-químicos disponiveis para representar estes dois resíduos, nos deparamos com valores muito próximos ou até mesmo semelhantes para ambos. Enquanto não for mapeado um, ou a combinação de mais de um, atributo físico-químico capaz de caracterizar de forma satisfatória as entradas, a codificação ortogonal ainda parece ser a melhor opção quando da utilização de RNAs.

Por fim, espera-se que o conhecimento disponibilizado seja analisado por profissionais de competência para tal, com a intenção de fomentar a discussão sobre a validade do mesmo bem como do processo pelo qual foi gerado. E, ainda, que o retorno da análise sirva para o aprimoramento do processo utilizado.

7.1 Trabalhos Futuros

Durante o desenvolvimento deste estudo, foram surgindo idéias sobre testes passíveis de serem analisados bem como sobre implementações que podem ser feitas. Porém, como não havia tempo disponível para testar todas as variações imaginadas, as mesmas foram sendo armazenadas. A seguir, segue a compilação destas idéias:

- avaliar a utilização de uma base de dados de seqüências não-homólogas com mais padrões, visando uma melhor exploração do espaço de entrada (*e.g.* a base do projeto EVA [56]);
- avaliar outros atributos físico-químicos, visando extração de conhecimento e, se possível, um acréscimo de exatidão na previsão (*e.g.* os atributos documentados em [11]);
- executar um processo de minimização de energia sobre as proteínas utilizadas no treinamento, visando comparar os resultados da RNA antes da minimização e após (*e.g.* utilizar o pacote de ferramentas denominado TINKER [69]);
- implementar e avaliar um tratamento para os padrões que foram desconsiderados devido

ao fato das suas regras não serem estatisticamente válidas, visando aproveitar a informação contida nos mesmos (*e.g.* criar uma regra que inclua a todos; ou submeter cada padrão à todas as regras estatisticamente validas, escolhendo a que melhor o atende e classifica-lo como pertencente a esta regra);

- implementar e avaliar, no FAGNIS, uma partição do neurônio em um número variável de regiões, visando uma melhor adequação à distribuição do espaço neural (*e.g.* uma partição em regiões lineares determinadas pela distribuição dos pontos; uma partição em maior número de regiões lineares mas em pontos fixos);

- implementar, no FAGNIS, o cálculo do intervalo de confiança da regra, visando fornecer o grau de certeza da regra extraída;

- testar e avaliar a extração de regras utilizando-se os ângulos diedros ϕ e ψ, visando informações mais precisas, sem a perda inerente à discretização ocorrida quando os mesmos são transformados em classes (*e.g.* testes com valores no intervalo de [-180; +180]; conversão para *seno* e *cosseno*).

Bibliografia

1 VOET, D.; VOET, J. G.; PRATT, C. W. *Fundamentos de Bioquímica*. Porto Alegre: ARTMED, 2000.

2 ANFINSEN, C. B. Principles that govern the folding of protein chains. *Science*, v. 181, p. 223–230, 1973.

3 MOUNT, D. *Bioinformatics: Sequence and Genome Analysis*. Woodbury, NY: Cold Springs Harbor Laboratory Press, 2001.

4 HOLBROOK, S. R.; MUSKAL, S. M.; KIN, S. Predicting protein structural features with artificial neural networks. In: HUNTER, L. (Ed.). *Artificial Intelligence and Molecular Biology*. [S.l.]: AAAI Press, 1993. p. 162–194.

5 CECHIN, A. L. *The Extraction of Fuzzy Rules from Neural Networks*. Tese — Tübingen, 1998.

6 TAKAGI, T.; SUGENO, M. Fuzzy identification on systems and its application to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 15, p. 116–132, 1985.

7 LEHNINGER, A. L. *Principles of Biochemistry*. 3. ed. New York: Worth, 2000.

8 SCHLICK, T. *Molecular Modeling and Simulation*. [S.l.]: Springer-Verlag, 2002.

9 WU, C. H.; MCLARTY, J. W. *Methods in Computational Biology and Biochemestry: Neural Networks and Genome Informatics*. [S.l.]: Elsevier, 2000.

10 KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. *J. Molec. Biol.*, v. 157, p. 105–132, 1982.

11 EXPASY - Expert Protein Analysis System. Disponível em: www.expasy.org/cgi-bin/protscale.pl. Acesso em: 15 mar. 2003.

12 STRYER, L. *Bioquímica*. 4. ed. [S.l.]: Koogan, 1996.

13 RAMACHANDRAN, G.; RAMAKRISHNAN, C.; SASISEKHARAN, V. Stereochemistry of polypeptide chain configurations. *J. Molec. Biol.*, n. 7, p. 95–99, 1963.

14 HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2. ed. [S.l.]: Bookman, 2001.

15 HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2001.

16 BALDI, P.; BRUNAK, S. *Bioinformatics: The Machine Learning Approach*. Cambridge, Massachusetts and London, England: The MIT Press, 2001.

17 MATTHEWS, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, v. 405, p. 442–451, 1975.

18 ANDREWS, R.; DIEDERICH, J.; TICKLE, A. A survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based Syst.*, v. 8, n. 6, p. 373–389, 1995.

19 TICKLE, A. B. et al. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks*, v. 9, n. 6, p. 1057–1068, november 1998.

20 ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965.

21 CHOU, P. Y.; FASMAN, G. Conformational parameters for amino acids in helical, beta-sheet and randon coil regions calculated from proteins. *Biochemistry*, v. 13, p. 211–222, 1974.

22 CHOU, P. Y.; FASMAN, G. Prediction of protein conformation. *Biochemistry*, v. 13, p. 222–245, 1974.

23 CHOU, P. Y.; FASMAN, G. Prediction of the secondary structure of proteins from their amino acid sequence. *Advanced Enzymology*, v. 47, p. 45–148, 1978.

24 YI, T.-M.; LANDER, E. S. Protein secondary structure prediction using nearest-neighbor methods. *J. Molec. Biol.*, v. 232, p. 1117–1129, 1993.

25 NISHIKAWA, K.; OOI, T. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim. Biophys. Acta*, v. 871, p. 45–54, 1986.

26 LEVIN, J.; ROBSON, B.; GARNIER, J. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, v. 205, p. 303–308, 1986.

27 ZHANG, X.; MESIROV, J.; WALTZ, D. Hybrid sistem for protein secondary structure prediction. *J. Molec. Biol.*, v. 225, p. 1049–1063, 1992.

28 SALZBERG, S.; COST, S. Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Molec. Biol.*, v. 227, p. 371–374, 1992.

29 SALAMOV, A.; SOLOVYEV, V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Molec. Biol.*, v. 247, p. 11–15, 1995.

30 SALAMOV, A.; SOLOVYEV, V. Protein secondary structure prediction using local alignments. *J. Molec. Biol.*, v. 268, p. 31–36, 1997.

31 LEVIN, J. Exploring the limits of nearest-neighbour secondary structure prediction. *Prot. Eng.*, v. 10, p. 771–776, 1997.

32 SOFTBERRY. Disponível em: http://www.softberry.com/berry.phtml?topic=protein. Acesso em: 13 abr. 2003.

33 GIBRAT, J.; GARNIER, J.; ROBSON, B. Further developments of protein secondary structure prediction using information theory. *J. Molec. Biol.*, v. 198, p. 425–443, 1987.

34 STOLORZ, P.; LAPEDES, A.; XIA, Y. Predicting protein secondary structure using neural net and statistical methods. *J. Molec. Biol.*, v. 225, p. 363–377, 1992.

35 RAYMER, M. L.; KUHN, L. A.; PUNCH, W. F. Knowledge discovery in biological datasets using a hybrid bayes classifier/evolutionary algorithm. In: *Proceedings of the IEEE 2nd Internation Simposium on Bioinformatics and Bioengineering*. [S.l.: s.n.], 2001.

36 ASAI, K.; HAYAMIZY, S.; HANDA, K. Prediction of protein secondary structure by the hidden markov model. *Bioinformatics*, v. 9, p. 141–146, 1993.

37 STULTZ, C.; WHITE, J.; SMITH, T. Structural analysis based on state-space modeling. *Prot. Sci.*, v. 2, p. 305–314, 1993.

38 FRANCESCO, V. D.; GARNIER, J.; MUNSON, P. Protein topology recognition from secondary structure sequences: Application of the hidden markov models to the alpha class proteins. *J. Molec. Biol.*, v. 267, n. 2, p. 446–463, Mar 1997.

39 FRANCESCO, V. D. et al. Fold recognition using predicted secondary structure sequences and hidden Markov models of proteins folds. *Proteins, Suppl*, v. 1, n. 1, p. 123–128, 1997.

40 HARGBO, J.; ELOFSSON, A. A study of hidden markov models that use predicted secondary structures for fold recognition. *Proteins: Struct. Funct. Genet.*, v. 36, n. 1, p. 68–87, 1999.

41 LIN, T.; WANG, G.; WANG, Y. Prediction of beta-turns in proteins using the first-order markov models. *J. Chem. Inf. Comp. Sci.*, v. 42, n. 1, p. 123–133, 2002.

42 HUA, S.; SUN, Z. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J. Molec. Biol.*, v. 308, p. 397–407, 2001.

43 QIAN, N.; SEJNOWSKI, T. Predicting the secondary structure of globular proteins using neural networks models. *J. Molec. Biol.*, v. 202, p. 865–884, 1988.

44 MACLIN, R.; SHAVLIK, J. Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, v. 11, p. 195–215, 1992.

45 TOWELL, G. G.; SHAVLIK, J. W.; NOORDENIER, M. O. Refinement of approximate domain theories by knowledge based neural network. In: *AAAI-90, Proceedings of the 8th National Conference on AI.* [s.n.], 1990. v. 2, p. 861–866. Disponível em: <citeseer.ist.psu.edu/towell90refinement.html">.

46 ROST, B.; SANDER, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Molec. Biol.*, v. 232, p. 584–599, 1993.

47 BALDI, P. et al. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, v. 15, n. 11, p. 937–946, 1999.

48 CUFF, J. A.; BARTON, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, v. 34, n. 4, p. 508–19, Mar 1999.

49 POLLASTRI, G. et al. Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. *Proteins*, v. 47, n. 2, p. 228–235, 2001.

50 RIIS, S.; KROGH, A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol*, v. 3, p. 163–183, 1996.

51 RIIS, S.; KROGH, A. *Prediction of beta sheets in proteins*. Boston,MA: MIT Press, 1996. 917-923 p.

52 ROST, B. Learning from evolution to predict protein structure. In: *BCEC97: Proceedings of Sweden Bio-Computing and Emergent Computation*. [S.l.: s.n.], 1997. p. 87–101.

53 ZHU, H.; YOSHIHARA, I.; YAMAMORI, K. Prediction of protein secondary structure by multi-modal neural networks. In: *IJCNN02: Proceedings of the International Conference on Neural Networks*. [S.l.: s.n.], 2002. v. 1, p. 280–285.

54 LAMONT, O.; LIANG, H. H.; BELLGARD, M. Data representation influences protein secondary structure prediction using artificial neural networks. In: *Proceedings of the 7th Australian and New Zeland Intelligent Information Systems Conference*. [S.l.: s.n.], 2001. p. 411–416.

55 SSPRO. Disponível em: http://promoter.ics.uci.edu/BRNN-PRED/. Acesso em: 13 abr. 2003.

56 EVA - Evaluation Large Scale Project. Disponível em: http://maple.bioc.columbia.edu/eva. Acesso em: 15 mar. 2003.

57 HAYASHI, Y.; NAKAI, M. Automated extraction of fuzzy if-then rules using neural networks. *Transactions of IEE of Japan*, v. 110, p. 198–206, 1990.

58 TAKAGI, H.; HAYASHI, I. NN-driven fuzzy reasoning. *International Journal of Approximate Reasoning*, v. 5, p. 191–212, 1991.

59 HAYASHI, I. et al. Construction of fuzzy inference rules by ndf and ndfl. *International Journal of Approximate Reasoning*, v. 6, p. 241–266, 1992.

60 POECHMUELLER, W. et al. Rbf and cbf neural network learning procedures. In: *Proceedings of the International Conference on Neural Networks*. Orlando, USA: [s.n.], 1994. p. 407–412.

61 GUJARATI, D. N. *Basic Econometrics*. 3. ed. [S.l.]: McGraw-Hill Inc, 2000.

62 ROST, B. Review: Protein secondary structure prediction continues to rise. *J. Structural Biol.*, v. 134, p. 204–218, 2001.

63 PDB - Protein Data Bank. Disponível em: http://www.rcsb.org/pdb/. Acesso em: 15 mar. 2003.

64 DSSP - Define Secondary Structure of Proteins. Disponível em: http://www.cmbi.kun.nl/gv/dssp/. Acesso em: 27 mar. 2003.

65 KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, v. 22, n. 12, p. 2577–2637, 1983.

66 RIIS, S. K. Combining neural networks for protein secondary structure prediction. In: *Proceedings of the IEEE Internation Conference on Neural Networks*. [S.l.: s.n.], 1995. v. 4, p. 1744–1748.

67 RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: *Proc. of the IEEE Intl. Conf. on Neural Networks*. San Francisco, CA: [s.n.], 1993. p. 586–591.

68 HINTON, G. E.; SEJNOWSKI, T. J.; ACKLEY, D. H. *Boltzmann Machines: Constraint Satisfaction Networks that Learn*. [S.l.], may 1984.

69 TINKER. Disponível em: http://dasher.wustl.edu/tinker/. Acesso em: 10 dez. 2003.

Apêndice A

Base QS106

TABELA A.1 – Relação de proteínas definida por Qian & Sejnowski e utilizadas neste estudo

Código da Proteína	Código Atualizado	Comentário(s)
1abp	1abe	
1acx		
1apr	2apr	
1aza	2aza	
1azu		
1bp2		
1cac	1ca2	
1cc5		
1ccr		
1cpv	5cpv	
1crn		
1ctx		
1cy3	2cy3	
1cyc		
1ecd		
1est		
1fc2		
1fdh		
1fdx	1dur	
1fxl		
1gcn		
1gcr	4gcr	
1gf1		
1gf2		
1gp1		
1hds		
1hip		
1hmq	2hmq	

Código da Proteína	Código Atualizado	Comentário(s)
1ig2	2ig2	
1ige		não encontrada no PDB
1ins	4ins	
1ldx	2ldx	
1lzl		
1lzm	2lzm	
1lzt		
1mbd		
1mbs		
1mlt		
1nxb		
1p2p		
1pfc		
1ppd		
1ppt		
1pyp		
1rei		
1rhd		
1rn3	3rn3	
1sn3	2sn3	
1tim		
1tgs		
2act		
2adk	3adk	
2alp		
2ape	4ape	
2app	3app	
2b5c	1cyo	
2cab		
2ccy		
2cdv		
2cyp		
2dhb		
2fd1	5fd1	
2gch		
2gn5		
2grs	3grs	
2icb	3icb	
2kai		
2lh1		
2lhb		
2mcp		
2mdh	4mdh	
2mt2		obsoleta
2pab		
2rh		

Código da Proteína	Código Atualizado	Comentário(s)
2sbt		
2sga		
2sns		
2sod		
2ssi	3ssi	
2stv		
2taa		
2tbv		
3c2c		
3cna		
3fxc	4fxc	
3gpd		
3hhb		
3pcy		
3pgk		
3pgm		
3rp2		
3sgb		
3tln	8tln	
451c		
4cts		
4dfr		
4fxn	2fox	
4sbv		
5atc	5at1	
5cpa		
5ldh		
5pti		
5rxn		
6adh		
6api	8api	
8cat		