CLASSIFICAÇÃO DE PROTEÍNAS COM REDES NEURAIS ARTIFICIAIS

Tiago Amador Coelho
UFLA – Universidade Federal de Lavras
Cx Postal 3037 – CEP 37200-000 Lavras (MG)
tcoelho@comp.ufla.br

Resumo: Pelo fato de ainda existirem seqüências sem classificação nos bancos de dados públicos, os métodos tradicionais de classificação de seqüências se mostram deficientes. O objetivo do presente trabalho é implementar um esquema de codificação de seqüências de aminoácidos a fim de se construir um classificador de proteínas baseado em Redes Neurais Artificiais utilizando os vetores resultantes da codificação implementada, de modo a ser um complemento aos métodos tradicionais de classificação.

Palavras Chaves: classificação de proteínas, redes neurais artificiais, COG

PROTEIN CLASSIFICATION WITH NEURAL NETWORKS

Abstract: For the fact that there are still unrated sequences in public data banks, traditional methods for classifying sequences are working poor. The objective of this work is to implement a scheme of encoding sequences of amino acids in order to build a classifier of proteins based on Artificial Neural Networks using the vectors resulting from the consolidation implemented in order to be a complement to traditional methods of classification.

Key words: protein classification, neural networks, COG

1 Instrodução

No início da década de 80 com o desenvolvimento de técnicas relativamente rápidas para o seqüenciamento do DNA, onde houve um aumento no número de seqüências (Figura 1). Essas seqüências de nucleotídeos e aminoácidos foram armazenados em alguns bancos de dados.

Os principais bancos de dados são GenBank, EMBL-Bank(European Molecular Biology Laboratory – Nucleotide Sequence Database), COG (Cluster of Orthologous Groups), GO (Gene Ontology), e o DDBJ(DNA Data Bank of Japan).

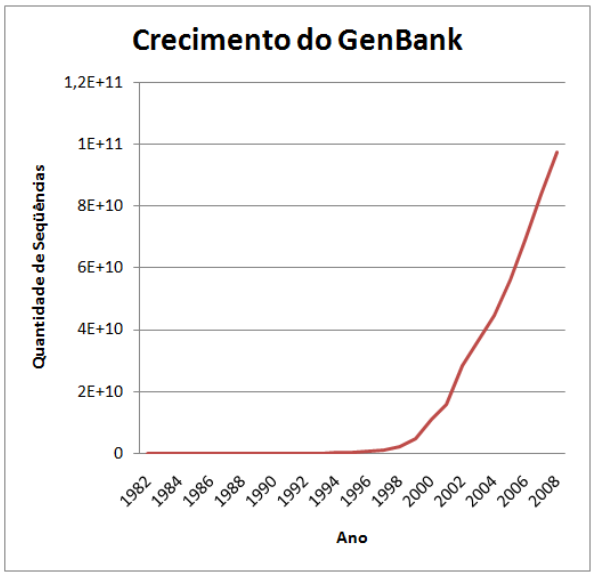

Figura 1 - Crescimento do GenBank 1982 - 2008

Inúmeros bancos de dados e métodos computacionais de acesso público (na sua grande maioria) ou privados, foram e estão sendo criados na tentativa de organizar e permitir acesso eficiente e rápido às informações geradas pelos projetos de larga escala, bem como a análise comparativa dessa quantidade maciça de dados (CASTANHO, 2005). A

criação e manutenção de banco de dados biológicos são por si só um desafio, devido não só à imensa quantidade de dados, mas sobretudo à dificuldade de desenvolver esquemas e estruturas que representem de forma exata ou bastante aproximada a complexa relação existente entre os diversos componentes dos sistemas biológicos (CASTANHO, 2005).

A partir do seqüenciamento de um genoma, a geração de dados tem como objetivo, a predição do conjunto de proteínas existentes no organismo em questão e a funcionalidade que cada proteína desempenha, para melhor entender o funcionamento do organismo (RODRIGUES, 2007). Existem dois métodos que podem ser seguidos: o laboratorial e o computacional. O primeiro método é o mais confiável, onde serão realizados testes em laboratório para a predição do conjunto de proteínas, entretanto é um processo muito dispendioso e demorado. O segundo método, é indicado para um grande quantidade de seqüências, já que é pretendido um resultado em um prazo de tempo menor e com um certo grau de confiabilidade.

A comparação de seqüências é a mais fundamental operação de análise de proteínas, indicando a similaridade entre elas, pode-se sugerir relações envolvendo estrutura, função e evolução, sendo essas proteínas originárias de um mesmo ancestral comum.

As proteínas são atualmente classificadas de acordo com a ocorrência de padrões conservados de aminoácidos que definem os domínios (Rodrigues, 2007).mpletos. Cada COG consiste de uma proteína individual ou de um grupo “paralogs” de pelo menos 3 genomas que possuem funções conservadas ao longo da evolução. (TATUSOV, 1997)

Atualmente ainda existem um grande número de seqüências não classificadas nestes bancos de dados,sendo importante a sua classificação. Adicionalmente, seqüências anotadas em uma classe podem ter sua classificação modificada pelo fato de um novo domínio, presente na proteína, ter sido identificado recentemente.

2 Banco de Dados GOG

O banco de dados COG (Cluster of Orthologous Groups) (TATUSOV, 2000) disponibilizado por *National Center for Biotechnology Information* (NCBI), compreende grupos de proteínas preditas codificados por genomas procarióticos e mais recentemente também eucarióticos, cujos genomas foram integramente seqüenciados. Ele representa uma tentativa de classificação filogenética destas proteínas, caracterizando-se como uma fonte de informação em Genômica Funcional e Evolutiva. Através de inúmeras páginas navegáveis o usuário tem acesso a diversos dados pré-computados, como por exemplo, os padrões filogenéticos, as classificações funcionais, e listas de grupos de genes ortólogos (COGs) por categoria funcional ou por via metabólica e co-ocorrência de genomas em GOGs (ALMEIDA, 2007).

O banco de dados original continha proteínas de cinco genomas de bactérias, um de arquea e um de eucarioto, constituindo 720 COGs. Em seguida, um sexto genoma bacteriano foi adicionado, aumentando o número de COGs para 860. O estado atual do Banco de dados COG consiste em 2091 COGs e inclui proteínas de 21 genomas completos (TATUSOV, 1997).

Sua ferramenta de busca mais importante é o COGNITOR, programa através do qual se determina a qual (ou quais) GOG(s) pertence uma nova seqüência protéica. Suas principais limitações são: mecanismos rígidos de busca e obtenção de dados, os quais são baseados em resultados pré-computados sendo acessíveis somente através de tediosa navegação; dados bastantes desatualizados em relação ao número de genomas integralmente seqüenciados; impossibilidade de realização de consultas maciças e/ou complexas seja por nome ou número do COG, categoria funcional, espécie ou por comparação de seqüência, através do COGNITOR (ALMEIDA, 2007).

3 Proteínas

As proteínas são as moléculas orgânicas mais abundantes e importantes nas células e perfazem 50% ou mais de seu peso seco. São encontradas em todas as partes de todas as células, uma vez que são fundamentais sob todos os aspectos da estrutura e função celulares. Existem muitas espécies diferentes de proteínas, cada uma especializada para uma função biológica diversa. Além disso, a maior parte da informação genética é expressa pelas proteínas. (MARZZOCO, 2007)

Pertencem à classe dos peptídeos, pois são formadas por aminoácidos ligados entre si por ligações peptídicas. Uma ligação peptídica é a união do grupo amino (-NH 2) de um aminoácido com o grupo carboxila (-COOH) de outro aminoácido, através da formação de uma amida.

São os constituintes básicos da vida: tanto que seu nome deriva da palavra grega "proteios", que significa "em primeiro lugar". Nos animais, as proteínas correspondem a cerca de 80% do peso dos músculos desidratados, cerca de 70% da pele e 90% do sangue

seco. Mesmo nos vegetais as proteínas estão presentes. (CHAMPE, 2006)

Segundo Marzzoco (2007), a importância das proteínas, está relacionada com suas funções no organismo, e não com sua quantidade. Todas as enzimas conhecidas, por exemplo, são proteínas; muitas vezes, as enzimas existem em porções muito pequenas. Mesmo assim, estas substâncias catalisam todas as reações metabólicas e capacitam aos organismos a construção de outras moléculas - proteínas, ácidos nucléicos, carboidratos e lipídios - que são necessárias para a vida.

3.1 Composição

Todas as proteínas, independentemente de sua função ou espécie de origem, são formadas a partir de um conjunto básico de vinte aminoácidos, arranjados em várias seqüências específicas (PASQUIER, 1999).

3.2 Estrutura Primária

Refere-se ao número e identidade dos aminoácidos que compõem a molécula e ao ordenamento ou seqüência dessas unidades na cadeia polipeptídica. A união peptídica somente permite a formação de estruturas lineares e por isso, as cadeias não apresentam ramificações, como na Figura 2.

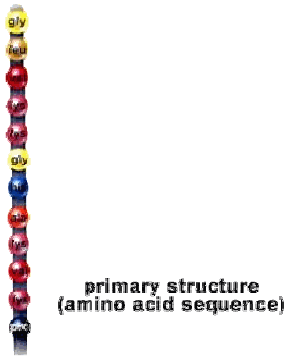

Figura 2 - Estrutura primária da proteína

3.3 Estrutura Secundária

Segundo Nelson (2006), a medida que o comprimento das cadeias vai aumentando e em função das condições físico-químicas do meio, se cria a estrutura secundária (Figura 3), que é a disposição espacial regular, repetitiva, que a cadeia polipeptídica pode adotar, geralmente mantida por ligações de hidrogênio. Podemos ter:

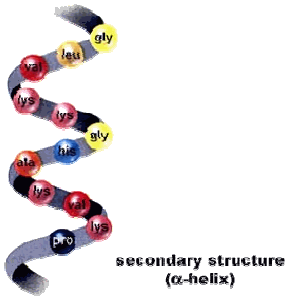

Figura 3 - Estrutura secundária da proteína (hélice - α)

3.4 Estrutura Terciária

É a estrutura da maioria das proteínas globulares, aparece a partir das hélices, que voltam a enrolar-se. É uma estrutura tridimensional (Figura 4) completa que forma-se a partir das forças de atração ou repulsão eletrostática, das pontes de hidrogênio, das forças de Van der Waals e das pontes dissulfeto existentes entre os resíduos de aminoácidos que formam as cadeias. (NELSON, 2006)

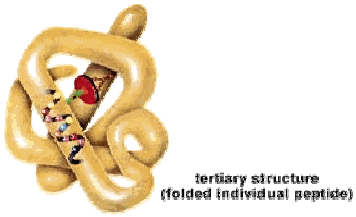

Figura 4 - Estrutura terciária da proteína

3.5 Estrutura Quaternária

São estruturas de caráter oligomérico, que estão compostas por várias moléculas separadas, mas entrelaçadas em estrutura terciária (Figura 5). Se aplica somente a proteínas constituídas por duas ou mais cadeias polipeptídicas e se refere a disposição espacial dessas cadeias e as ligações que se estabelecem entre elas – pontes de hidrogênio, atrações eletrostáticas, interações hidrofóbicas, pontes dissulfeto entre cisteínas de cadeias diferentes. Um exemplo deste tipo de estrutura é a hemoglobina que é composta por quatro subunidades semelhantes à mioglobina. (NELSON, 2006)

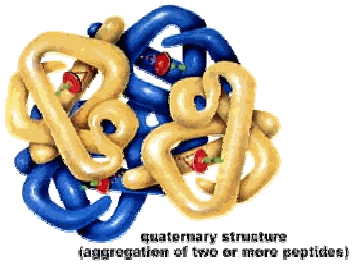

Figura 5 - Estrutura quaternária da proteína

4 Redes Neurais Artificiais

4.1 Neurônio Artificial

Redes Neurais Artificiais (RNA) é um modelo baseado na natureza, mais especificamente no cérebro humano (BARRETO, 2002). Assim como o sistema nervoso é composto por bilhões de células nervosas, a rede neural artificial também seria formada por unidades que nada mais são que pequenos módulos que simulam o funcionamento de um neurônio. Estes módulos devem funcionar de acordo com os elementos em que foram inspirados, recebendo e retransmitindo informações.

Segundo Lemes (2000), o fisiologista Warrem S. MacCulloch e Walter Pitts desenvolveram um modelo que propõe elementos computacionais, introduzindo assim a principal referencia da teoria de Redes Neurais Artificiais. O modelo proposto é bem simples comparando-o com um neurônio biológico, que possui uma complexa estrutura e grande número de detalhes.

O neurônio McCulloch-Pitts é um dispositivo binário, a sua saída representa um dos dois estados possíveis, ou seja, 0 ou 1. Como segue na Figura 6:

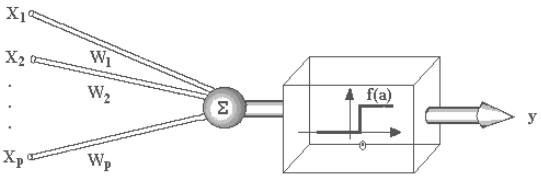

Figura 6 - Modelo de McCulloch e Pitts

As entradas (X_i) do neurônio é em binário, e possuem ganhos (W_i), que podem ser excitatórias ou inibitórias.

$$\Phi = \sum_{i=1}^{n} X_i W_i$$

O resultado da entrada do neurônio serve de argumento para a função de ativação, para dar a resposta do neurônio.

No modelo geral de neurônio, sendo uma generalização do modelo de McCulloch e Pitts, as entradas XiWi são combinadas usando uma função Φ, para produzir um estado de ativação do neurônio que através da função de ativação η (Figura 7). Um valor auxiliar θ (bias) é utilizado para representar a polarização (BARRETO, 2002).

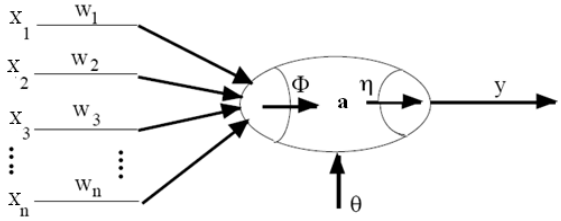

Figura 7 - Esquema de um Neurônio Artificial

4.2 Funções de Ativação Limiar (Degrau)

Utilizada no modelo de McCulloch e Pits, função que modela a característica "tudo-ou-nada" (MENDES, 2008). A função limiar é descrita da seguinte forma:

$$f(a) = \begin{cases} 1, se\ a \geq 0; \\ 0, se\ a \leq 0; \end{cases}$$

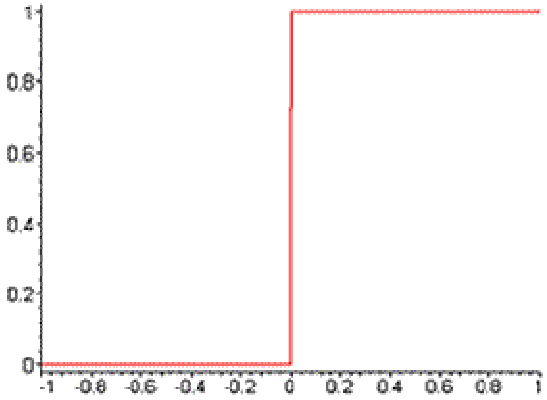

Figura 8 - Gráfico da Função de Limiar

4.3 Treinamento

Para que uma Rede Neural Artificial possa fornecer resultados convenientes, é necessário que passe por uma fase de treinamento.

A fase de aprendizagem consiste em um processo interativo de ajuste de parâmetros da rede, os pesos das conexões entre as unidades de processamento, que guardam ao final do processo, o conhecimento que a

rede adquiriu do ambiente em que está operando. Um fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente, durante a aprendizagem (Figura 9) (BARRETO, 2002).

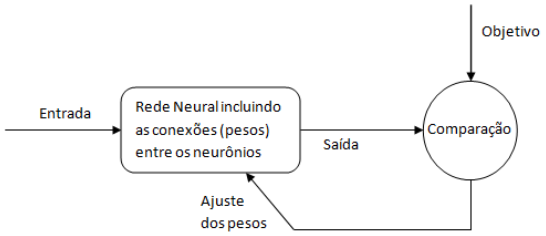

Figura 9 - Treinamento de uma rede neural

Nesse contexto existem os seguintes paradigmas de aprendizado:

- Aprendizado Supervisionado, quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;
- Aprendizado Não Supervisionado (auto-organização), quando não existe uma agente externo indicando a resposta desejada para os padrões de entrada;

A Regra Delta e o *Backpropagation* são exemplos de algoritmos supervisionados. Para o algoritmo não supervisionado, somente os padrões de entrada estão disponíveis na rede, ao contrário do algoritmo supervisionado, cujo o conjunto de treinamento possui pares de entrada e saída. No aprendizado supervisionado, a medida de desempenho é baseada no conjunto de respostas desejadas usando um critério de erro conhecido. No aprendizado por reforço (um caos particular do aprendizado supervisionado) a única informação de realimentação fornecida a rede é se uma determinada saída está correta ou não.

Denomina-se ciclo uma apresentação de todos os N pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos num ciclo pode ser executado de dois modos:

- Modo Padrão, a correção dos pesos acontece a cada apresentação, à rede, de um exemplo do conjunto de treinamento.
- Modo Ciclo, apenas uma correção é feita por ciclo.

4.4 Perceptron

Em 1958 Rosemblatt criou o modelo Perceptron, no qual os neurônios eram organizados em camadas de entrada e saída (Figura 10), onde os pesos das conexões eram adaptados durante treinamento afim de atingir a eficiência sináptica. (LEMES, 2000).

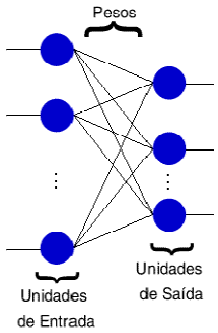

Figura 10 - Modelo do Perceptron

A limitação desta Rede Neural se encontra na reduzida gama de problemas que consegue tratar: classificação de conjuntos linearmente separáveis (Figura 11).

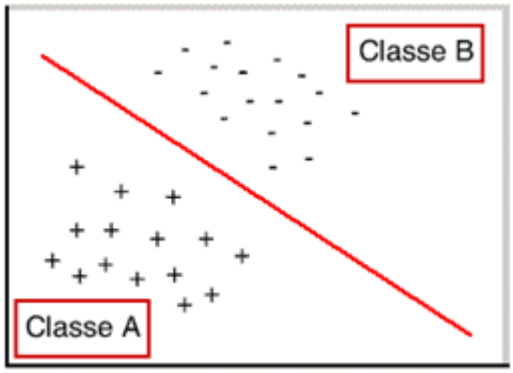

Figura 11 - Problema Linearmente Separável

4.5 Multilayer Perceptron

A forma de arranjar os perceptrons em camadas é denominado Multilayer Perceptron. O multilayer perceptron foi concebido para resolver problemas mais complexos, os quais não poderiam ser resolvidos pelo modelo de neurônio básico, problemas linearmente não separáveis. Os neurônios internos são de suma importância na rede neural, pois provou-se que sem estes torna-se impossível a resolução de problemas linearmente não separáveis (Figura 12) (BISHOP,2005).

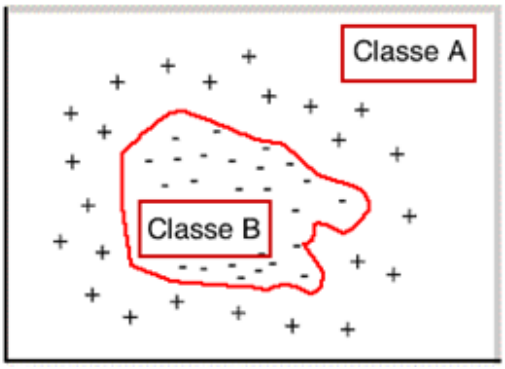

Figura 12 - Problema Linearmente não Separável

Usualmente as camadas são classificadas em três grupos:

- Camada de Entrada: onde os padrões são apresentados à rede;
- Camadas Intermediárias ou Ocultas: onde é feita a maior parte do processamento, através das conexões ponderadas, podem ser consideradas como extratoras de características;
- Camada de Saída: onde o resultado final é concluído e apresentado.

A Figura 13 mostra uma RNA com uma camada de entrada, duas camadas escondidas e uma camada de saída.

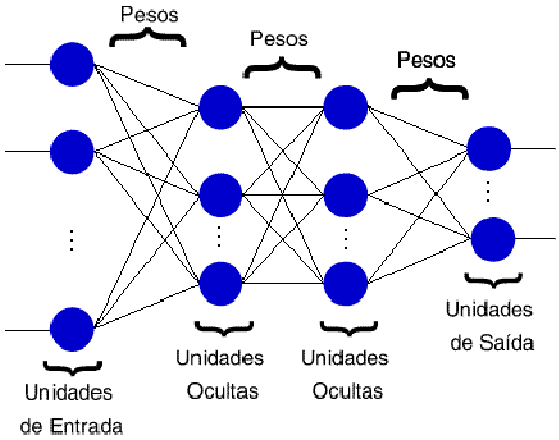

Figura 13 - Organização em Camadas

4.6 Algumas Aplicações de RNA em Bioinformática

A utilização de Redes Neurais Artificiais em alguns problemas já vem sendo aplicada em alguns problemas como:

- Reconhecimento de Sinais(Promotores, *Start Codon, Stop Codon*);
- Identificação de Assinaturas;
- Identificação de Repetições e de Regiões de Baixa Complexidade;
- Similaridade entre Seqüências;
- Anásile de Cromatogramas;
- Análise de experimentos com expressões de genes;
- Predição de estrutura secundárias de proteínas;
- Análise de regiões extra-gênicas em DNA;
- Extração de relações entre elementos de uma seqüência.

(MENDES, 2008).

4.7 Problema na Aplicação das RNA à Bioinformática

Normalmente métodos computacionais utilizam seqüências sem sua estrutura primária como entrada de dados. É fácil perceber que a quantidade de aminoácidos em um conjunto de seqüências protéicas não é o mesmo, resultando em uma diferença de dimensionalidade entre os dados. A Figura 14 mostra a quantidade de aminoácidos de todas as 6254 proteínas da bactéria *Acaryochloris marina.*

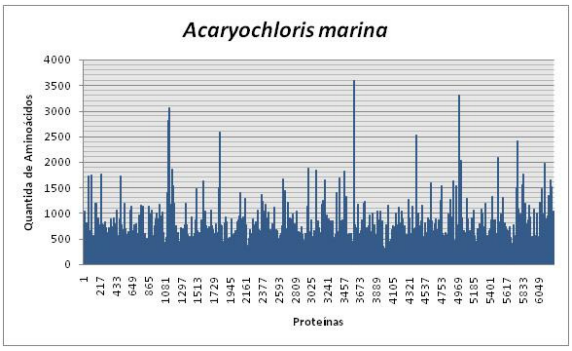

Figura 14 - diferença de dimensionabilidade entre as Proteínas do *Acaryochloris marina*

Como mostrado na seção 4.1 as entradas de uma RNA devem possuir valores numéricos e mesma dimensão, logo uma metodologia de codificação que converta seqüências de caracteres de diferentes dimensões em vetores numéricos de mesma dimensão é necessária. O método de codificação de seqüências conhecido como *Sequence Coding By Sliding Window* (SCSW) (RODRIGUES, 2007) pode ser aplicado para solucionar o problema em questão.

5 *Sequence Coding by Sliding Window* (SCSW)

Em 1986, Blaisdell propôs uma codificação que resolvia o problema da diferença de dimensionalidade,

convertendo seqüências de dimensões diferentes em vetores de mesma dimensão. (RODRIGUES, 2007).

Funcionamento da codificação:

- Dado uma seqüência S de tamanho N definida sobre um alfabeto α;
- Uma janela deslizante W_n de tamanho $1 \leq n \leq N$ é posicionada na posição 1 da seqüência S e vai sendo deslocada até a posição N – n + 1;
- Um vetor V_n de dimensão α^n é definido, onde cada posição corresponde a uma possível n – tupla dos elementos de α;
- A cada deslocamento de W_n em S a posição de V_n correspondente à n – tupla encontrada é incrementada de 1;
- Após W_n atingir a posição N – n + 1 em S, o vetor V_n conterá a quantidade de cada n – tupla da seqüência percorrida e, independentemente do tamanho da seqüência, o vetor V_n terá dimensão α^n.
- (BLAISDELL, 1986)

Para manter um padrão de nomenclatura, a codificação será denominada *Sequence Coding by Sliding Window* (SCSW) (RODRIGUES, 2007).

6 Metodologia

6.1 Obtenção dos Dados

A obtenção dos dados a serem utilizados neste trabalho foi realizada em duas etapas.

Na primeira foi selecionado o conjunto de proteínas dos organismos *Mycoplasma hyopneumoniae* e *Acaryochloris marina* disponibilizado em Genome Project . Todas as proteínas estão disponíveis no formato FASTA (Figura 15).

```
>gi|72080389|ref|YP_287447.1| F0F1 ATP synthase subunit A [Mycoplasma
hyopneumoniae 7448]
MMDFFRDWNQPQLFTLFILVFLVIILSIIIFFHIKKAKIDESPSAVVLFAESYLIFIDDL
VETAGEGYINKVKPYIFSLFTFFLLGNLLSLVGLEPISTSISVTLSLAFVSWFGIFVVG
AIYSRWKYLSEFAKNPLKIIGIPAPLISLSFRMVGNLISGSVLLLIIYSGVQWIYQKIP
LGTFGNFNLPIVLIFPPFLIYFDIVGSLIQSFIFVILTTSYWGMEVNQDEARLKINKKQ
LNLQKI
>gi|72080390|ref|YP_287448.1| F0F1 ATP synthase subunit C [Mycoplasma
hyopneumoniae 7448]
MNSIVNFSQQLIQNFQEVSQKTVADSSNLKAFAYLGAGLAMIGVIGVGAGQGYAA
GKACDAIARNPEAQKQVFRVLVIGTAISETSSIYALLVALILIFVG
```

Figura 15 - Seqüência gi: 72080351 *do Mycoplasma hyopneumoniae* 7448

Na segunda etapa o conjunto de proteínas foi separado de acordo com as classes funcionais do COG.

6.2 Redução do Alfabeto

A fim de diminuir o custo computacional durante o treinamento das RNAs, foi utilizado o denominado *Exchange group* (Wu et al., 1992) baseado na matriz de similaridade PAM (DAYHOFF,1978)., O alfabeto original de tamanho 20 foi reduzido para um alfabeto de tamanho 6, de acordo com a Tabela 1

Tabela 1 - *Exchange group*, redução do alfabeto tamanho 20 para 6

HRK	H
C	C
FYW	F
DNQE	D
STPAG	S
MILV	M

6.3 Codificação

Assim como o conjunto de proteínas da bactéria *Acaryochloris marina* (Figura 14), as proteínas da bactéria *Mycoplasma hyopneumoniae* possuem diferença de dimensionalidade quando é utilizada sua estrutura primária, Figura 26.

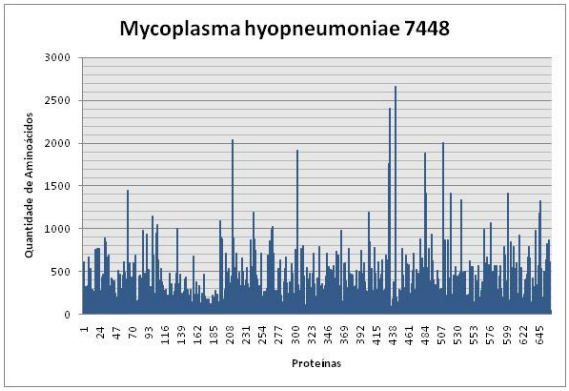

Figura 16 - Diferença de dimensionabilidade entre as proteínas do Mycoplasma hyopneumoniae

O método de codificação de seqüências Sequence Coding By Sliding Window (SCSW) foi implementado para extrair informações das seqüências, utilizando janela de tamanho 2 e o alfabeto reduzido para gerar vetores de dimensão 36.

6.4 Aplicação da Metodologia *One-Against-All* e Seleção dos Pontos da Margem de Separação

Para o treinamento e validação das RNAs os dados foram separados de modo que cada RNA seja um classificador para cada classe funcional do COG. Portanto cada classe será classificada contra todas as outras, *one-against-all.*

É fácil perceber que existe um desbalanceamento entre as classes funcionais do COG correspondentes a cada bactéria. Quando se aplica a metodologia *one-against-all* o desbalanceamento fica ainda mais evidente podendo fazer com a RNA fique tendenciosa durante o seu treinamento. Para realizar o balanceamento entre as classes, a metodologia *Condensed Nearest Neighbor* (CNN) foi aplicada onde a distância euclidiana foi utilizada como medida de distância entre os pontos.

6.5 Construção, Treinamento e Validação das Redes Neurais Artificiais

A construção, treinamento e validação das Redes Neurais Artificiais foram realizadas com o *ToolBox* de Redes Neurais Artificiais do Matlab 6.

A Figura 17 demonstra a topologia da RNA para cada classificador, onde na camada de entrada tinha 36 neurônios, na camada escondida a quantidade neurônios variava para cada classificador e na camada de saída existia apenas 1 neurônio.

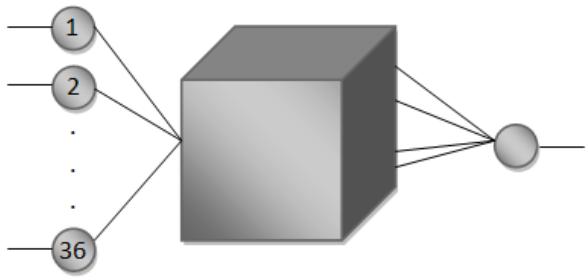

Figura 17 - Topologia da RNA utilizada

O treinamento foi realizado com o conjunto de proteínas da bactéria *Mycoplasma hyopneumoniae* e a validação com as proteínas da bactéria *Acaryochloris marina*.

O algoritmo de treinamento utilizado foi o *Backpropagation* com somente um nodo na camada escondida, durante 100 iterações. A classificação foi realizada com o limiar de 0,6, ou seja, maior que 0,6 indica pertinência à classe, menor que -0,6 indica pertinência à outra classe e caso contrário resultado indefinido.

7 – Resultados e Discussão

As seqüências utilizadas no teste das RNAs foram codificadas utilizando os mesmos parâmetros dos dados de treinamento a fim de manter a compatibilidade de dimensão desses dados.

Os resultados são mostrados na Tabela 2. Nesta tabela são representados os classificadores (colunas) e o resultado da classificação de acordo com a entrada recebida (linhas):

- Pertencente a classe: porcentagem de acerto que o classificador obteve com os dados de entrada pertencem à mesma classe do classificador.
- Não pertencente à classe: porcentagem de acerto que o classificador obteve com os dados de entrada pertencentes a classes diferentes à do classificador.

Tabela 2 - Tabela das Taxas de Acertos dos Classificares de Acordo com os Dados de Entrada

Dados de Entrada	Classificares								
	C	D	E	F	G	H	I	J	K
Pertencente a classe (acerto em %)	85,65	57,50	0,00	13,75	0,00	45,14	71,68	28,33	10,61
Pertencente a classe (erro em %)	5,65	20,00	69,29	72,50	100	46,29	21,24	28,33	74,92
Não pertencente a classe (acerto em %)	12,15	9,00	66,92	68,46	95,86	53,25	24,40	58,73	81,75
	L	M	O	P	R	S	T	U	V
Pertencente a classe (acerto em %)	0,00	2,16	22,52	35,04	7,57	11,86	6,97	18,52	1,10
Pertencente a classe (erro em %)	38,03	76,62	50,45	58,97	36,76	40,72	60,96	59,26	96,70
Não pertencente a classe (acerto em %)	72,72	70,33	55,20	81,40	33,99	44,68	48,83	63,03	94,00

O comitê de RNAs não possui uma boa especificidade pois, em muitos casos, obteve uma baixa taxa de acerto para uma classe específica. Por exemplo, as classes E F G M L R T V tiveram uma especificidade muito baixa. Uma justificativa a essa peculiaridade pode se dá pelo tamanho da janela da metodologia SCSW utilizada assim como os dados utilizados no treinamento com pouca representatividade.

Analisando a Tabela 2, percebe-se a existência de uma classificação indefinida dada pelos classificadores. Por exemplo, o "Classificador O" ao receber dados pertencente a classe, obteve 22,52% de acerto, 50,45% de erro e o restante interpretado pela classificador como indefinido.

8 – Conclusão

Neste trabalho foi construído um classificador proteínas utilizando RNA. E para avaliar e eficiência deste classificador, foram utilizados dados já classificados.

Dos resultados obtidos percebeu-se que as maiorias dos classificadores não obtiveram uma boa especificidade, já que não conseguiam classificar os dados que pertenciam a sua classe, corretamente. Por

outro lado os classificadores conseguiram classificar corretamente os dados que não pertenciam a sua classe.

Uma justificativa a essa peculiaridade pode se dá pelo tamanho da janela da metodologia SCSW utilizada.

9 – Propostas Futuras

Como propostas futuras de continuidade deste trabalho sugere-se que se invista nos seguintes problemas encontrados:

- Utilização do modelo SCSW com outros tamanhos de janela;
- Utilização do modelo extendido da metodologia SCSW;
- Utilização do Alfabeto de tamanho 20;
- Utilização de outros algoritmos de treinamento de RNAs.

10 - Referencias Bibliográficas

ALMEIDA, F. N.. "Implementação de um Banco de Dados de Proteomas de Bactérias Associadas a Plantas: Probacter". Petrópolis, 2007.

BARRETO, J. B. (2002). Introdução às Redes Neurais Artificiais. UFSC, 2002.

BISHOP, C. M. (2005). Neural Networks for Pattern Recognition. Pages 98 – 140.

BLAISDELL, B. E.. "A measure of the similarity of sets of sequences not requiring sequence alignment." Proc. Natl. Acad. Sci. USA, Vol. 83, pp. 5155-5159, Julho de 1986.

CASTANHO, M. (2005). Desenvolvimento de Abordagens Computacionais e Ferramentas para a Análise Comparativa de Genomas Microbianos. FIOCRUZ, 2005.

CHAMPE, P. C.; HARVEY, R. A.; FERRIER, D. R.; Bioquímica Ilustada 3ª Ed.. Editora Artmed, 2006.

DAYHOFF, M. O.. Survey of new data and computer methods of analysis. Atlas of protein sequence and structure, 5.(1978).

LEMES, N. H. T.. Redes Neurais Artificiais Volume I – Um Texto Básico. UNICOR, 2000.

MARZZOCO, A.; TORRES, B. B.; Bioquímica Básica 3ª Ed.. Editora Guanabara, 2007.

MENDES, D. Q.; Tutorial de Redes Neurais: Aplicações em Bioinformática. Disponível em: <http://www.lncc.br/~labinfo/tutorialRN/> - consultado em 17/05/2008.

NELSON, D. L.; LOX, M. M.; Lehninger Principios da Bioquinica 4ª Ed.. Editora Savier, 2006.

PASQUIER, C., HAMODRAKAS, S. J.. "An hierarchical artificial neural network system for the classification of transmembrane proteins." Protein Engineering, Vol. 12, No. 8, 631-634, August 1999

RODRIGUES, T. S. (2007). Codificação de Seqüências de Aminoácidos e suas Aplicação de Proteína com Redes Neurais Artificiais. UFMG, 2007.

TATUSOV, R.L.; GALPERIN, M. Y.; NATALE, D. A.; KOONIN, E.V.. The COG database: a Tool for Genome-Scale Analysis of Protein Functions and Evolution. Nucleic Acids Res. 2000 Jan 1;28(1):33-6.

TATUSOV, R. L., KOONIN, E. V., LIPMAN, D. J.. "A Genomic Perspective. on Protein Families," Science, vol. 278, pp. 631–637, 1997.

WU, C., WHITSON, G., MCLARTY, J., ERMONGKONCHAI, A., and CHANG, T.. Protein classification artificial neural system. Protein Science, (1):667–677. (1992)