Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp. 85–90, June 2–5, 2003 – Centro Universitário da FEI, São Paulo - SP - Brazil

# Aplicação de um Sistema Neural ao Problema de Classificação de Proteínas

Wagner Rodrigo Weinert[1], Heitor Silvério Lopes[2]
[1,2] Laboratório de Bioinformática / CPGEI
Centro Federal de Educação Tecnológica do Paraná (CEFET-PR)
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR), Brasil
E-mails: weinert@cpgei.cefetpr.br, hslopes@cpgei.cefetpr.br

## Abstract

*The prediction of protein function, based on its structure, is a hard problem in Molecular Biology. The number of known proteins has grown sharply, thanks to the several genomic sequencing efforts worldwide. This fact makes unfeasible the application of conventional laboratory techniques for function prediction of new proteins, therefore leading to the necessity of automatic classifiers for this purpose. Classification of an unknown protein means that it will be assigned to a known family, based on its structure (usually the primary structure), hence, supposing its function. This work presents the use of multi-layer perceptrons for protein classification. The main contribution is a novel way to encode data based on the hydrophobicity grade of amino acids of the proteomic chain. The classification of five different families are compared with a Hidden Markov Chain profile. Results show a good accuracy rate in the classification using the neural system and encourages future improvements.*

## 1. Introdução

A Biologia Molecular tem se desenvolvido rapidamente e uma enorme quantidade de dados é gerada constantemente. Entretanto, pouca informação se consegue extrair destes dados, logo a análise de dados tem se tornado o principal problema da biologia molecular [1].

O esforço internacional do Genoma Humano e sequenciamento de inúmeros outros organismos têm proporcionado um crescimento contínuo na descoberta de novas estruturas de proteínas. Este fato tem aumentado ainda mais a necessidade do desenvolvimento de métodos de classificação mais eficientes.

A utilização de redes neurais [2] no problema de classificação de proteínas tem trazido resultados promissores, como pode ser observado em [3] que propõe uma nova técnica para extrair aspectos de dados de proteínas e usa-os em conjunto com uma rede neural Bayesiana para classificar seqüências de proteínas, e em [4] que explora segmentos de seqüência ditos informativos para determinar estruturas e funções de classes de proteínas utilizando uma rede neural de três camadas com algoritmo de aprendizado *backpropagation*.

Um aspecto importante da aplicação de redes neurais na classificação de seqüências de proteínas é a codificação da seqüência que deve ser realizada de maneira a permitir o processamento desta em uma rede neural. Neste artigo é apresentada uma nova forma de codificação que transforma a seqüência de uma proteína (estrutura primária – Figura 1) num conjunto de valores reais que podem ser diretamente inseridos na primeira camada da rede neural, diferentemente da maioria dos métodos que extrai informações relevantes de seqüências de proteínas, como aspectos de similaridade [3], para então originar o conjunto de dados que será utilizado como entrada da primeira camada da rede neural.

## 2. Biologia molecular

As proteínas são constituídas de aminoácidos, sendo que estes se combinam através de ligações peptídicas fornecendo uma seqüência linear que contém informações necessárias para a geração de uma molécula protéica. As proteínas diferem umas das outras porque cada uma delas tem uma seqüência distinta de unidades de aminoácidos. Os aminoácidos são o alfabeto da estrutura protéica, pois eles podem ser agrupados em um número quase infinito de seqüências para formar um número igualmente infinito de diferentes proteínas [5].

A seqüência de aminoácidos de uma proteína está diretamente relacionada à sua função. As proteínas geralmente contêm subestruturas cruciais no interior da sua seqüência de aminoácidos, também conhecidas como *motifs* [6], as quais são essenciais para a execução de sua função biológica, funcionando também como identidade da proteína.

As proteínas são representadas por um alfabeto composto de 20 letras que representam os 20 aminoácidos que podem estar presentes em uma seqüência protéica, conforme Tabela 1.

Os aminoácidos diferem entre si pela estrutura da sua cadeia lateral [5]. As propriedades das cadeias laterais dos aminoácidos, principalmente o fato de algumas delas terem afinidade pela água e outras não, são importantes para a conformação das proteínas e, portanto, para a sua função.

Tabela 1: Aminoácidos e sua simbologia

| Letra | Símbolo | Nome do Aminoácido |
|---|---|---|
| A | Ala | Alanina |
| C | Cis ou Cys | Cisteína |
| D | Asp | Aspartato |
| E | Glu | Glutamato |
| F | Fen ou Phe | Fenilanina |
| G | Gli ou Gly | Glicina |
| H | His | Histidina |
| I | Ile | Isoleucina |
| K | Lis ou Lys | Lisina |
| L | Leu | Leucina |
| M | Met | Metionina |
| N | Asn | Asparagina |
| P | Pro | Prolina |
| Q | Gln | Glutamina |
| R | Arg | Arginina |
| S | Ser | Serina |
| T | Ter ou Thr | Treonina |
| V | Val | Valina |
| W | Trp | Triptofano |
| Y | Tir ou Tyr | Tirosina |

Os aminoácidos podem ser classificados de diversos modos. Um exemplo de classificação é a escala de hidrofobicidade de Kyte e Doolittle [7] que designa valores numéricos para hidrofobicidade de cada um dos 20 aminoácidos. A Tabela 2 mostra os 20 aminoácidos ordenados de acordo com esta escala, sendo que podem também ser agrupados em três subconjuntos: aminoácidos hidrofóbicos, neutros e hidrofílicos.

As proteínas possuem estruturas espaciais complexas que podem ser organizadas em quatro níveis de complexidade (Figura 1): primário (dado pela seqüência de aminoácidos e ligações peptídicas da molécula – nível estrutural mais simples e importante), secundário (dado pelo arranjo espacial de aminoácidos próximos entre si na seqüência primária), terciário (fornecido pelo arranjo espacial de aminoácidos distantes entre si na seqüência polipeptídica) e quaternário (dado pela distribuição espacial de uma cadeia polipeptídica no espaço, as subunidades da molécula).

Existem inúmeras bases de dados de acesso público disponíveis na Internet, algumas com mais, outras com menos informações. O estudo de caso apresentado na seção 7 utiliza informação referente à estrutura primária de um conjunto de proteínas do Protein Data Bank (PDB) (http://www.pdb.org). O PDB é um repositório internacional de dados públicos de estruturas tridimensionais de macromoléculas biológicas. Os conteúdos são originados de cristalografia de raios-X e experimentos de ressonância magnética nuclear.

A Figura 2 representa a estrutura primária da proteína 1C56 pertencente à família Toxin, onde o termo "família" refere-se a um conjunto de proteínas semelhantes que, em geral, compartilham estruturas e funções semelhantes.

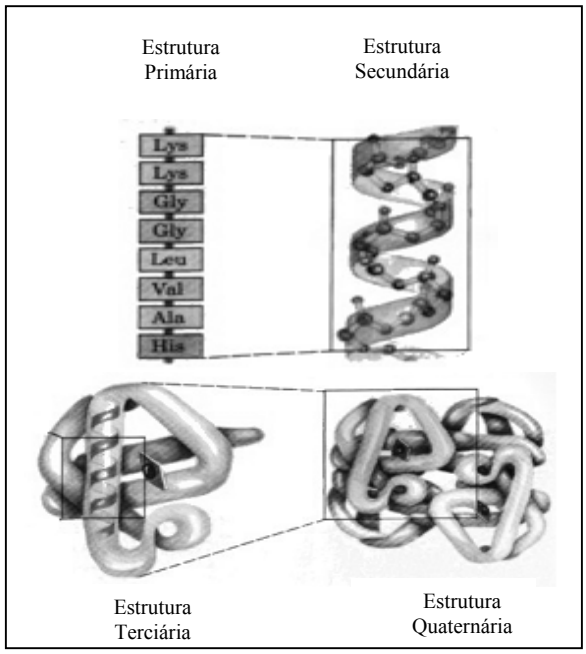


Figura 1: Estrutura primária, secundária, terciária e quaternária de uma proteína

Se uma proteína é classificada como pertencente a uma determinada família então se pode inferir a estrutura e principalmente a função desta proteína. Este processo é importante em muitos aspectos da Biologia Molecular.

WCSTCLDLACGASRECYDPCFKAFGRAHGKCMNNKCRCYT

Figura 2: Proteína 1C56 pertencente à família Toxin, com apenas 40 aminoácidos

## 3. Classificação de proteínas

Quase todas as proteínas têm estruturas similares com outras proteínas e, em diversos casos, partem de uma origem evolucionária comum [8], o que dificulta a sua classificação.

Embora não exista um padrão de classificação universal, as proteínas podem ser classificadas com base na sua composição, número de cadeias, forma ou função biológica [9]. Uma outra classificação é sugerida por [8] que apresenta um modelo de classificação baseado no domínio da proteína. As proteínas até então identificadas e classificadas estão catalogadas em vários bancos de dados de domínio público como, por exemplo, o PDB, o Swiss-Prot, o PIR e o InterPro.

Várias técnicas computacionais já foram empregadas na tarefa de classificação de proteínas, como por exemplo, redes neurais [3][4], transformações estruturais [10], algoritmos genéticos [11], programação genética [12] e cadeias de Markov [13], porém sempre com resultados e aplicabilidade limitados.

## 4. Codificação das estruturas protéicas

A codificação de seqüência é fundamental para a modelagem da arquitetura de rede neural. Uma boa representação dos dados de entrada é crucial para o sucesso da aprendizagem da rede.

As proteínas, em geral, são formadas por quantidades variadas de aminoácidos, não é raro encontrar proteínas compostas por mais de 3000 aminoácidos, o que dificulta a modelagem de uma arquitetura de rede capaz de receber esta seqüência de aminoácidos como conjunto de entrada de dados.

Neste trabalho é apresentada uma nova forma de codificação baseada unicamente na seqüência de aminoácidos que forma a proteína, vista computacionalmente como um *string*. Esta nova forma de codificação permite encontrar um modelo de arquitetura fixa para o conjunto de redes neurais que farão parte do sistema classificador, mesmo que o conjunto de treinamento destas redes seja composto por proteínas de tamanhos diferentes.

A quantidade de redes neurais necessárias para o desenvolvimento do sistema de classificação proposto é determinada pelo comprimento das proteínas pertencentes ao conjunto de treinamento e pelo tamanho de partição (segmento de aminoácidos) escolhido para o processo, como será detalhado mais adiante.

A codificação consiste na determinação de um alfabeto numérico de valores reais entre 0 e 1, excluindo o zero, que represente os 20 aminoácidos, seguindo a escala de hidrofobicidade de Kyte e Doolittle [7], como mostra a Tabela 2.

Tabela 2: Valores de hidrofobicidade, e codificação adotada em valores reais para o sistema neural

| Aminoácido (símbolo) | Valor K.D. | Valor Real | Categoria |
|---|---|---|---|
| I | +4,5 | 0,05 | Hidrofóbico |
| V | +4,2 | 0,10 | Hidrofóbico |
| L | +3,8 | 0,15 | Hidrofóbico |
| F | +2,8 | 0,20 | Hidrofóbico |
| C | +2,5 | 0,25 | Hidrofóbico |
| M | +1,9 | 0,30 | Hidrofóbico |
| A | +1,8 | 0,35 | Hidrofóbico |
| G | -0,4 | 0,40 | Neutro |
| T | -0,7 | 0,45 | Neutro |
| S | -0,8 | 0,50 | Neutro |
| W | -0,9 | 0,55 | Neutro |
| Y | -1,3 | 0,60 | Neutro |
| P | -1,6 | 0,65 | Neutro |
| H | -3,2 | 0,70 | Hidrofílico |
| Q | -3,5 | 0,75 | Hidrofílico |
| N | -3,5 | 0,80 | Hidrofílico |
| E | -3,5 | 0,85 | Hidrofílico |
| D | -3,5 | 0,90 | Hidrofílico |
| K | -3,9 | 0,95 | Hidrofílico |
| R | -4,0 | 1,00 | Hidrofílico |

## 5. O sistema neural

Uma vez determinada a forma de codificação, passa-se então à construção do sistema neural. O primeiro passo é selecionar o conjunto de dados que fará parte do treinamento. Por exemplo, *n* famílias compostas de *m* proteínas cada uma, lembrando que *m* varia de família para família.

Tendo-se selecionado o conjunto de treinamento parte-se para definição do tamanho do segmento *t* de aminoácidos que será utilizado no sistema neural, na prática esse tamanho de segmento será a quantidade de neurônios da primeira camada de cada rede neural que compõe o sistema.

Faz-se, então, uma reformulação do conjunto de treinamento. Cada proteína é dividida em *k* subconjuntos de aminoácidos, pela equação (1) onde *x* refere-se a quantidade de aminoácidos da proteína, se esta divisão não resultar num valor inteiro, a parte fracionária referente ao final Tabela 2: Valores de hidrofobicidade, e codificação adotada em valores reais para o sistema neural da cadeia protéica é desprezada, assim como as proteínas que tenham um comprimento menor do que *t*. O maior valor de *k* entre todas as proteínas determina a quantidade de redes *u* que compõem o sistema neural. Durante essa reformulação os símbolos que representam os aminoácidos dentro da seqüência protéica são substituídos por seus referidos valores reais seguindo a Tabela 2.

$$k = (\frac{x}{t}) \quad (1)$$

Tendo-se o valor de *u, t* e *n* e utilizando uma rede do tipo Perceptron de Múltiplas Camadas (MLP) com *backpropagation*, define-se a arquitetura do sistema neural, representado de forma genérica na Figura 3.

Uma vez que o conjunto de treinamento é dividido em diversos blocos de segmentos de aminoácidos (veja exemplo na Figura 4), cada um desses blocos faz parte do conjunto de treinamento de uma das redes neurais que compõem o sistema neural.Pelo fato das proteínas possuírem diferentes quantidades de aminoácidos em sua composição o conjunto de treinamento de cada rede é formado por diferentes quantidades de amostras, exigindo uma política de pesos diferenciada para cada rede como será detalhado a seguir.

Observando-se o exemplo da Figura 4 pode-se compreender melhor esta nova abordagem de codificação. Considere-se neste exemplo um conjunto de treinamento formado por 2 famílias de proteínas, onde cada família é composta por 3 proteínas e a maior cadeia protéica têm 245 aminoácidos. Primeiramente, determina-se o tamanho de segmento *t,* por exemplo, 40. Então, se extrai *u* da divisão inteira de (245 / 40). Conclui-se então que o sistema neural, para este caso, é composto de 6 redes (*u*) com 40 neurônios na primeira camada de cada uma delas. Partindo destes parâmetros de configuração reformula-se o conjunto de treinamento.

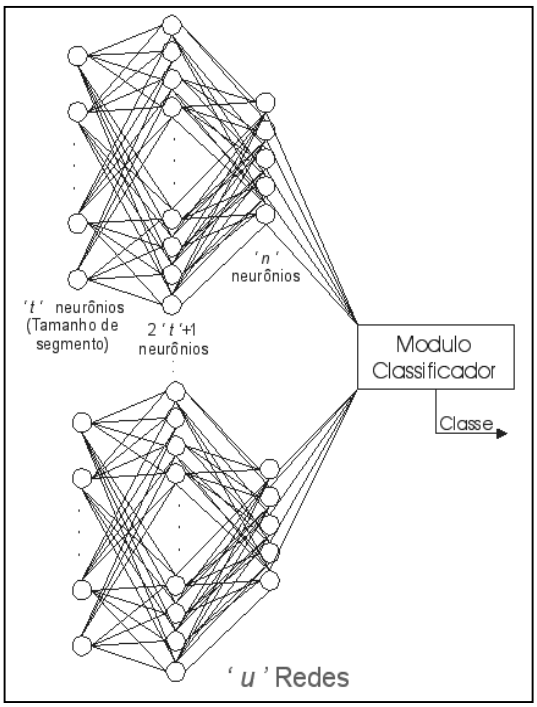


Figura 3: Representação genérica do sistema neural

| | | Segmentos (u) | | | | | |
|---|---|---|---|---|---|---|---|
| **Família (n)** | **Proteína (m)** | **1** | **2** | **3** | **4** | **5** | **6** |
| 1 | 1 | | | | | | |
| | 2 | | | | | | |
| | 3 | | | | | | |
| 2 | 1 | | | | | | |
| | 2 | | | | | | |
| | 3 | | | | | | |
| | Total | 6 | 5 | 4 | 3 | 3 | 2 |
| | Pesos | 1 | 0,8 | 0,7 | 0,5 | 0,5 | 0,3 |

Figura 4: Exemplo de conjunto de treinamento

No exemplo da Figura 4, a proteína 1 da família 1 é composta originalmente por 122 aminoácidos, após o rearranjo, ela foi dividida em 3 segmentos de 40 aminoácidos cada um, onde cada segmento é apresentado a uma rede em ordem crescente, ou seja, o segmento 1 à rede 1, o segmento 2 à rede 2 e assim sucessivamente. Assim, é possível perceber que o conjunto de treinamento de cada rede é formado por uma quantidade de amostras diferenciada (rede 1 – 6 amostras, rede 4 – 3 amostras), o que leva à aplicação de uma política de pesos às saídas destas redes quando submetidas a um caso de teste de classificação.

Esta política de pesos está diretamente relacionada com a quantidade de amostras do conjunto de treinamento de cada rede, ou seja, cada rede tem sua quantidade de casos amostrais de treinamento dividido pela quantidade de casos amostrais da rede 1, logo a rede 1 sempre terá peso 1. As redes seguintes terão pesos proporcionais (no exemplo, rede 2 – peso 0,8, rede 4 – peso 0,5). Vale lembrar que esta política de pesos é aplicada somente no processo de classificação, ou seja, teste da rede, não em seu treinamento e que todas as redes agem de forma independente uma da outra, assim podem ser processadas em separado.

Para escolha da quantidade de neurônios da camada escondida de cada rede utiliza-se a fórmula $2t + 1$ [2], onde $t$ é o número de neurônios da camada de entrada, resultando neste exemplo (Figura 4) em 81 neurônios e a camada de saída é determinada pela quantidade de classes que se deseja classificar (famílias), neste exemplo, 2 famílias, num formato binário como mostra a Tabela 3.

Quando se analisa o vetor de saída a fim de se obter a resposta da rede para um determinado caso de teste, varre-se o vetor e se extrai o maior valor deste, juntamente com sua posição, sendo que esta indica a classe na qual a proteína é classificada. Por exemplo, se um conjunto de treinamento fosse formado por 5 classes e em um determinado caso de teste apresentasse como resposta o vetor [0,30 0,20 0,96 0,10 0,21], o determinado caso de teste seria classificado como pertencente à classe 3 (maior valor).

Tabela 3: Classes e padrões de saída

| Classes (Famílias) | Padrões de Saída |
|---|---|
| Família 1 | [ 1 0 ] |
| Família 2 | [ 0 1 ] |

## 6. Procedimento de classificação

Seleciona-se a proteína que se deseja classificar, dividindo-a em segmentos e trocando os símbolos de seu alfabeto pelos seus respectivos valores reais, de modo análogo ao processo aplicado ao conjunto de treinamento. Então ela é apresentada ao sistema neural. Por exemplo, seguindo a Figura 4, uma proteína de comprimento 123 é submetida ao processo de classificação. Esta proteína é dividida em 3 segmentos onde o primeiro é apresentado à rede 1, o segundo à rede 2 e o terceiro à rede 3. Supondo que as respostas das redes 1, 2 e 3 sejam [0,98 0,15], [0,75 0,63] e [0,50 0,70] respectivamente. Essas respostas são então enviadas ao módulo classificador (Figura 3), que aplica os pesos de cada rede a cada vetor de resposta, realizando ao final a soma dos três vetores e fornecendo a classificação obtida. Neste exemplo: [0,98*1,00 0,15*1,00] + [0,75*0,80 0,63*0,80] + [0,50*0,70 0,70*0,70] = [1,93 1,14] classificando a referida proteína como sendo pertencente à família 1.

O resultado desta nova forma de codificação para o problema de classificação de proteínas com a utilização de redes neurais é apresentado num estudo de caso a seguir.

## 7. Estudo de caso

Para a realização deste estudo de caso extraiu-se 5 famílias de proteínas do PDB (Oxigen Transport, Hidrolase (o-glicosyl), Lyase, Toxin e Imuno Globulin). Estas 5 famílias foram escolhidas pelo seguinte critério: primeiramente realizou-se uma poda no banco de dados que separou todas as famílias que possuíam mais de 150 proteínas em seus conjuntos. Destas famílias, 5 foram escolhidas ao acaso. Então, 100 amostras de cada uma das 5 famílias foram escolhidas aleatoriamente e formaram o conjunto de treinamento do sistema neural que fora construído como explicado anteriormente.

Este conjunto de treinamento deu origem a um sistema neural com as seguintes características: 70 redes neurais tipo MLP que seguem a lei de aprendizado do *backpropagation*, com 40 neurônios na primeira camada, 1 camadas intermediárias com 81 neurônios e uma camada de saída de 5 neurônios.

As redes foram treinadas por 300 épocas com uma taxa de aprendizado $\alpha = 0,1$. Para a implementação do sistema neural utilizou-se o pacote de redes neurais do MATLAB 5.3. O módulo classificador (Figura 3) foi implementado em Delphi 5.0 e tem como entrada a pré-classificação realizada por cada uma das redes para um determinado conjunto de teste. Este módulo aplica a política de pesos sobre os resultados parciais de cada rede e fornece a classificação final.

O conjunto de teste foi composto de 50 amostras escolhidas ao acaso de cada família, sendo que estas naturalmente não faziam parte do conjunto de treinamento.

Para avaliar o desempenho do sistema neural proposto, foram comparados os resultados de classificação do sistema neural com os resultados apresentados pelo pacote de *software* HMMER 2.2 (http://hmmer.wustl.edu) que utiliza Cadeias Ocultas de Markov (HMMs) para analisar seqüências biológicas [13]. HMMs são uma técnica de modelagem estatística geral para problemas "lineares" e têm sido usadas em aplicações para a análise estrutural de proteínas.

O processo classificatório realizado pelo HMMER 2.2 se subdivide em três etapas.

A primeira etapa consiste em encontrar um alinhamento múltiplo para o conjunto de treinamento. O pacote não oferece uma ferramenta específica para realizar este alinhamento. O ClustalX 1.81 (http://inn-prot.weizmann.ac.il/software/ClustalX.html) é um *software* de referência para esta tarefa, e pode ser utilizado. Foram gerados 5 arquivos de alinhamento múltiplo (um para cada família - Oxigen Transport, Hidrolase (o-glicosyl), Lyase, Toxin e Imuno Globulin - compostos de 100 proteínas cada um).

A segunda etapa consiste da criação do modelo estatístico que representa o conjunto de treinamento. Cada um dos 5 arquivos contendo o alinhamento múltiplo das 5 famílias de proteínas é apresentado ao *software hmmbuild*, que faz parte do HMMER 2.2, e cria um modelo estatístico único para as 5 famílias que compõem o conjunto de treinamento. O modelo pode opcionalmente passar por um processo de otimização (*hmmcalibrate*) que pode melhorar a precisão da classificação realizada pelo módulo *hmmpfam*.

A terceira etapa é o processo de teste do modelo estatístico, ou seja, verificação da taxa de acerto dado um conjunto de teste. Para este processo utilizou-se o mesmo conjunto de teste aplicado ao sistema neural. Então o *software hmmpfam* confronta cada um dos 5 arquivos de teste contendo 50 amostras cada, com o modelo gerado e, finalmente computam-se os resultados.

As percentagens de acerto do sistema neural e do HMMER 2.2 no processo classificatório, utilizando os mesmos conjuntos de treinamento e teste, estão apresentados na Tabela 4.

Tabela 4: Taxas de acerto para o estudo de caso

| Famílias | sistema neural | HMMER 2.2 |
|---|---|---|
| Oxigen Transport | 84% | 84% |
| Hidrolase (o-g.) | 92% | 100% |
| Lyase | 80% | 54% |
| Toxin | 76% | 76% |
| Imuno Globulin | 96% | 100% |
| média | 85,5 % | 82,8 % |

## 8. Conclusões

Apresentou-se aqui uma nova técnica de codificação para a resolução do problema de classificação de proteínas baseando-se na sua estrutura primária, com a utilização de redes neurais MLP.

Esta nova forma de codificação leva em consideração somente a seqüência de aminoácidos que forma a proteína seguindo a escala de hidrofobicidade de Kyte e Doolittle. A idéia da distribuição do problema para um conjunto de redes ao invés de apenas uma rede neural juntamente com a abordagem de ponderação permitiu a aplicação eficaz desta forma de codificação.

Os resultados obtidos são extremamente satisfatórios, apesar de não ter ocorrido nenhuma taxa de acerto de 100% com sistema neural como ocorreu com HMMER 2.2. As taxas de acerto na classificação se mantiveram altas, diferentemente do HMMER 2.2 que teve um índice de 54% de acerto para a família Lyase. Observando-se as médias de acerto de cada sistema de classificação, percebe-se que, para este estudo de caso o desempenho das duas abordagens é equivalente, sendo que se obteve uma pequena melhora com sistema neural A repetição deste experimento (dados não mostrados aqui) revelam a robustez do sistema neural em relação à outra abordagem, sendo menos sensível a variações no conjunto de treinamento. Considerando a complexidade do problema abordado, os resultados encorajam a realização de novos experimentos e estudos sobre esta nova forma de codificação e classificação de proteínas.

## Referências


[ 1 ] Y. Hu. Biopattern Discovery by genetic programming. *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 152-157, 1998.

[ 2 ] L. Fausett. *Fundamentals of Neural Networks***.** Upper Saddle River: Prentice-Hall, Inc., New Jarsey, 1994.

[ 3 ] J. T. L. Wang, Q. Ma, D. Shasha, C. H. Wu. Application of neural networks to biological data mining: A case study in protein sequence classification. *International Conference on Knowledge Discovery and Data Mining*, pages. 305-309, 2000.

[ 4 ] C. H. Wu, G. M. Whitson, G. J. Montllor. PROCANS: A protein classification system using a neural network. *Proceedings International Joint Conference on Neural Networks*, 2:91-96, 1990.

[ 5 ] A Lehninger. *Princípios de Bioquímica***.** Savier Editora de Livros Médicos, São Paulo, 1991.

[ 6 ] T. K. Attwood, M. E. Beck, A. J. Bleasby, K. Degtyarenko, D. J. P. Smith. Progress with the PRINTS protein fingerprint database. *Nucleic Acids Research*, 24(1):182-188, 1996.

[ 7 ] J. Kyte, R. Doolittle. A simple method for displaying the hydropathic character of proteins. *Journal of Molecular Biology*, 157:105-132, 1982.

[ 8 ] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia. A structural classification of proteins database for the investigation of sequences and structures. *Journal Molecular Biology*,247:536-540, 1995.

[ 9 ] A. L. Lehninger, D. L. Nelson, M. M. Cox. *Principles of Biochemistry with an Extended Discussion of Oxigen – Binding Proteins*. 2. ed. Worth Publishers Inc., New York, 1998.

[10] T. Ohkawa, D. Namihira, N. Komoda, A. Kidera, H. Nakamura. Protein structure classification by structural transformation. *Proceedings of IEEE International Joint Symposium on Intelligence and Systems***,** pages 23-29, 1996.

[11] S. Chiba, K. Sugawara, T. Watanabe. Classification and function estimation of protein by using data compression and genetic algorithms. *Proceedings of the 2001 Congress on Evolutionary Computation*, 2:839-844, 2001.

[12] J. R. Koza. Classifying protein segments as transmembrane domains using genetic programming and architecture-altering operations. In: Baeck, T., Fogel, D. B., and Michalewicz, Z. (editors) *Handbook of Evolutionary Computation*. Bristol, UK: Institute of Physics Publishing, 6(1):1-5, 1997.

[13] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755-763, 1998.


Rev. 15/04/2003