Protein Classification Using Artificial Neural Networks with Different Protein Encoding Methods

Andre Luis Debiaso Rossi and Maria Angelica de Oliveira Camargo-Brunetto
Computer Science Department - State University of Londrina
andrerossidc@yahoo.com.br , angelica@uel.br

Abstract

The fast growth of annotated biological data implies in the need of developing new techniques and tools to classify these data, in such way that they can be useful. Protein classification is one relevant task in this context. This paper presents different models of neural network, aiming to compare the influence of the protein sequence encoding method in the performance of the Neural network to classify proteins. Besides, it is proposed two methods of protein sequence encoding, that were tested with several neural network, for classifying proteins using two approaches: based on families of proteins and based on function of proteins. The results of performance of the neural networks are presented and compared with other works in the area.

1. Introduction

The challenge for treating or preventing genetic diseases is one of the main reason to the research in Molecular Biology. The advance in this area, as well as the results of Genom Projects has implied in an increasing quantity of biological data, mainly DNA, RNA and proteins [18, 16].

But would be usefulless so much available data dissociated of information and knowledgment about them. So, it is necessary mining and organizing data in order to give meaning to them.

Several techniques are used to perform specific tasks, as sequencing, annotation and classification of proteins. Sequencing and annotation are performed in large scale through Bioinformatics technologies that are essential to handle biological data.

The protein classification is an important area of Bionformatic, since it is a way of organizing biological data resulting from the sequencing of these polypeptideos. Proteins can be classified by different aspects, as for example, by its function in the cell or by its families or superfamilies.

Discovering structure and function of proteins is usually solved by algorithms that seeking for homologies between molecular sequences. However, the fast growth in the biological data sequenced has difficulted the use of algorithms based on alignment as base technique. Another problem presented by [4] is that such kind of algorithms have limitations in classifying protein families with low similarity. More recentlly, alternative techniques have been used, as Artificial Neural Networks (ANN) and Genetic Algorithms.

Artificial Neural Networks (ANNs) have been shown a powerful technique to solve problems of classification and clustering in a wide field of applications. In Bioinformatics the task of protein classification has been investigated for several authors, showing promissing results [21]. An important issue concerning with the performance of ANN in the task of protein classification is the protein sequence encoding . In this paper are presented three methods of protein encoding sequence to be used in ANN to protein classification. The main contributions of this paper are the purpose of two modified existent algorithms to protein sequence encoding as well as comparing the performance of ANN using three different methods including the two ones.

2. Biological and Computational Foundations

It is presented background of the main areas in this research: Molecular Biology and Artificial Neural Networks.

2.1. Molecular Biology

The main issues of study in Molecular Biology are genetic material and the proteins [15].

Proteins have a wide number of biological functions, including catalizing chemical reactions, controlling membrane permeability and controlling genic function. They determine the shape and the structure of the cell [14].

Proteins are composed by amino acids, that can be connected through a special link named peptide bond. During this link there is lost of one water molecul.

*Thanks to Prof. Dr A. L. M. de Oliveira by his help on elucidate biological concepts

0-7695-2976-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ISDA.2007.81

Each protein has a unique sequence of amino acids, which determines the protein function [3]. The alphabet of the protein structure is composed by 20 amino acids, represented by 20 letters, and can be grouped in an almost unbounded number of sequences in order to compose an uncountable number of proteins.

2.2. Artificial Neural Networks

An Artificial Neural Network is a computational model inspired in the functioning of the human brain. It is composed by a set of artificial neurons (known as processing units) that are interconnected with other neurons. Each connection has a weight associated that represents the influence from one neuron on the other. The first formal model of a artificial neuron was proposed in 1943 by McCulloch and Pitts. They show that such model was able to perform the computation of any computable function using a finite number of artificial neurons and sinaptic weights adjustable [9].

Each processing unit is composed by three elements, namelly: **Sinapses** - Weight associated to each input of the neuron that determines if that input will contribute to active or not the neuron. Each synapse is responsable for the storing of knowledgment; **Acumulator** - performs the ponderate sum of the input multiplied by the correspondent weights; and **Activation Function** - determines the activation level based on the result given by the acumulator, bounding the output value of each neuron. Normally the output interval is [0,1] or [-1,1].

A neuron can be expressed by the following equations

$\mu_k = \sum_{j=1}^{m} w_{kj} x_j$ and $y_k = \varphi(\mu_k)$

where , x_j is a input sign , w_{kj} is the weight from the neuron k to the neuron j; μ_k is the result of the acumulator, φ is the activation function and y_k is the output of the neuron.

There are several possible activation function. The choice of a suitable function depends on the problem that the neuron is intended to solve [8]. The sigmoid functions are commonly used, which have as example the logistic function or hyperbolic tangent.

Using this formal model of artificial neuron, several ANN models have been proposed, including Perceptron and MultiLayer Perceptron (MLP). This is perhaps the most used model in a wide variety of applications. A MLP is a feedforward network of interconnected neurons, organized in layers: the input layer, one or more hidden layers (where most processing effort takes place) and the output layer. The topology of a MLP is specified by the number of layers and by the number of neurons of each layer. Besides the topology, a MLP uses a activation function and a learning (training) procedure. There are different learning procedures that can be used in a MLP, and it seems backpropagation is the most popular ones. In the backpropagation algorithm, a set of input-output patterns is presented to the ANN and the computed result is compared to the actual value (provided by an expert). The weight is updated, taking into account the mean square error (MSE) from the difference between the actual value and the computed value.

3. Protein Classification

Protein classification can be done under different approaches, being the most commonly used (i) Classification based on a hierarchy of the molecular taxonomy, composed by superfamilies, families and sub-families [5] and (ii) Classification based on functional annotation, that determines what is the function of a protein in a cell.

When it is found a protein with unknown function, which structure is similar to a protein with known function, it can be inferred that it belongs to the same superfamily/family, or shares similar structures and has the same function [16]. With this kind of classification, frequently it is possible to infer the function of a protein sequence [22]. The performance of classifiers can be measured using the metrics precision, sensibility and specificity. Such metrics are defined according to the following equations: (1), (2) and (3).

$$precision = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

$$sensibilility = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$specificity = \frac{TN}{TN + FP} \times 100\% \quad (3)$$

where TP=True Positive, TN= True Negative, FP=False positive and FN=false negative .

These equations have been used in our research to analyze the results delivered by the neural networks for the task of protein classification.

3.1. Protein Classification Methods

Protein classification can be done throught direct or indirect modelling. The former uses a training set of sequences to build a model that characterizes a family or a function of interest. In the indirect modelling, it is used direct modelling as pre-processing, aiming to extract useful features from the sequences, that are transformed in arrays of fixed length, that can be used to train a ANN, for example [2].

BLAST (*Basic Local Alignment Search Tool*) [1] is a tool that uses direct modelling, performing a search of homologie between sequences. This software explores the local alignment in pairs to measure the similarity between sequences. The classification is done based on the alignment which had the greatest punctuation.

Another method that uses direct modelling is the HMM *Hidden Markov Models* that is widely used for probabilistic modelling of family of proteins. It uses probabilistic values to score how much an unkown protein belongs to a given family [2].

3.2. Related Works on Protein Classification with ANN

Nowadays, the most common used approach to discover the structure and the function of proteins are the algorithms that use seek of homologies between sequences. However, the fast increase in the number of sequenced biological data has difficulted the use of this kind of approach [21].

Another problem presented by [4] is that algorithms based on seek of homologies between sequences and methods based on search of patterns are limited or have difficulties in classifying families of proteins that have weak similarity.

The use of ANN for the task of protein classification has been motivated by the results found by [19], [16] and [18]. Some of these methods have obtained accuracy equal or greater than other classification methods.

The process of protein classification using ANN can be divided in three parts: (i) pre-processing: protein sequence encoding; (ii) processing the protein classification using the ANN and (iv) post-processing: decoding the output (if necessary) to determine the class.

3.2.1. Protein Sequence Encoding to Generate the Input of ANNs

The protein encoding aims to convert the amino acids sequence in input arrays to the ANN [20]. The encoding method is one of most important phasis in order to obtain high accuracy and a best performance of the ANN in the task of protein classification [19].

The amino acids sequences can be interpreted in different ways, depending on the biological meaning or on its linear sequence, from what they are encoded. In the works of [21] and [19], it is used a kind of hashing function called *n-gram* The use of *n-grams* makes possible that all the encoded sequences are represented by a same number of parameters, in such way that the number of neurons in the input layer is fixed.

The *n-gram* is a way of representing a long chain of string in many substrings of length n. According to [13] this function is used with the following general aims: a) to minimize the processing time; b) to guarantee the non-ocurrence of false negative; c) to minimize the number of false positives. The *n-grams* have been used in text indexing, particularly substring matching. For example, using *bigram (2-gram)* to extract substrings from *string* ABCDEFGHIJ, the result is the following substrings: AB, BC, CD, DE, EF, FG, GH, HI, IJ.

The *n-grams* are also useful to classification based on location of regions of similarity of proteins [19]. A new protein encoding method was proposed by [18] that uses hydrofobicity scale of amino acids elabored by Kite [12].

3.2.2. Trainning an ANN to Classify Protein Sequence Encoded

There are different ANN models to solve a wide variety of problems. The choice of one model depends on its suitability to the problem adressed. The MultiLayer Perceptron (MLP) is the most used in the task of protein classification, as is presented in the works of [19] and [18]. Another ANN model that was already used to classify proteins in families is a Bayesian neural network, known as BNN [17].

It is also possible to use ANN with non-supervisioned learning, when the number and composition of clusters are not known. A work presented by [7] has been used Kohonen Self-Organizing Maps to perform clustering of protein families and for filogenetic classification.

3.2.3. Post-Processing

Sometimes the output delivered by the ANN can not be interpreted directly. It requires a kind of post-processing in order to give semantic to the output.

In the work of [18], each ANN generates five outputs, according to five possible functional classes of proteins. As the number of ANN is related to the lenght of the sequence of amino acids encoded, each protein is classified by a proper number of ANN. So, at the end, it was necessary to combine the output of all ANNs in order to delivery the final output. In this case, each ANN had a proportional weight to the ammount of processed proteins, so the output presented should be submitted to a function evaluation.

Another way used in [19] was to separate the superfamilies in modules of functional classes. So the number of outputs is the number of superfamilies presented during the training of that network. The number of superfamilies choosen for each module was a value between 100 and 200, once with a smaller value, a great number of false positives would be obtained, and with a greater number of superfamilies, the ANN would take a lot of time to be trained.

Several researches present comparative results with different methods of classification. In their study, [16] compare the performance of a Bayesian Neural Network (BNN) with other three techniques used to classify protein, as well as their combination. The techniques compared were: the BLAST, that is based in the alignment of sequences, the software SAM [10] that is based on Hidden Markov Models (HMM) and the SAM-T99 [11], also based on HMM,

but with the model built interactivelly. The comparisons have been done to classify proteins in the following superfamilies:*globin*, *kinase, ras* and *ribitol*. The BNN proposed presented better precision and specificity than BLAST and SAM-T99 and it was notable faster than BLAST.

In other work, [18] proposes an ANN, which inputs are protein sequence encoded using hydrofobicity scale. The authors compared their results with that produced by the software HMMER [1], that is based on the HMM [6].

4. The ANNs Modelled Using Different Protein Encoding Sequence Methods

In this section it is presented the study performed with three ANN models, being each model with a different method of protein encoding sequence. Each ANN model has been tested to perform classification by family and by function. It has been chosen the same functional class used in the work of [18] aiming to compare the results. For the classification of proteins based in families, it was used the same families of the work of [20].

4.1. The Protein Sequence Encoding Methods Analyzed

In this work we analyze three protein sequence encoding methods, to be used as input of Artificial Neural Networks. The first is the application of the function **2-gram** on the **l**inear **s**equence of amino acids, according to [21], that we call **lsa2**. The second method is an extension of the work of [18], that uses hydrofobicity scale. In the original purpose, several ANN are necessary to perform the classification, once the input encoding sequences are of variable lenght. We have proposed to apply the 2-gram function over the original encoding sequence so that only ANN is necessary. We refers this method by **hyd2**. Finally, it is proposed an hybrid method to protein sequence encoding that uses information from the two later methods, performing the concatenation of the two resultant input matrices.This method has been named **lsa2hyd2**.

4.2. The ANNs Modelled

The ANNs were modelled based on the MultiLayer Perceptron (MLP) model. The input of the ANN has the same format so the **lsa2** as the **hyd2**: a matrix of 400 columns by m rows. The number of columns is obtained, considering 2-gram and the number of possible amino acids that is 20. So, the number of possible sub-strings is given by 20^2. The number of rows is the number of proteins presented to the ANN to be classified.

[1] http://hmmer.wustl.edu

For the method **lsa2hyd2**, the input of the ANN is a matrix with 800 columns by m rows. This number of columns resulted from the concatenation of the columns of the **lsa2** and **hyd2**.The number of rows has the same meaning as in other ANN models. This method has been proposed aiming to test the performance of the ANN with a greater number of information about the protein to be classified.

For the protein classification in functional groups it has been selected 100 sequences belonging to each functional group and 100 belonging to none of the classified groups, to do the role of negative sample to the neural network. For the protein classification in families, it has been used 39 negative samples. The number of positive samples for each family has been defined as 137, 196, 58 and 34 for the families *cytochrome b*, *cytochrome c*, *flavodoxin* and *invertebrate defensin*, respectivelly.

The training set was established as 2/3 of the samples and the testing set as 1/3 of the samples. The criteria to accept a classification as correct were based on the following: for each test sample, it was computed the mean square error (MSE), and this should be less or equal 0.2. Besides, this criteria must be repeated for at least 80% of the repetitions performed.

In order to find the best configuration of the MLP, several simulations have been done, with different topologies and variation of the parameters. Considering the best precision rate and small values of MSE, it has been choosen the following configurations: For protein classification based on its function: A MLP with one hidden layer with 5 neurons, having been performed tests with the functional class *lyase*, the sla2 method, with the parameter learning rate, η=0.1, $\eta = 0.7$ and momentum term $\gamma = 0$, $\gamma = 0.5$, 20 repetitions of 50 epochs each. For protein classification based on families - a MLP with one hidden layer with 10 neurons. The test was performed with the family *flavodoxin* , the **sla2** method, with $\eta = 0.1, \eta = 0.7$ and $\gamma = 0, \gamma = 0.5$, 10 repetions of 200 epochs each.

4.3. The Tests Performed

The functional classes used were: *toxin*, *hydrolase (o.g)*, *immunoglobulin*, *oxygen transport* and *lyase*. The database Protein Data Bank (PDB) was used to get the sequences.

The families selected from the work of [20] were: *cytochrome b* (*cyt b*), *cytochrome c* (*cyt c*), *flavodoxin* (*flav*) and *invertebrate defensin family (invert)*. For this kind of classification, it was used the databse Swiss-Prot/TrEMBL.

Both databases used offers several formats for extracting the sequences, and it was selected FASTA format, that is available in both databases, once this format is simple, presenting the linear sequence of amino acids, the only necessary information to test the three methods of protein sequence encoding.

5. Results and Discussion

For both approaches of protein classification: based on families and based on protein function the same methodology of analysis was used.

In order to evaluate sensibility and specificity of the ANNs tested, it has been used different values to the parameter learning rate (η) and to the momentum term (γ) for each functional class and the encoding method. On tables 1 and 4 these values are represented by the pair (η, γ) on the intersection between each functional class and encoding method. The results of protein classification based on functional classes are presented on the tables 2 and 3 to sensibility and specificity, respectivelly. The results of protein classification based on families of proteins are presented on the tables 5 and 6.

Table 1. Configuration of ANNs for protein functional classification

	lsa2	*hyd2*	*lsa2hyd2*
hydrolase	(0.7, 0.7)	(0.4, 0.7)	(0.7, 0.5)
immunoglobulin	(0.4, 0.0)	(0.4, 1.0)	(0.7, 0.5)
lyase	(0.4, 1.0)	(0.7, 0.7)	(0.7, 0.7)
oxygen transport	(0.4, 0.2)	(0.7, 0.7)	(0.7, 1.0)
toxin	(0.7, 1.0)	(0.7, 0.2)	(0.2, 1.0)

Table 2. Sensibility for each functional class using different encoding methods

	lsa2	*hyd2*	*lsa2hyd2*
hydrolase	97.06%	93.75%	100.00%
immunoglobulin	97.14%	93.30%	100.00%
lyase	92.60%	94.12%	93.94%
oxygen transport	96.88%	100.00%	100.00%
toxin	96.30%	96.97%	94.29%

Table 3. Specificity for each functional class using different encoding methods

	lsa2	*hyd2*	*lsa2hyd2*
hydrolase	100.00%	96.88%	86.11%
immunoglobulin	96.55%	91.90%	96.77%
lyase	75.68%	83.33%	93.55%
oxygen transport	84.38%	96.67%	93.75%
toxin	86.49%	100.00%	93.10%

Table 4. Configuration of the ANN for classifying proteins in families

	lsa2	*hyd2*	*lsa2hyd2*
cyt b	(0.7, 0.7)	(0.7, 0.7)	(0.4, 0.0)
cyt c	(0.4, 0.5)	(0.4, 0.5)	(0.7, 0.2)
flav	(0.7, 0.0)	(0.4, 0.7)	(0.4, 0.5)
invert	(0.4, 0.5)	(0.4, 0.5)	(0.7, 0.5)

Table 5. Sensibility achieved for each family and encoding method

	lsa2	*hyd2*	*lsa2hyd2*	*[20]*
cyt b	97.73%	100%	97.73%	100%
cyt c	100%	100%	88.33%	100%
flav	100%	83.33%	88.89%	100%
invert	100%	100%	100%	*

From the results, it is not possible to point out a unique encoding method as the best for any functional class, considering both sensibility and specificity. For specificity, none of the encoding methods has been achieved more than 90% for all the classes. In general, the functional class *lyase* has presented smaller rates of specificity for the encoding methods **lsa2** and **hyd2**, with the best rate being 93.55% with the method **lsa2hyd2**.

For the protein classification in families, the method **lsa2** must be outlined by its sensibility, that has been achieved 100% for the three families and 97.73% for the *cytochrome b*. The family *invertebrate* was the unique to achieve 100% of sensibility independent of the encoding method used. In the specificity, the classification of the family *cytochrome b* did not changed for the different encoding methods, achieving 100% for all families.

For specificity, none of the encoding method has been outlined. However the **hyd2** can be considered the most stable, that can be observed analysing the family *cytochrome c*.

On the table 7 it is presented the rate of correct results obtained by our networks (**hyd2** and **lsa2hyd2**) and the ones obtained by [18] (**hyd**) and [21] (**lsa2**). For all functional classes and all encoding methods analyzed, our neural network presented better or equal results.

Comparing the results of classification of proteins based on families, it is verified that for the sensibility, our networks have presented similar results to those reported by [21], [20] and the later has presented best results for the sensibility (see table 5).

Table 6. Specificity achieved for each family and encoding method

	lsa2	*hyd2*	*lsa2hid2*	*[20]*
cyt b	100%	100%	100%	99.61%
cyt c	64.71%	92.31%	52.94%	99.95%
flav	84.62%	84.62%	92.31%	99.99%
invert	92.86%	92.86%	92.86%	*

Table 7. Correct results rate for the different ANN with different protein encoding method

	lsa2	*hyd2*	*lsa2hyd2*	*hyd*
hydrolase	100.00%	100.00%	100.00%	92%
immunog.	100.00%	100.00%	100.00%	96 %
lyase	95.31%	95.31%	100.00%	80%
oxygen	100.00%	100.00%	95.31%	84 %
toxin	100.00%	100.00%	100.00%	76%

6. Conclusion

The use of artificial neural networks has been considered as a promissor approach to classify proteins, once they are able to generalize after the learning process. In this work three different protein sequence encoding methods have been tested, as input of MLP neural network with the backpropagation algorithm. The results have confirmed the influence of the encoding method over the accurace of the ANNs as also have shown the ability of them in such task. The main contributions of this work were: to propose two methods of protein sequence encoding, to analyze ANNs with different encoding sequence and to achieve competitive results comparing the performance of the ANNs to the existed ones.

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] K. Blekas, D. I. Fotiadis, and A. Likas. Motif-based protein sequence classification using neural network. *Journal of Computational Biology*, 12(1):64–82, 2005.

[3] N. P. Carneiro, A. A. Carneiro, C. T. Guimarães, and E. Paiva. Desvendando o Código Genético. *Biotecnologia Ciência e Desenvolvimento*, (17):50–58, 2000.

[4] J. Chen and N. S. Chaudhari. Protein Family Classification Using Second-Order Recurrent Neural Networks. *Genome Informatics*, 14:520 521, 2003.

[5] M. Dayhoff, R. Schwartz, and B. Orcutt. Atlas of Protein Sequence and Structure. *Natl. Biomed. Res. Found.*, 5, 1978.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[7] E. A. Ferran and P. Ferrara. Clustering proteins into families using artificial neural networks. *Comput. Appl. Biosci.*, 8(1):39–44, 1992.

[8] M. T. Hagan, H. B. Demuth, and M. H. Beale. *Neural Network Design*. PWS Publishing Co., Boston, 1 edition, 1996.

[9] S. Haykin. *Redes Neurais: Princípios e Práticas*. Bookman, Porto Alegre, 2 edition, 2001.

[10] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, 12(2):95–107, 1996.

[11] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[12] J. Kyte and R. Doolittle. A Simple Method for Displaying the Hydropathic Character of Proteins. *Journal of Molecular Biology*, (157):105–132, 1982.

[13] B. Mahleko, A. Wombacher, and P. Fankhauser. Process-annotated Service Discovery facilitated by an n-gram based index. In *IEEE Intl. Conference on e- Technology, e-Commerce and e-Service*, pages 2–8, Washington, 2005. IEEE Computer Society.

[14] L. Rosseti. Célula e seus Constituintes Moleculares. In A. Zaha, editor, *Biologia Molecular Básica*, chapter 1, pages 10–25. Mercado Aberto, Porto Alegre, 2 edition, 2000.

[15] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Co, Porto Alegre, 1997.

[16] J. T. L. Wang, Q. Ma, D.Shasha, and C. H. Wu. New techniques for extracting features from protein sequences. *IBM Systems Journal*, 40:426–441, Jan. 2001.

[17] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. Application of neural networks to biological data mining: a case study in protein sequence classification. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 305–309, New York, NY, USA, 2000. ACM Press.

[18] W. R. Weinert and H. S. Lopes. Aplicação de um Sistema Neural ao Problema de Classificação de Proteínas. In *Anais do VI Congresso Brasileiro de Redes Neurais*, volume 1, pages 85–90, São Paulo, 2003.

[19] C. Wu, G. Whitson, J. McLarty, A. Ermongkonchai, and T.-C. Chang. Protein Classification artificial neural system. *Protein Science*, I:667–677, 1992.

[20] C. Wu, S. Zhap, H. Chen, C. Lo, and J. McLarty. Motif Identification Neural Design for Rapid and Sensitive Protein Family Search. *CABIOS*, 12:109–118, 1996.

[21] C. H. Wu, A. Ermongkonchai, and T.-C. Chang. Protein classification using a neural network database system. In *ANNA '91: Proceedings of the conference on Analysis of neural network applications*, page 29 41, New York, NY, USA, 1991. ACM Press.

[22] C. H. Wu, H. Huang, L.-S. L. Yeh, and W. C. Barker. Protein family classification and functional annotation. *Computational Biology and Chemistry*, 27:37–47, 2003.