Neural Network Approach to Locate Motifs in Biosequences*

Marcelino Campos and Damián López

Departamento de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia,
Camino de Vera s/n, 46022 Valencia, Spain
{mcampos, dlopez}@dsic.upv.es

Abstract. In this work we tackle the task of detecting biological motifs, i.e. subsequences with an associated function. This task is important in bioinformatics because it is related to the prediction of the behaviour of the whole protein. Artificial neural networks are used to, somewhat, translate the sequence of amino acids of the protein into a code that shows the subsequences where the presence of the studied motif is expected. The experimentation performed prove the good performance of our approach.

1 Introduction

The quantity of biological data is increasing each day. Processing of this data implies sometimes to detect certain subsequences (domains or motifs) with some functional features.

Coiled coil motif is involved in protein interaction. It is known the role of this motif in some biological processes such as protein transport and membrane fusions and the infection of cells by parasites [12][2].

Briefly, the coiled coil is an ubiquitous protein folding and assembly motif made of α-helices wrapping around each other forming a super-coil. Coiled coil motifs are usually made of seven-residue repeats $(abcdefg)_n$, called heptads, in which hydrophobic core occurs mostly at positions a and d. The interaction between two α-helices in a coiled coil involves these hydrophobic residues. Its simplicity and regularity results in a highly versatile protein interaction mechanism (see Figure 1). Furthermore, this is the most extensively studied protein motif.

Several programs for predicting coiled coil domains have been described. The most relevant to large-scale annotations are *coils* [7] (probably the most widely used), *paircoil* [1] and *multicoil* [13]. All these programs are based on the probability of appearance of every amino acid in each position of the characteristic heptad. This information is extracted from known coiled coil motifs and stored in a matrix. This approach is known as a PSSM approach (Position Specific Scoring Matrix). Multicoil is the most specialized one, and aims to detect double or triple coiled coil domains.

* Work supported by the Spanish CICYT under contract TIC2003-09319-C03-02.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 214–221, 2005.

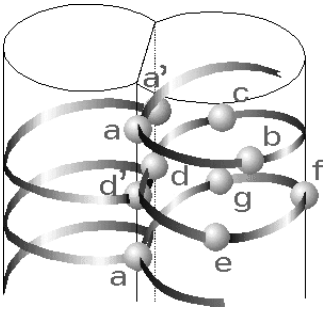

Fig. 1. We show a schematic representation of the coiled coil structure. Note that two α-helix are involved. Hydrophobic residues at the a and d positions are spatially close one each other because the helix structure. Their interaction result in a simple protein to protein fusion mechanism.

Lupas et al. [7] take into account that even very short proteins have stable coiled coils containing four or five heptads. The general scheme performs the analysis of the protein sequence using a sliding window of 21-35 amino acids. In that way, a score for each amino acid in the sequence of the protein is obtained using the probabilities stored in the PSSM. Berger's approach is the same but it considers correlation between amino acids where Lupas' consider probabilities of appearance. Berger et al. claim that the approach is useful to discard false positives detected by the Lupas' approach.

Hidden Markov Models and grammatical inference approaches has also been used in order to detect the presence of this motif [3,6,5]. Nevertheless, the problem of locating general coiled coil motifs is far of being solved. Several authors have noted several important coiled-proteins that are not detected when the previous approaches are used (among others, fusion-membrane proteins of the human and simian immunodeficiency virus or Ebola virus [10]). Thus, several other works propose solutions for more specific instances of the problem [11][10].

In our work, we use artificial neural networks to detect the subsequences which probably correspond to coiled coil. The experimentation carried out show that the performance of our approach is suitable for the task. This work is structured as follows: on section 2 we explain our neural net based approach and the process to select the parameters and topology of the net. Section 3 presents the experimentation that proves the validity of our approach. The conclusions of the work and some lines of future work end this paper.

2 Neural-Based Pattern Recognition

In our work we use Multilayer Perceptrons (MLPs). These neural nets are widely applied in pattern recognition tasks. For this purpose, the number of cells in the output layer is determined by the number of classes (C) involved in the task. In the same way, the input layer must hold the input patterns, and therefore the size

of this layer depends on the data representation. Classification is based on the creation of boundaries between classes. These boundaries can be approximated by hyperplanes. Each unit in the hidden layer(s) of MLPs forms a hyperplane in the pattern space. If a sigmoid activation function [9] is used, MLPs can form smooth decision boundaries which are suitable to perform classification tasks. The activation level of an output unit can be interpreted as an approximation of the a posteriori probability that the input pattern belongs to the corresponding class. Therefore, an input pattern can be classified in the class $i^\star$ with maximum a posteriori probability:

$$i^\star = \underset{i \in C}{\operatorname{argmax}} \Pr(i|x) \approx \underset{i \in C}{\operatorname{argmax}} g_i(x, \omega)\,,$$

where $g_i(x, \omega)$ is the i-th output of the MLP given the input pattern, x, and the set of parameters of the MLP, ω.

2.1 Input Data

In order to test our approach we used the *SwissProt Database* (release 40, April 2003). Each entry in the database contains the protein sequence and annotations for its known motifs (domains). Some of these motifs are annotated as *potential*, which means that have not yet been confirmed. We extracted from the database those entries corresponding non-potential coiled coil proteins, resulting in a set of 350 sequences (containing 720 coiled coil motifs).

Proteins are composed by a sequence of amino acids. When the amino acids are codified with one symbol, then protein sequences can be considered strings over an alphabet of 23 symbols: 20 amino acids, the glutamic and aspartic acids, plus a wildcard symbol. The wildcard symbol appear in the sequences whenever the true amino acid is not yet confirmed.

In order to standardize the input (the length of the proteins is not constant), the database was used to extract the set of segments of a given length (k). This parameter will closely determine the size of the input layer of the MLP. For each of these segments three output classes (three neurons in the output layer) were established in the following way:

- Class -1 whenever the segment does not overlap with a coiled coil motif
- Class 1 whenever the segment overlap but is not wholly contained in a coiled coil motif
- Class 2 whenever the segment is wholly contained in a coiled coil motif.

Three different numerical representations of the input data were tested. The first one considered the ordinal of each symbol, resulting in an input layer of k nodes. The second codification considered the symbols as a vector of 23 bits, obtaining an input layer of $23k$ nodes. The third codification is divided into two steps: first we used the Dayhoff codification (see Figure 1) to reduce the size of the input alphabet. Then we used the vector-based representation of the second representation. This option reduced the size of the input layer to $8k$.

Table 1. Dayhoff Amino acid codifications. This codification uses physic-chemical properties of the amino acids to group them into seven classes. Therefore, it is biologically justified.

amino acid	Dayhoff
C	a
A, P, G, S, T	b
N, Q, D, E	c
R, H, K,	d
L, V, M, I	e
F, W, Y	f
B, Z	g
X	x

2.2 Neural Network Topologies

The training of the MLPs was carried out with the software package *SNNS Stuttgart Neural Network Simulator* [14]. In order to properly use MLPs as classifiers we need to take some considerations. The more suitable input codification, the size of the input layer and the learning algorithm were studied in this order. To select the more suitable parameters, we randomly extracted 287 out of 350 sequences in the database to train different MLPs. The remaining 63 sequences were used to validate the resulting neural nets. The best results were obtained using the $23k$ codification, length of the segments $k = 28$ and backpropagation learning algorithm with learning rate of 0.1. Increasing number of nodes in the hidden layer of the MLPs (20, 40, 60, 80, 100, 200, 300 and 500 nodes), as well as MLPs with two hidden layers of 40 and 20 nodes respectively were also considered. Best results were obtained with one single hidden layer of 500 nodes.

3 Experimentation

For each test segment, it was expected that the MLP outputs 2 whenever the segment belonged with high probability to a coiled coil motif, -1 if the segment did not belong to a coiled coil motif and 1 otherwise. When a protein was analyzed, the different segments were processed sequentially. The output shows the appearance probability of a coiled coil motif (see Figure 2). In order to obtain statistically significant results, five balanced random partitions of the data were done (80% to train and 20% to test). Therefore, our final experiment entailed five runs obtaining a global 12.25% classification error rate. Confusion matrix is shown in Table 2.

3.1 Postprocessing

It is important to note that the result of any motif forecasting method ought to be confirmed in the laboratory. Thus, on the one hand it is important to reduce

Table 2. Confusion matrix for the final experiment

Classes	-1	1	2
-1	54.824 (89,36%)	3.752 (6,12%)	2.779 (4,53%)
1	4.045 (12,48%)	26.621 (82,13%)	1.747 (5,39%)
2	2.635 (5,61%)	2.312 (4,92%)	42.034 (89,47%)

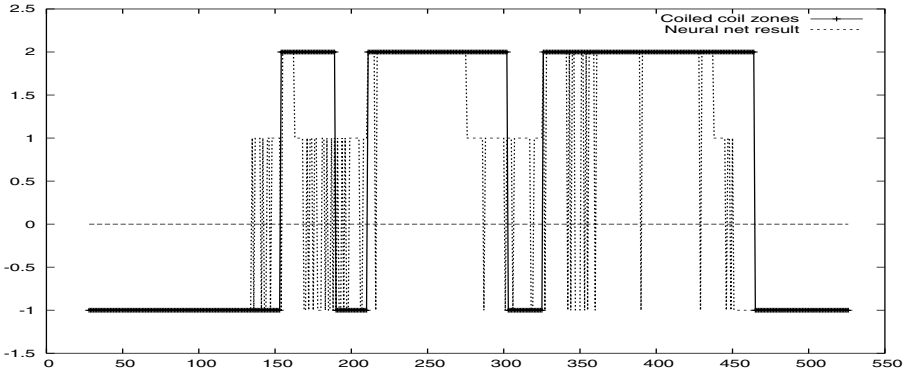

Fig. 2. Processing of a protein sequence. Note that the output fairly approximates the coiled coil database annotations.

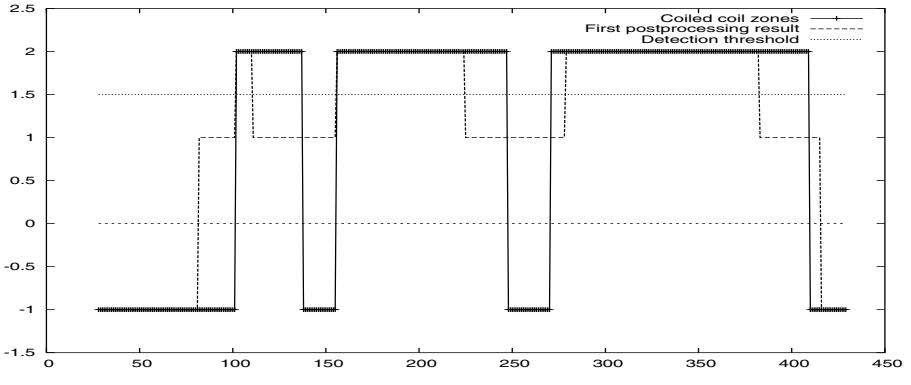

Fig. 3. Result of the first postprocessing method is shown. Note that in order a change to be considered, more than 7 consecutive amino acids with the same output are needed. The noise level of the postprocessed output is also highly reduced.

the rate of *false positive* detection. On the other hand, it is more important to roughly detect the more motifs the best rather than to accurately detect only some of them.

Before to analyze our approach in this way, we post-processed the output of our method. To do this, two procedures were tested.

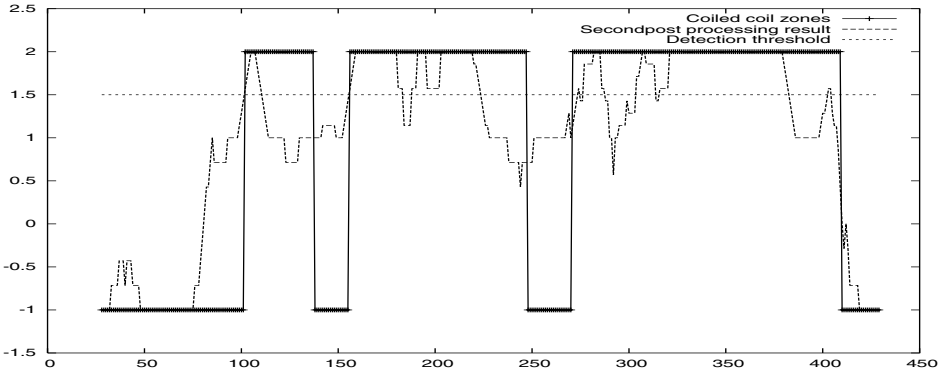

Fig. 4. Result of the second postprocessing method is shown. Note that there is no trouble with several predictions for a single coiled coil annotation.

The first one take into account that certain number of heptad repeats are needed to give stability to a coiled coil motif. Thus, the postprocessing did not considered those changes of length lower or equal to one heptad. Besides, this postprocess of the output reduced the noise level, because most of them was produced by very short predictions. Figure 3 shows the postprocessing of the output shown in Figure 2.

The second post-processing procedure is based on a smoothing of the output signal. To do this a one-heptad-length sliding window was considered to average the value of each output value. Figure 4 shows the postprocessing of the output shown in Figure 2.

In order to analyze the motif detection error, three categories were established:

- Error: Annotated coiled coil motif that has not been detected. The error detection rate is defined as the number of errors among the total number of coiled coil motifs in the database.
- False positive: prediction that overlap with no coiled coil annotation. Therefore, false positive detection rate is considered as the number of false predictions among the total number of coiled coil annotations in the database.
- Correct detection: Annotated coiled coil motifs that overlap with some coiled coil predictions.

Finally, we considered as a coiled coil prediction those regions with output over a value of 1.5, that is, the average between the probable and high probable coiled coil output. The results obtained are shown in Table 3. In order to compare our results with the most known prediction algorithms [7][1], we run available versions of the algorithms ([8][4] respectively) using the default parameter values. The results are also shown in Table 3.

The results obtained differ from each other considering the postprocessing procedure. On the one hand the first postprocessing procedure produces very

Table 3. Comparative experimental results. The error rate and the false positives detection rate is shown for each method tested.

	error rate	FP
MLP (1st postprocessing)	22,50%	1,80%
MLP (2nd postprocessing)	7,50%	14,44%
coils	19,30%	15,83%
paircoil	21,38%	10,27%

low rate of false positive detection. This reduction does not produce significant increase of the error rate (it is quite similar to the Lupas' and Berger's methods error rate). The second postprocessing procedure obtains better error rate than any other approach (the false positive rate is also similar to the Lupas' and Berger's methods rate).

4 Conclusions

We propose a neural net based method to detect coiled coil motifs from biosequences. This motif is related to protein interaction. Motif location is important in bioinformatics because it is related to the prediction of the behaviour of the whole protein.

MLPs are used to, somewhat, translate the sequence of amino acids of the protein into a code that shows the subsequences where the presence of the studied domain is expected. The output of the neural net is then postprocessed to obtain a motif location forecast. Two postprocessing procedures were tested. The behaviour of these procedure were different one each other.

In any case the results are improved respect previous prediction methods. This is proved by the experimentation carried out. We can select the postprocessing to obtain very low rate of false positive detection or low rate of error detection. The reduction of false positive rate is highly biologically demanded because it reduces the experimental effort. Comparison with two well-known coiled coil prediction algorithms [7][1] is shown.

It is very important to note that the database contains annotations only for those proteins that contain a coiled coil region. Furthermore, it is not assured that the coiled coil motifs are accurately annotated. Furthermore, there not exist any negative annotation, that is, information concerning non-coiled coil subsequences. This have to be considered as an important drawback. Of course, the availability of non-coiled protein sequences should improve our results. This sequences could be obtained by considering protein structural information.

Coiled coil is a well characterized motif. Its structure is the key stone of the most used prediction algorithms. MLPs could also be used to predict the location of other motifs whose structure is poorly known.

References

1. B. Berger, D.B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicting coiled coils by use of pairwise residue correlation. Proc. Natl. Acac. Sci., 92:8259 8263, 1995.
2. D.C. Chan and P.S. Kim. Hiv entry and its inhibition. Cell, 93:681684, 1998.
3. M. Delorenzi and T. Speed. An hmm model for coiled-coil domains and a comparison with pssm-based predictions. Bioinformatics, 18(4):617625, 2002.
4. PAIRCOIL implementation by the authors, 1995. http://theory.lcs.mit.edu/bab/computing.
5. D. Lopez, A. Cano, M. Vazquez de Parga, B. Calles, J.M. Sempere, T. Perez, M. Campos, Jose Ruiz, and Pedro Garcia. Motifs discovery by k-tss grammatical inference. 2005. Submitted and accepte at the 19th IJCAI'05.
6. Damian Lopez, Antonio Cano, Manuel Vazquez de Parga, Belen Calles, Jose M. Sempere, Tomas Perez, Jose Ruiz, and Pedro Garcia. Detection of functional motifs in biosequences: A grammatical inference approach. In X. Messeguer and G. Valiente, editors, Proceedings of the 5th Annual Spanish Bioinformatics Conference, pages 7275. Univ. Politécnica de Catalunya. ISBN: 84-7653-863-4, 2004.
7. A. Lupas, M. Van Dyke, and J. Stock. Predicting coiled coild from protein sequences. Science, 252:11621164, 1991.
8. Source Code NCOILS, 1999. http://www.russell.embl.de/cgi-bin/coils-svr.pl.
9. D. E. Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton. Schemata and sequential thought processes in pdp models. In J. L. McClelland, D. E. Rumelhart, others, and eder, editors, Parallel Distributed Processing: Volume 2: Psychological and Biological Models, pages 757. MIT Press, Cambridge, MA, 1986.
10. M. Singh, B. Berger, and P.S. Kim. Learncoil-vmf: Computational evidence for coiled-coil-like motifs in many viral membrane fusion proteins. J. Mol. Biol., 290:10311041, 1999.
11. M. Singh, B. Berger, P.S. Kim, J.M. Berger, and A.G. Cochran. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. Proc. Natl. Acac. Sci., 95:27382743, 1998.
12. J.J. Skehel and D.C. Wiley. Coiled coils in both intracellular vesicle and viral membrane fusion. Cell, 95:871874, 1998.
13. E. Wolf, P.S. Kim, and B. Berger. Multicoil: a program for predicting two- and three-stranded coiled coils. Protein Science, 6:11791189, 1997.
14. Andreas Zell, Niels Mache, Ralf Huebner, Michael Schmalzl, Tilman Sommer, and Thomas Korb. SNNS: Stuttgart neural network simulator. Technical report, Stuttgart, 1992.