SELENOPROTEIN DISCOVERY USING NEURAL NETWORKS

Miguel Lupi Alves
Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
mlupi@netcabo.pt

Carlos Lourenço
Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
and
Centro de Lógica e Computação, Departamento de Matemática, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
csl@di.fc.ul.pt

ABSTRACT

Identifying SECIS sequences in the genome is an essential step in Selenoprotein discovery. We use a competition neural network in a pre-processing stage to encode potential SECIS sequences, and a multilayer perceptron trained with backpropagation to predict SECIS. We thus propose a new approach within SECIS studies by combining more classical bioinformatics techniques with the learning and generalization capabilities of neural networks. Selenoproteins are thought to be responsible for most biomedical effects of dietary selenium and are essential to mammals. They play an important role in cancer prevention, immune function, aging, male reproduction, and other physiological and pathophysiological processes. Their prediction and study are of undeniable value. The ability of artificial neural networks to be fault tolerant, to adapt easily to new environments without specialized programming and to handle fuzzy, probabilistic, noisy or inconsistent data allowed us to provide an alternative to stricter pattern matching and score based methods. Our results compare well with the ones obtained with previously available methods.

KEYWORDS

Neural Networks, Bioinformatics, Pattern Classification, SECIS, Selenocysteine, Selenoproteins

1. INTRODUCTION

Selenium is a micronutrient found in proteins on the eubacterial, archaeal and eukaryotic domains of life. It is present in selenoproteins in the form of Selenocysteine (Sec), the 21st amino acid (Hatfield and Gladyshev, 2002).

It is known that supplementation of the human diet with Selenium potentially offers effective means of preventing or diminishing human maladies. Dietary Selenium plays an important role in cancer prevention, immune function, slowing the aging process, male reproduction and other physiological and pathophysiological processes (Kryukov et al., 1999).

Selenium deficiency results in decreased levels of selenium-containing proteins (Kryukov et al., 2003) which are associated with decreased survival rates of HIV-infected patients (Kryukov et al., 2003), white muscle disease –a degenerative disease of the cardiac and skeletal muscles, mostly a veterinary problem but seen occasionally in malnourished people– and endemic juvenile cardiomyopathy (Gu et al., 1997). In human clinical trials, increasing the dietary supplement of Selenium reduced the rate of prostate, lung and colon cancer incidence in 48-63%. Selenium is also essential for mammalian development. For example, disruption of the mouse Sec-tRNA gene results in early embryonic lethality (Kryukov et al., 1999).

It should be noted that the functions of the majority of selenoproteins are not known. Characterization of their function is an obvious direction in selenoprotein research. (Hatfield and Gladyshev, 2002)

Unfortunately, it is especially difficult to predict selenoproteins, since Selenocysteine is the only amino acid genetically coded. The fact that UGA serves as a stop as well as Sec codon, raises an important question: How can the cells' translational machinery distinguish these two functions? Something has to tell the translational machinery of the cells to continue and not terminate translation at the UGA codon. This is the function of the stem-loop structure in the 3' untranslated regions of eukaryotic mRNAs that encode selenoproteins, known as Selenocysteine Insertion Sequence (SECIS) elements (Kryukov et al, 2003; Castellano et al., 2001).

Gene prediction programs rely on the standard stop codons. Because of the non-standard use of the UGA codon, they are unable to identify selenoproteins (Castellano et al., 2001). One solution for this problem could be to search for occurrences of SECIS structural patterns (Castellano et al., 2004). But these patterns are too common in genomic material and searches produce large number of predictions (Castellano et al., 2001 and 2004). The distance between a SECIS element and a UGA codon influences whether the UGA is read as a stop or a Selenocysteine codon. However, this is not the only factor that affects the interpretation (Gu et al., 1997). Yet another example is the fact that rodent and sheep selenoprotein W and *Schistosoma* GPX mRNAs are known to use UGA codons to specify both selenocysteine and termination (Gu et al., 1997). However, SECIS still provides an identifier that can help in the annotation of uncharacterized selenoprotein genes (Hatfield and Gladyshev, 2002).

As depicted in Fig. 1, SECIS elements are composed of two helices separated by an internal loop. An additional ministem may appear, if the apical loop is large enough. This seemingly stabilizes the SECIS element. Furthermore the SECIS elements are classified as form 1, if without ministem, and form 2 otherwise. The SECIS core structure, Quartet, is located at the base of the second helix. Preceding it, in the majority of cases, we can find in the apical loop two Adenosine nucleotides (Hatfield and Gladyshev, 2002).

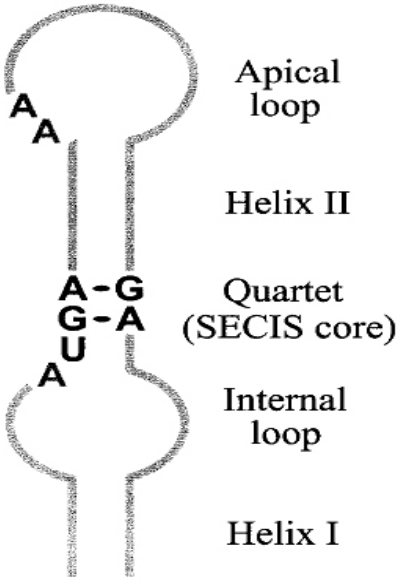

Figure 1: SECIS element consensus structure (Hatfield and Gladyshev, 2002).

SECIS elements have different structures in eukaryotes, archaea and bacteria. Nevertheless, when searching for potential Selenoprotein genes, it is known that the SECIS elements are located in the 3´-untranslated regions of all eukaryotic and archaeal Selenoprotein genes. In bacteria they are located in the coding regions immediately downstream of the UGA codons (Hatfield and Gladyshev, 2002).

The Selenoprotein prediction protocol using classical methods is illustrated in Fig. 2.

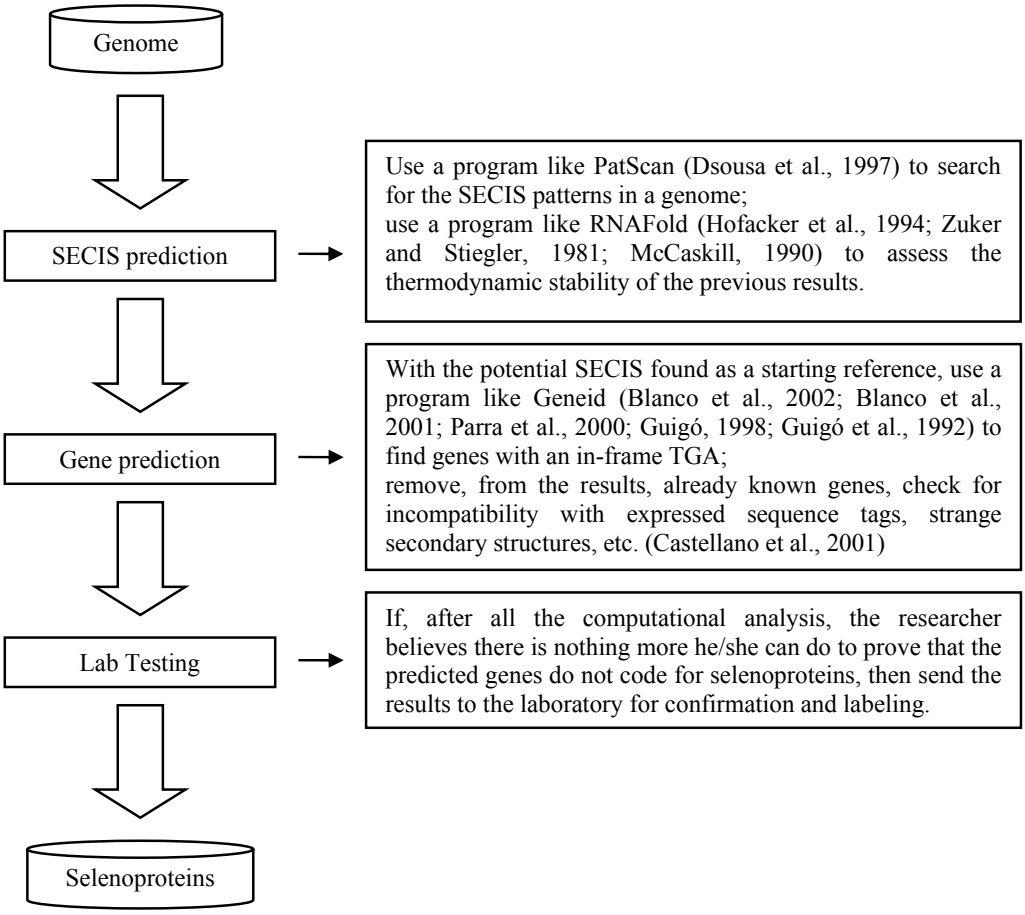

Figure 2: Selenoprotein prediction protocol overview.

For the computational prediction of selenoproteins the first step is to find potential SECIS by providing PatScan with a pattern and sequence files in FASTA format (Pearson et al, 1988). Since a mere pattern is not sufficient to identify positively a SECIS element and the SECIS patterns are too common in genomes, it is necessary to use other methods to improve PatScan results, such as applying RNAFold to the PatScan output in order to assess the thermodynamic stability of each potential SECIS, using the protocol described in (Kryukov et al., 1999). The thermodynamically stable SECIS location and Geneid are used to predict genes that may be interrupted by in-frame TGA codons. The putative distance between the SECIS element and the potential selenoprotein coding gene has changed over time. For *Drosophila Melanogaster* 500bp downstream was used in (Gu et al., 1997), but this may be too restrictive. In mammals, distances larger than 4000bp are possible (Castellano et al., 2001). From the genes found, one removes those that are already known, those that are incompatible with expressed sequence tags, those that have unlikely structures and anything that to the best of your knowledge cannot possibly code a selenoprotein. Whatever is still left in the end may be a selenoprotein and should request further analysis in a laboratory.

The first step of the protocol presented in Figure 2 still results in many false predictions, revealing that the simple thermodynamic evaluation may not be enough to restrict the false predictions to a minimum possible, and may even be excluding some potential SECIS that should suffer further investigation.

The complexity of this problem demands a system that is robust, fault tolerant, easily adaptable to new environments without specialized programming, and which can deal with fuzzy, probabilistic, noisy or inconsistent data. In view of this, we have used the neural computation paradigm to build a successful alternative in the prediction of SECIS elements. The neural network approach does not necessarily replace RNAFold usage, but it may complement it. Indeed, the neural protocol would enter the diagram of Fig. 2 at about the location of RNAFold in that diagram.

2. METHODS

2.1 Materials

The materials used were:

a) 35876 potential SECIS extracted from the first 19 large genome scaffolds of *Drosophila Melanogaster* (Castellano et al., 2001);
b) 1415 thermodynamically stable potential SECIS extracted from the previous ones using RNAFold (Castellano et al., 2001; Hofacker et al., 1994; Zuker and Stiegler, 1981;McCaskill, 1990);
c) 267 several species real SECIS from RFam (Griffiths-Jones, 2003);
d) 14182 potential SECIS extracted from the fourth revision of the *Drosophila Melanogaster* genome, the latest version at this time; the genome was obtained from the Berkeley *Drosophila* Genome Project web site (http://www.fruitfly.org/);
e) 46625 potential SECIS extracted from the human mRNA assembled at GenBank and available at the UCSC Genome Bioinformatics site (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/).

The potential SECIS on d) and e) were extracted following the recommendations expressed in (Castellano et al., 2001) using PatScan (Dsouza et al., 1997) with the patterns:

r1={at,ta,gc,cg,tg,gt} p1=5...15 p2=1...7 a tga n p3=9...12 p4=0...2 aa p5=6...17 r1~p3[2,1,1] n gan p6=3...9 r1~p1[2,1,1]

2.2 Test Set Preparation

Two types of tests were made, one using a test set —herein called abridged test set— with just 9600 sequences mainly extracted from the first version of the *Drosophila Melanogaster* genome, and the other with the latest releases of full genomes. A slight variation of the abridged test set was also considered originally, consisting in deleting from it the 267 real SECIS from several species. This shorter test set was also used in the first type of tests.

The abridged test set was build with:

- 1415 *Drosophila Melanogaster* thermodynamically stable SECIS;
- 3 real *Drosophila Melanogaster* SECIS;
- 7915 *Drosophila Melanogaster* thermodynamically unstable SECIS;
- 267 real SECIS from several species, where the 3 real *Drosophila Melanogaster* SECIS appear once again, but actually originate 4 entries since one of the *Drosophila M.* SECIS is associated with two different selenoproteins.

The abridged test set and the test sets built from full genomes were split into 14 columns. Each column represents one of the PatScan patterns and originates a test sub-set.

From these 14 test sub-sets the third and fourth are discarded since they only contain fixed values. The first and last test sub-sets are also discarded since the results of thousands of tests revealed us a negative

impact on the neural networks predictions. This is probably because the beginning and end of the SECIS sequences change far too much.

Using a small script, the nucleotide symbols of the remaining 10 test sub-sets are then substituted by the Hydrogen bonding symbols, S for strong bondings and W for weak ones. A standard sparse coding is adopted at this early encoding stage.

This still results in an input vector that is too large, affecting the learning ability of the neural network used in the prediction. Furthermore, the data must be in an adequate numerical format. Hence, the 10 test sub-sets are fed into 10 independent competition neural networks (see e.g. pages 217-221 of Hertz et al., 1991) for encoding and compression using GenEncode (a competition neural network program created for this work ; available on request) with the following configuration per network:

PatScan Pattern	Qty.Neurons	Learning Epochs
2	2	30
5	2	30
6	1	30
7	2	30
8	1	30
9	2	30
10	2	30
11	2	30
12	2	30
13	2	30

All the remaining program options were left with default settings.

GenEncode produces 10 encoded test sub-sets. The encoded values represent the distances between the final neuron weight vectors and each of the input vectors. If a pattern features sequences with up to 14 nucleotides in length, and these are represented by only 2 neurons, this provides a high compression ratio.

Using JoinSECIS (a script used to join the text columns from different files into a single one; available on request) we join all the 10 encoded test sub-sets into a single test set, the final one, which is to be fed into a multilayer perceptron (MLP).

With unsupervised learning, the competition neural network must find for itself features, regularities, correlations, or categories in the input and code for them in the pre-processed output which is the MLP's input.

Unsupervised learning is only useful here because there is redundancy in the data. Otherwise data would look like random noise (see pages 197-199 of Hertz et al., 1991).

2.3 SECIS Predictions

The power of multilayer networks was realized long ago (see page 115 of Hertz et al., 1991), since they have the capacity to compute a wide range of functions (see pages 149-171 of Rojas, 1996) and are able to generalize over training examples both in approximation and classification tasks (see e.g. Anderson, 1996). The lack of a good training method for these networks and the demonstration of the limitations of single-layer neural networks was a significant factor in the decline of interest in neural networks in the 1970s. The discovery by several researchers independently and widespread dissemination of an effective general method of training a multilayer neural network, played a major role in the reemergence of neural networks as a tool for solving a wide variety of problems. The main method is known as Backpropagation (of errors) (Hertz et al., 1991; Rojas, 1996; Fausett, 1994).

The Backpropagation algorithm looks for the minimum of an error function in neuron weight space using the method of gradient descent. A combination of weights which minimizes the error function is considered to be a solution of the learning problem. Since this method requires computation of the error function's

gradient at each iteration, we must guarantee the continuity and differentiability of the error function (see pages 149-171 of Rojas, 1996).

We use a multilayer perceptron in the task of predicting which input vectors −originated from DNA or RNA sequences that have been selected through a previous pattern matching process− correspond to real SECIS elements. Recall that the input to the MLP is prepared with the help of a separate competition network. As explained in Section 2.2, the input vectors to the MLP are sequences from the encoded set produced by GenEncode.

The Backpropagation learning algorithm is used to train the MLP. The actual architecture we adopt for the MLP is illustrated in Fig. 3.

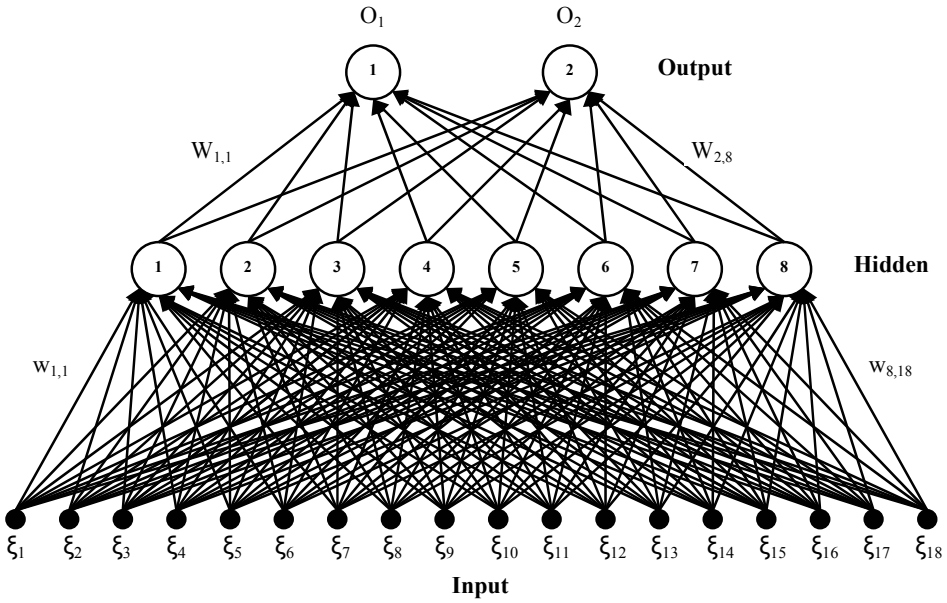

Figure 3: Multilayer Perceptron architecture. Here ξ is the input vector, w and W are connection weights, and O is the output vector. For clarity, only a few connections are labeled by the respective weights. Sigmoidal activation functions are used for all neurons.

During training, the network is taught to respond ($O_1 = 0$, $O_2 = 1$) if the input stimulus' target is a SECIS, and ($O_1 = 1$, $O_2 = 0$) otherwise. At each iteration, the connection weights are updated according to the Backpropagation algorithm. Once learning has stopped, the weights become fixed and the network can be tested for classification performance. For each input stimulus, either seen during training or not, the respective output is given by the activities of neurons O_1 and O_2. These outputs are then subjected to a Euclidean distance calculation between them and the two possible "perfect" answers, (0,1) for predicted SECIS, (1,0) otherwise. The smallest distance wins, meaning that if the output is closer to (0,1) then the input vector is considered to correspond to a SECIS; otherwise it is classified as non-SECIS.

All scripts and programs, as well as the neural networks described, are available on request.

3. RESULTS

The selenoprotein prediction protocol presented in Figure 2 was developed over the years through international cooperation of several researchers, and features what are believed to be the best tools for the task. Therefore when we compare our results with those of RNAFold we are comparing them with the best available. The two methods are quite different, since RNAFold is used to assess the thermodynamic stability of the sequences while the MLP is used to classify the sequences through biological knowledge acquired

from the sequences during the learning process. A direct comparison between the two cannot be fully undertaken with presently available data. Our evaluation falls on which method provides the least amount of false positive predictions while finding all the real SECIS as well as on the possibility of improving the existing results by finding new real SECIS that had been discarded by previous methods.

Prior to running the learning algorithm, learning parameters have to be chosen. In our implementation, we used the following default settings: learning rate of 0.1; momentum 0.9; uniform [0,1] weights initialization. Furthermore, learning sets have to be prepared. This is done following the same procedure as outlined in Section 2.2 for the test set preparation.

Reaching the proper encoding procedure, as presented in Section 2.2, and neural network architecture, as presented in Section 2.3, was a computational intensive process. Hence we started to test the learning procedure with a sample of 9333 potential SECIS, the same as described in Section 2.2 for the abridged test set but without the 267 real SECIS from several species, from the materials used in (Castellano et al., 2001). These were used as a first test set.

With a learning set constituted by 263 real SECIS from several species, not including *Drosophila M.*, and 300 thermodynamically unstable SECIS (as predicted by RNAFold) from *Drosophila M.*, the MLP was able to predict 773 potential SECIS from within this first test set, including the 3 real ones from *Drosophila M.*. For comparison, RNAFold predicted 1415 SECIS, also including the 3 real ones from *Drosophila M.*.

However, given that until now only three real SECIS were found on *Drosophila M.* and therefore this constitutes a rather thin statistical argument, we decided to add 267 real SECIS from several species into this sample, creating a new test set with 9600 sequences as described in Section 2.2 ―the so-called abridged test set. Using this new test set and a new learning set constituted by 120 thermodynamically stable SECIS from *Drosophila M.* (as predicted by RNAFold), not including the 3 real *Drosophila M.* SECIS, and 300 thermodynamically unstable SECIS (as predicted by RNAFold) from *Drosophila M.*, the MLP was able to predict 1155 potential *Drosophila M.* SECIS including the 3 real ones from *Drosophila M.* and 194 real SECIS from several species. This means that 74% of the real SECIS on the test set were found while keeping a low rate of false predictions.

Furthermore, in both tests around 40% of the potential SECIS predicted by MLP are different from those predicted by RNAFold, revealing in this way a new space with potential for discoveries.

We have also started to test the encoding and neural network techniques with full genomes. Using a learning set constituted by 263 real SECIS from several species, not including *Drosophila M.*, and 300 thermodynamically unstable SECIS (as predicted by RNAFold) from *Drosophila M.*, with the following results:

- 2248 potential SECIS predicted against 1415 predicted by RNAFold on the first release of Drosophila M. genome;
- 904 potential SECIS predicted on the latest version of *Drosophila M*. genome (4th release);
- 3146 potential SECIS predicted by MLP on the latest version of the human genome (see Section 2.1).

Using a learning set constituted by 280 thermodynamically stable SECIS (as predicted by RNAFold) and 300 thermodynamically unstable SECIS (also RNAFold-predicted) from the first version of the *Drosophila M.* genome, applied to the abridged test set, 1726 input vectors were predicted as SECIS; among them, 224 were from other species (not *Drosophila M.*). Note that 267 real SECIS from other species were included in the representative test set.

Our tests were iterated several thousand times with varying learning parameters and starting conditions, but only a selection of the most interesting results is shown here. All other results are available on request.

This is ongoing work, and therefore we expect to release more detailed results and comparisons as soon as possible. However, the work done until know already allows some insight on the potential of this method. We are currently applying RNAFold to potential SECIS extracted from other genomes in order to improve our comparison with MLP. We are also awaiting lab results.

4. CONCLUSION

The results show that the Multilayer Perceptron is able to ascertain the biological information needed to provide good predictions from the encoded input data. Furthermore, a detailed analysis reveals that a very significant amount of the predictions do not match those from RNAFold, which increases the potential for new discoveries, while remaining low in the quantity of false positive SECIS predictions. Even though for the first version of the *Drosophila M.* genome the MLP generated more false predictions than RNAFold, those predictions were qualitatively different. Furthermore, the first version of the *Drosophila M.* genome is very different from the latest release, since it has twice as much sequences as the latest release. This suggests that there may be many errors in it that could have misled the MLP. The tests with the latest version of the *Drosophila M.* genome are much more promising and we continue to carry them on.

Regarding cross-species predictions, the MLP taught with only *Drosophila M.* SECIS and non-SECIS (as predicted by RNAFold) was still able to find 74% of the real SECIS from the RFam database which contains other species (see Griffiths-Jones, 2003, for the database). Notably, by relaxing the obligation of finding all 3 real SECIS from *Drosophila M.*, the MLP actually found 84% of the real SECIS from the RFam database. In the latter case, only 2 of the 3 *Drosophila M.* SECIS were detected. Reciprocally, the usage of other species' real SECIS in the learning process of the MLP to identify *Drosophila M.* sequences also revealed the good potential of this method for cross-species predictions. The neural networks shown here brought out their potential to outperform RNAFold in the prediction of potential SECIS even when taught by the results of RNAFold itself.

The potential of neural networks for assisting in the identification of new selenoproteins is evident, but there is still much work to be done. Encoding improvements, more biological information about the potential SECIS sequences, better understanding of what the neural network uses from the potential SECIS sequences, and multilayer perceptron learning process improvements, are just a few aspects that can still be looked at. As SECIS research proceeds at a steady pace, we should also expect more feedback from the labs in the process of assessing the quality of the bioinformatics predictions described above.

ACKNOWLEDGEMENTS

We thank Roderic Guigó, Sergi Castellano, Enrique Blanco and Charles Chapple from the Genome Bioinformatics Lab, Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra, Barcelona, Spain, for all their assistance and materials provided.
C. L. acknowledges the partial support of Fundação para a Ciência e a Tecnologia and EU FEDER via the Center for Logic and Computation and the project ConTComp (POCTI/MAT/45978/2002).

REFERENCES

Book

Anderson, J.A., 1996. *An Introduction to Neural Networks*. The MIT Press, Massachusetts, USA.

Fausett, L., 1994. *Fundamentals of Neural Networks, Architectures, Algorithms, and applications*. Prentice Hall, Upper Saddle River, New Jersey, USA.

Hertz, J. et al., 1991. *Introduction to the theory of neural computation*. Perseus Books, Massachusetts, USA.

Rojas, R., 1996. *Neural Networks - A Systematic Introduction*. Springer-Verlag, Berlin, Germany.

Journal

Adams et al., 2000. The Genome Sequence of Drosophila Melanogaster. *Science*, Vol. 287, pp 2185-2195.

Blanco et al., 2002. Using geneid to Identify Genes. *Current Protocols in Bioinformatics*, unit 4.3.

Castellano et al., 2001. In silico identification of novel selenoproteins in the *Drosophila Melanogaster* genome. *EMBO reports*, Vol. 21, No. 81, pp 697-702.

Castellano et al., 2004, Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO reports*, Vol. 5, No. 1, pp 71-77.

Dsouza et al., 1997. Searching for patterns in genomic data. *TIG*, Vol. 13, No. 12, pp 497-498.

Griffiths-Jones, S., 2003. Rfam: an RNA family database. *Nucleic Acids Research*, Vol. 31, No. 1, 439-441.

Guigó, R., 1998. Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, Vol. 5, pp 681-702.

Guigó et al., 1992. Prediction of gene structure. *Journal of Molecular Biology*, Vol. 226, pp 141-157.

Gu et al., 1997. Conserved features of Selenocysteine insertion sequence (SECIS) elements in selenoprotein W cDNAs from five species. *Gene*, Vol. 193, pp 187-196.

Hatfield, D. L. and Gladyshev, V. N., 2002. Minireview, How selenium has altered our understanding of the genetic code. *Molecular and Cellular Biology*, Vol. 22, No. 11, pp 3565-3576.

Hofacker, et al., 1994. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, Vol. 125, pp 167-188.

Kryukov et al., 1999. New Mammalian Selenocysteine-containing Proteins Identified with an Algorithm that Searches for Selenocysteine Insertion Sequence Elements. *The Journal of Biological Chemistry*, Vol. 274, No. 48, pp 33888-33897.

Kryukov et al., 2003. Characterization of Mammalian Selenoproteomes, *Science*, Vol. 300, 30 May.

McCaskill, J.S., 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, Vol. 29, pp 1105-1119.

Parra et al., 2000. Geneid in Drosophila. *Genome Research*, Vol. 4, No. 10, pp 511-515.

Zuker, M. and Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucl Acid Res*, Vol. 9, pp 133-148.

Conference paper or contributed volume

Blanco et al., 2001. Gene Prediction in Post-Genomic Era. *IXth ISMB (Poster)*, Copenhagen, Denmark.

Pearson et al, 1988. Improved Tools for Biological Sequence Analysis. *Proceedings of the National Academy of Sciences*, USA. Vol. 85, pp 2444-2448.