Computational Biosequence Analysis by Neural Networks

S. Brunak and J. Engelbrecht

Department of Physical Chemistry
and
Computational Neural Network Center (CONNECT)
Building 206
The Technical University of Denmark
DK-2800 Lyngby, Denmark

Abstract

The use of artificial neural networks for biological sequence analysis has recently been strongly intensified. This paper describes work on the difficult coding/non-coding classification problem in human DNA — an important step in the information processing of the living cell. The network approach was utilized not only for generalization purposes, but also as a tool for obtaining knowledge about previously unknown local features of the DNA sequence.

1 Introduction

The study of molecular computing engines based on the analysis of biological sequences of chemical symbols — one dimensional descriptions of proteins, RNA and DNA — is faced with the accumulation of quantitative information at an ever increasing rate. The chief source of this information is DNA sequencing and — thereby — the associated amino acids in proteins; but data sets covering functionality and three-dimensional structure of macromolecules are also growing rapidly.

Considerable advances in the development of software and hardware capable of mapping, analyzing and comparing complex sequences of chemical letters — including those of bacteria, yeast, and man — are needed to keep pace with the enormous expansion in the databases produced by the new efficient experimental techniques of molecular biology. The impact of modern methods of sequence analysis has generated an environment in which computer–based methods of analysis form an integral and vital part of the research process.

In the last five years the use of non-linear neural networks[9] for understanding and modeling the relation between the symbolic content of biological sequences and macromolecular structure and function has been strongly intensified. Complex biological mechanisms are often very non-linear: small changes in the content of chemical components may cause large changes in the products of reaction chains and in molecular function and form. Neural networks are in general mostly used due to their ability to generalize, that is to 'recycle' non–linear regularities in a training set of examples to new cases. The utilization of neural networks is often questioned because of the difficulty in extracting knowledge from them following training. However, as shown below it is possible to 'invert' the trained network and in the pattern of its adjustable parameters to identify what sequence features are related to the specific classification task carried out by the network. By monitoring the training process it is also possible to detect abnormal examples deviating strongly from prevailing patterns in the training set. Abnormal examples may arise either from the application of weak classification strategies, or simply due to classification errors introduced randomly devoid of any systematics at all. In addition to errors caused by simple misprints in the databases, the network method has been able to detect errors caused by incorrect interpretation of experiments.

2 Removal of non-coding regions in human pre-mRNA

Most eukaryotic genes contain several non–coding regions, (see Green, 1986 for a review, or the contribution from T. Schneider in this volume). In the nucleus these regions, known as introns, are excised from the pre–mRNA and the mature mRNA is formed by concatenation of the coding regions, the exons. Subse-

0-8186-3420-0/93 $03.00

quently the mRNA leaves the nucleus to be translated according to the genetic code into sequences of amino acids. In a computer or compiler analogy this removal corresponds to the stripping of comments from ordinary computer programs written in high level languages such as Fortran or C prior to the generation of the machine code. In these computer languages the beginning and end of comments are unambiguously defined by symbol patterns such as `C`, `/*` and `*/`. In the cell the splicing process recognizes sequences at the exon–intron and intron–exon borders and is mediated by a group of small nuclear RNAs (snRNAs) complexed with protein as ribonucleoprotein particles (snRNPs), [11, 7]. In the pre–mRNA sequence special regions hold very ambiguous information on the location of the introns or 'comments'. The exon–intron border, the donor site, can be characterized by a consensus sequence[13], $\frac{C}{A}$AG/GT$\frac{C}{A}$AGT. However, this consensus sequence is insufficient in distinguishing, with any reasonable degree of accuracy, between potential sites selected by the splicing machinery and those which are not. When the splicing process proceeds in the cell donor sites may therefore be defined by additional sequence characteristics recognized by the snRNPs or other known or unknown factors. The intron–exon border is even more weakly defined by a consensus sequence[8, 10]. It is characterized by a 10-50 bp long polypyrimidine tract with strong overabundance of nucleotides `C` and `T` and very few `AG` dinucleotides. An `AG` dinucleotide terminates the intron.

The general problem of classifying fragments of DNA sequence according to its function has been addressed by several methods, primarily either based on the compilation of tables of codon usage[16, 17], or detection of local non-randomness in the sequence[5, 15]. This problem is far easier when sequence fragments are "pure", that is, when they in their entirety stem from one specific class only. If the objective is to locate introns in unannotated DNA a high degree of accuracy in regions of transition between the different categories is of prime importance.

Recently a neural network method for prediction of splice sites in human pre-mRNA has been published[3, 1, 2]. This method combines a local search for potential donor and acceptor sites with a global evaluation of jumps in the coding/non-coding signals. By suppressing splice site assignments in regions of constant "exon-ness" a large number of false positive assignments in the interior of exons and introns could be eliminated, and likewise in regions of sharp transition in the coding/non-coding signal weak splice site assignments could be enhanced. Compared to conventional weight matrix methods the ratio between true and false positives could be improved by an order of magnitude. This work presents an analysis of the classification strategies used by the network with the aim of revealing differences in the local and global sequence characteristics of coding and non-coding DNA.

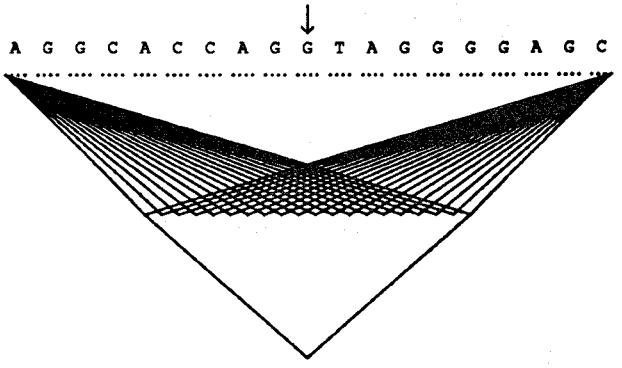

Figure 1: A small version of the feed-forward neural network used for classifying the middle nucleotide in one of two categories: coding or non–coding. The particular architecture shown has nineteen nucleotides visible to it in the input layer, twenty non–linear hidden units and one non-linear output unit.

3 Inverting the coding/non–coding network

The neural network used was of the feedforward type[12, 9], see Figure 1. It was equipped with an input layer scanning the sequence of nucleotides, a hidden layer, and an output layer delivering classifications of the nucleotide configurations in the window. The sliding window covered 301 bp of sequence, based on which the network classified the central nucleotide in one of two categories: coding or non-coding. To each letter in the window four units were associated and connected to 200 hidden units which in turn were connected to one output unit. The coding/non-coding network was very large, and contained a total of 241,201 adjustable parameters.

The functional steps of the network operation are as follows: The input layer reads a sequence fragment

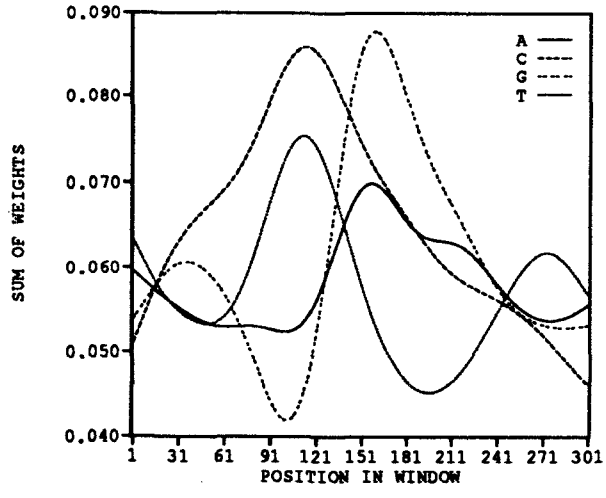

Figure 2: For the nucleotides A, C, G, and T these four curves give the average size of the weights in the coding/non–coding neural network as function of the position in the input window. Each point on the curves represents an average over 200 values of the input–to–hidden weights for each position in the window.

of 301 nucleotides which is coded as a binary string consisting of 301 blocks of four elements. An A is represented by the binary pattern 1000, C by 0100, G by 0010 and T by 0001. This string of 1204 numerals is broadcasted to the 200 hidden units each of which computes a weighted sum which subsequently is made the argument of a non-linear function possessing a sigmoidal (*i.e.* S-shaped) form. The value of the latter is limited downwards by zero and upwards by one. With the 200 real numbered activities in the hidden layer this step is repeated for the output unit, and finally, the resulting real-numbered activity is compared with a cutoff value, which separates the two output category assignments from each other.

For this network the 200 weights connecting the hidden units with the output unit had a large positive average value, meaning that input window configurations giving rise to high activities in the hidden units were likely to be classified as coding, while window configurations inhibiting activity in the hidden layer were likely to be classified as non-coding.

The size of the parameters connecting the input layer with the hidden layer was strongly correlated with both the window position and type of nucleotide, as can be seen in Figure 2.

Strong network parameters in specific regions of the input window mean that a matching sequence composition has a high chance of being classified as coding. Window configurations with high adenine content in the righthand side, high cytosine content in the left-hand side, and/or anti-correlated gradients in the guanine and thymine contents were likely to be classified as coding. Numerical estimates could be made by presenting 10^7 randomly generated window configurations to the network. 3.8% of these were classified as coding. They had: a) adenine and guanine contents in the right part exceeding the contents in the left part by 1.1% and 1.8%, respectively; b) cytosine and thymine contents in the left part exceeding the contents in the right part by 0.9% and 2.0%, respectively.

The prototype piece of DNA is compatible with the base composition in the region surrounding the acceptor site, where the polypyrimidine tract in the terminal part of the intron lowers the A and G content. Hence, in this region the network was able to perform the coding/non-coding classification with high confidence.

But what about the region surrounding the donor site in the initial part of the intron ? By measuring the jump in the frequencies at the donor sites in 95 human genes extracted from GenBank (all entries contained the complete coding sequence as well as the complete RNA transcript) it was demonstrated that the features of the prototype also can be found here. In the last 50 bp of the exons and in the first 50 bp of the introns the jump in the C content (from 28.8% down to 24.4%) and in the G content (from 27.9% up to 33.6%) was most significant and in fine agreement with the high average strengths of the network weights for C and G shown in figure 2. The change in the C content followed the jump in the average C content in exons and introns, while the G content in the terminal part of the exons was even lower than the average for exons, 28.5%, see Brunak *et al.*, 1991. The A content decreased from 24.4% down to 19.7%, while the T content increased slightly from 20.9% up to 22.5%. (This is not in agreement with figure 2 because of the conflicting situation with the terminal part of the introns).

Only the absolute size of the jumps depended on the length of the interval used when summing. The general picture holds out to about 100 bp, and also if the near vicinity of the donor site covered by the donor site consensus sequence $\frac{C}{A}AG/GT\frac{G}{A}AGT$ was excluded.

Together these numbers show that the initial part of human introns contains what could be termed a G/C rich tract, with a clear deficiency of adenine and thymine in the initial part of the introns. Figure 3 gives the Shannon information content as function of

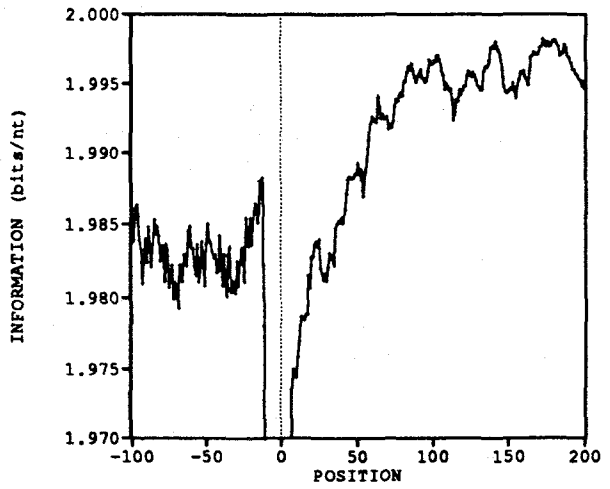

Figure 3: Shannon information content as function of position in the exon-to-intron transition region. The base-specificity in the G/C-tract lowers the information content in the initial 50 bp of the introns to a level below average for both exons (positions −100 through −1) and intron, where the value eventually comes very close to 2.0. The values for the information content were averaged over 10 neighbors. The unit of information is bits *per* nucleotide.

position in the exon-to-intron transition region. From the figure it may be concluded that the highest degree of systematics in this region is found in the initial 50 bp of the introns. The average value per nucleotide in the initial 50 bp of the introns is 1.85 bits and in the terminal 50 bp of the exons, 1.95 bits.

When frequencies were compiled for each of the a set of the 95 human sequences extracted from GenBank it showed that this systematics of the context of coding nucleotides is rather robust[4]. Apart from a few sequences with abnormally large internal exons the systematics was quite regular, indicating that this persistent systematics may be involved in the selection of proper splice sites and/or exons through binding or guidance of some of the molecules involved in spliceosome assembly[11].

4 Discussion

The complex biological mechanism where introns are removed from the genes is a crucial step in the information processing of the living cell. Just like in ordinary information processing in ordinary computers the biological computation where DNA is copied (transcribed) into pre–mRNA followed by intron removal and subsequent translation into a sequence of amino acids *information* is lost during the process. It is not possible to run the operation backwards; the sequence of the mature mRNA cannot be inferred from the protein sequence and the location and content of the introns cannot be deduced from the mRNA.

Neural networks have been shown here to be effective in identifying compositional gradients in DNA sequences which may support a recent model for splice site selection in vertebrates involving scanning[14]. The base specificities found by network inspection are to a large extent features which are present at each single transition from coding to non–coding sequence regions and not rather weak characteristics enhanced by obtaining statistics from a large sample of sequences. This indicates that the persistent systematics may be important for some of the steps in the processing of pre–mRNA.

The use of neural networks is often questioned because of the difficulty in extracting knowledge from them following training. This work shows however that it may be feasible to reveal previously unknown *local* features of the training data even for networks with as many as 241,201 adjustable parameters.

Acknowledgements

This work was supported in part by the Danish Natural Science Research Council under grants No. J.nr. 11-8168 and 5.26-1818.

References

[1] Brunak, S., Engelbrecth J. and Knudsen, S., "Cleaning up gene databases", *Nature*, Vol. 343, 123, 1990.

[2] Brunak, S., Engelbrecth J. and Knudsen, S., "Neural Network Detects Errors in the assignment of pre-mRNA splice sites", *Nucl. Acids Res.*, Vol. 18, 4797-4801, 1990.

[3] Brunak, S., Engelbrecth J. and Knudsen, S., "Prediction of human mRNA donor and acceptor sites from the DNA sequence", *J. Mol. Biol.*, Vol. 220, 49-65, 1991

[4] Engelbrecth J., Knudsen, S. and Brunak, S., "G/C rich tract in 5' end of human introns", *J. Mol. Biol.*, Vol. 227, 108-113, 1992.

[5] Fickett, J., "Recognition of Protein Coding Regions in DNA Sequences", *Nucl. Acids Res.*, Vol. 10, 5303-5318, 1982.

[6] Green, M. R., "Pre–mRNA Splicing," *Ann. Rev. Genet.*, Vol. 20, 671-708, 1986.

[7] Guthrie, C., "Messinger RNA splicing in yeast: Clues to why the spliceosome is a ribonucleoprotein", *Science*, Vol. 253, 157-163, 1991.

[8] Harris, N. L. and Senapathy, P., "Distribution and Consensus of Branch Point Signals in Eucaryotic Genes: a Computerized Statistical Analysis", *Nucl. Acids Res.*, Vol. 18, 3015-3019, 1990.

[9] Hertz, J., Krogh, A. and Palmer, R.G., "Introduction to the theory of neural computation", Santa Fe Institute, Studies in the Sciences of Complexity, Addison–Wesley, 1991.

[10] Lukashin, A. V, Engelbrecht, J. and Brunak,S., "Multiple Alignment Using Simulated Annealing: Branch Point Definition in Human mRNA Splicing", *Nucl. Acids Res.*, Vol. 20, 2511-2516, 1992.

[11] Maniatis, T. & Reed, R., "The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing", *Nature*, Vol. 325, 673-678, 1987.

[12] Minsky, M. and Papert, S., "Perceptrons", MIT Press, Cambridge, Massachusetts, 1969, 1988.

[13] Mount, S.M., "A catalogue of splice junction sequences", *Nucl. Acids Res.*, Vol. 10, 459-472, 1982.

[14] Niwa, M., MacDonald, C.C. and Berget, S.M., "Are vertebrate exons scanned during splice-site selection", *Nature*, Vol. 360, 277-280, 1992.

[15] Smith, T.F., Waterman, M.S., and Sadler, J.R. "Statistical characterization of nuclei acid sequences functional domains", *Nucl. Acids Res.*, Vol. 11, 2205-2220, 1993.

[16] Staden, R. "Computer methods to locate signals in nucleic acids sequences", *Nucl. Acids Res.*, Vol. 12, 505-519, 1984.

[17] Staden, R. "Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes", *Nucl. Acids Res.*, Vol. 12, 551-567, 1984.