A New Specification to Gene Signals Sensors by Neural Self Organizing Feature Map (SOFM)

Mariusz Zoltowski: ul. Bazynskiego 27/2, 80 309 Gdansk, Poland, zoltm@ieee.org

Biomed. Signals Analysis Lab., Colleg. Medicum of Nicolaus Copernicus Univ. of Torun, Poland.

ABSTRACT

Two different paradigms by the goals in gene-finding research have been recognized: 1) to offer computational aid in the annotation of the large volume of genomic data and 2) to provide a computational model helpful in elucidating the mechanisms involved in transcription, splicing, polyadynalation and other important processes on the pathway from genome to proteome [1]. New findings in gene regulation appear to focus a new interest in the latter paradigm approaches [1, 2, 3 and 4].

Therefore, a single weight matrix for the genomic patterns consensus scoring [5, 6] can be substituted by SOFM clusters matrices. This should result in the detection improvement of gene functional sites or signals, and therefore a gain evaluation across the known Burset's and Guigó's collection of the genes of 570 vertebrates is provided by a percentile measure on an exemplary site detection statistics. Such an improvement is important in both the "extrinsic" and "intrinsic" approaches [14]. In the "signal" case of the latter [14], a demand is addressed for a neural approach which translates into likelihood scoring. This includes, but is not limited to, applications with HMM (Hidden Markov Model) derived gene finders [7]. The approach is also scalable into a clusters-based solution to genes recognition with the capability of integrating DNA-string-contained knowledge in a novel way.

1. INTRODUCTION

In an attempt to analyse genes by computer (Fig.1), further improvements on gene structure elements detection and elucidation of the gene expression machinery would seem to be required [14], in view of new findings in gene regulation which include alternative splicing, alternative polyadenylation, alternative transcription initiation, RNA editing, twintrons, overlapping genes, trans-splicing etc. [1].Such gene functional sites can be recognized by so-called signal sensors or their combination.

The genes are strings of A, T, G and C nucleotides, arranged in protein coding exons and non-coding introns of the DNA double helix strands. Their intron-less (by splicing) assembly, ready to translate into protein, is known as expression. A gene parse is the most likely phrase among gene structure terms which are identified by content and signal sensors (Fig.1, [14]). During a sequential parse evidence related to gene structure from signal or content sensors is being contributed part by part. Hence a likelihood score (soft score) is required rather than a categorical choice (hard score) which can turn out to be erroneous. The more adequate the evidence in its score, the better the whole recognition.

Three subsequent nucleotides, as not implicitly singled out from a DNA sequence, introduce 3 options of being read as coding for protein amino acids codons which are called open reading frames (OFR-s).

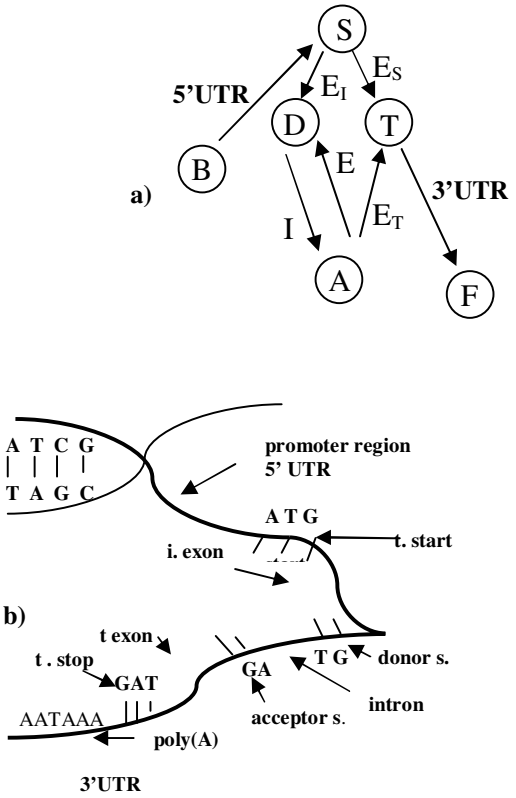


Fig.1. a) - By signal and content [9, 13-14] linguistic diagram for possible events while parsing a sequence of a candidate multi-exon gene [12], in the direction from 5' (upstream) to 3' (downstream). The arcs represented contents are: 5'UTR, E_I–initial exon, E–internal exon, I–intron, E_T– terminal exon, 3'UTR. The nodes represented

signals: B–beginning, S–translation start, D-donor splice site, A-acceptor, T-translation stop, F-sequence end;
b)- A gene structure by DNA stretch bases sequence between start and stop codons.

Among different tools used to recognize gene nucleotides patterns, including binding sites of transcription machinery and other gene functional elements, a profile matrix is still of great importance [3, 5 and 6].

Let $\mathbf{S} = \mathbf{s_1},...,\mathbf{s_N}$ be a set of DNA aligned sequences, N patterns instances of length L i.e. $\mathbf{s}_k = s_{k1},...,s_{kL}$, k=1, ..., N. A profile $\mathbf{P}_{4xL}$ is usually derived as $\mathbf{P}_{nt\,j} = \frac{1}{N}\sum_{j=1}^{N} i_{nt}(s_{kj})$, where *nt* = A, T, C, G; j=1, …, L and

$$i_{nt}(s) = \begin{cases} 1 \; if \;\; s = nt \\ 0 \; otherwise \end{cases} \qquad (1)$$

The profile, which is also a position-weighted matrix (PWM), is commonly used to recognize gene functional elements. This is done by scoring an annotated sequence versus the profile in which the each i-th column entry to PWM estimates the probability $p_i(nt_i)$ of finding at position *i* - a nucleotide $nt_i \in \{A,T,C,G\}$. Hence, for any candidate template sequence, the net consensus score of a site is approximated by the product of the probabilities, i.e. by likelihood scoring of template nucleotides, i.e.

$$p_{PWM} = \prod_i p_i(nt_i) \qquad (2).$$

However, within a sequence pattern, a single PWM does not account for correlations of A, T, C and G bases. To account for them, more sophisticated statistical approaches, such as Maximal Dependence Decomposition (MDD, [8]) or multi-layered neural nets approaches [9], have been devised. Neural networks trained by backward-propagation algorithm performance are well recognized [9], yet their outcomes neither guarantee optimality globally nor can be stated by rules clarifying the results, though some recent approaches appear to alleviate this disadvantage [9]. Also they do not simply translate into a likelihood scoring [7]. It is for this reason that the attempt is made to extend simple PWM scoring to multi PWM-s by SOFM clustering of gene data as a novelty. Since such a scoring has been attached to the nodes and arcs of gene model (Fig.1), an optimal gene parse can be found [5, 7-8, 12].

2. MATERIAL AND METHODS

The idea behind this approach is to substitute a gene site single PWM profile, e.g. provided in [5], by profile clusters constructed by a self-organizing feature map (SOFM), which is well suited to detecting correlations and regularities in its input waveforms and adapting its future responses to that input accordingly [10, 11]. To represent an input space of its excitation ***x***, **SOFM** performs by its neuronal unit ***w*** together with its neighborhood $N_i(d)$ containing all neurons within a range of radius *d*, which are configured according to the Kohonen rule [10, 11];

$$\mathbf{w}_i(t+1) = (1-\alpha)\,\mathbf{w}_i(t) + \alpha\,\mathbf{x}(t) \qquad (3),$$

where *i* stands for the winning neuron, i.e. the one which is closest to the input ***x***. SOFM learning is completed in two phases; the ordering phase and the tuning phase [10, 11].

A binary code known as BIN4 [9] can be used to represent an A, T, C and G nucleotides DNA pattern of length L. Accordingly, BIN4 codes of nucleotides are; [A]=[1 0 0 0], [T]=[0 1 0 0], [C]=[0 0 1 0] & [G]=[0 0 0 1].

It can be observed that after a learning set samples clustering, both the coordinate-wise orthogonality of BIN4 and the learning rule of (5), which is also a mean value formula with a controlling learning factor α, imply plausible properties of SOFM out coming cluster, i.e.;

1. Each neuronal vector ***w*** of length 4xL becomes a position- weighted feature matrix (PWFM)

$$\mathbf{P}^F = [p_i^F(nt)]_{4xL},\; nt = A,T,C\,and\,G \quad i = 1,...,L,$$

$$\sum_{nt=A,T,C,G} p_i^F(nt) = 1 \qquad (4),$$

with $p_i^F(nt)$ relevant to nucleotide *nt* frequency in the learning set.

2. SOFM response to the input sample is the winning vector of the highest consensus score class according to its cluster PWFM.

3. Correlations among the clusters become dependent on the topology of the SOFM network.

To get an idea of an order of value of *N* in the learning set division into *N* parts-clusters classes, an entropy index is devised: $I_C(\mathbf{P}_j^F, N) =$

$$= \sum_{j=1}^{N} p_j^c \prod_{i=1}^{L} 2 \wedge -\left(\sum_{nt=A,T,C,G} p_{ij}(nt)\log_2[p_{ij}(nt)] + 2 \right)$$

(5), with "∧" power operator.

By matrices $\mathbf{P}_j^F$ which are *N* in number, index $I_c(\mathbf{P}_j^F, N)$ of (5) answers the question of how many

states of symbols in the learning sequences are in comparison to the ones with a uniform distribution of A, T, C and G nucleotides. By p_j^c the frequency is relevant to how often a j- cluster neuron wins in response to learning data being trained, i.e. this is a cluster class probability.

3. RESULTS

Given the set of 570 vertebrates' genes, an I_c index versus clusters number is shown in Fig.2. A linear topology of SOFM has been chosen to minimize mutual correlations among clusters neurons. Each template length is set as short as L=12 to compare it against the PWM considered by Guigó [5] for the translation start case. The TS signal includes an ATG string, but the ATG string does not imply a TS. Therefore, after the real functional site pattern has been SOFM-learnt by acquisition of its PWMS-s feature clusters, the set is scanned along each gene for all candidate patterns, which are assigned their likelihood scoring. Candidates for the TS are identified by an ATG codon of methionine amino acid before they are scored by the SOFM PWM-s to establish their priority. The candidates which do not belong to signal or content sensor category are expected to score less by this sensor than in the category which they belong to. In the x – y plane of Fig.3, the cumulate percentile curve (CPC) informs, by its y- CPC number, how many true functional sites score equally to, or greater than, x- percentile of the overall statistics of the candidate functional sites, i.e. of those by the ATG-only consensus pattern, in the TS signal case. That improves remarkably (Fig.3) with more clusters by N. Shapes for true TS signals appearing relevant to an exponential Poisson-like distribution of possible scorings and a not true TS case by linear CPC curve describing uniform-like scorings are given in Fig.4.

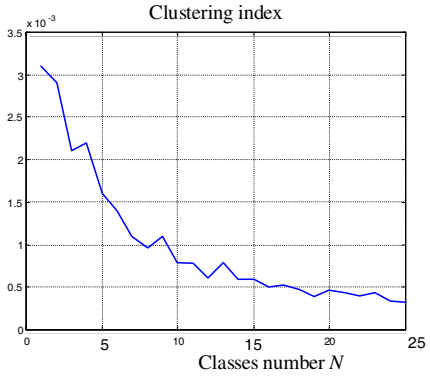


Fig.2. Clustering index I_c versus the number N of classes, "ATG" start codon and the Burse & Guigó set case.

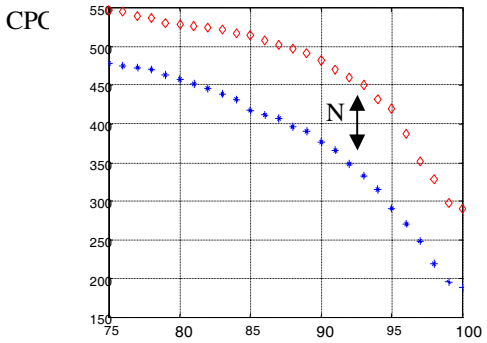


Fig.3. CPC versus percentile, translation sites (TS) case. Improvement by SOFM cluster (diamond, constraining N ≅ 40) versus single PWM (star). True TS-sites trained SOFM for their recognition.

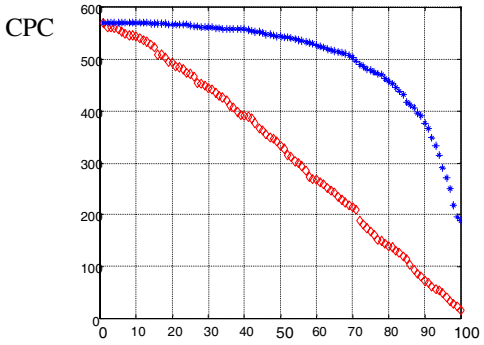


Fig.4. TS CPC versus percentile. The SOFM trained bottom line of scoring with not true TS-s (N=450 non-TS sites number equal to 46 602). Top curve as in Fig.3 – True TS-s clusters PWM-s scoring case, N≅40.

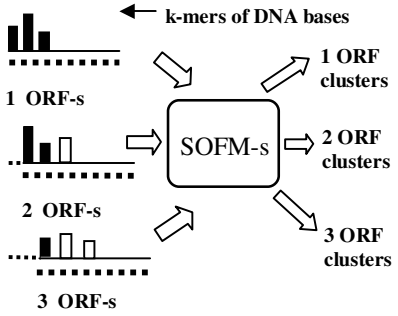
4. DISCUSSION AND PERSPECTIVES

Within gene recognition related to an artificial intelligence (AI) paradigm, SOFM out-coming vector clusters, which also serve weighting matrices for scoring, simply appear as winning neurons response. Though this approach may be related to the statistical ones [14] no statistical motivation has been a priori attached while the winning neuron scoring matrix self-selection can benefit from parallel architectures. The candidates scoring of gene functional sites allow a "soft" decision to true gene reconstruction in comparison to other neuronal approaches. Since genes' AI algorithms are originally exponentially hard [5], a "right" initial guess to "gene parsing" by dynamic programming is important to prune possibly many alternative nodes when a gene parse by AI tree data structure [5] is at view. This is especially so with approaches which are "extrinsic" or "intrinsic" by content [14]. Also, in connection with robust

and fast HMM algorithm of Kulp for gene annotation which is "intrinsic" by signals, a demand for a neural network result translating into likelihood for scoring is addressed [7].

The presented approach is extendable to whole gene recognition by gene structure-dedicated signals and content sensors aggregation. This concerns sensors which refer not only to gene functional sites signal sensors such as the TS or stop, the splicing donor or acceptor junctions but also to exonic regions in each of the three ORF-s, or to intronic genomic fragments being cut off during splicing (Fig. 1).

Given a scoring likelihood of the exciting A, T, C, G string, gene structure parse can be accomplished (Fig.1, [5, 7, 8, 13 and 14]). An improvement is clear by a CPC figure for a site case showing also, by Fig.5, the cis-control role [14] to distinguish the true from the not true sites. Also the winning neuron pattern can be a DNA part which codes for a protein motif which accounts for gene coded structural elements within a secondary gene structure case, such as a helix, sheet or coil. This seems to be an issue that can not be directly handled in a statistical way, as a problem with regular patterns case consistent with chaos approach rather than statistically. However, further evaluations are needed to see if the predicted advantages come with practice in a weight-balanced way. On the other hand, HMM probabilities seem to show an instance case of Fig.5 showing how period-3 HMM-like statistics, i.e. the one in three ORF-s, can be derived by SOFM clusters for a parse by the dynamic algorithm of Viterbi [7]. Also an optimization problem of set of clusters has resulted [1].



.

Fig.5. Principle of content sensors construction. Signal sensors k-mers are around gene functional sites.

5. REFERENCES

[1] Hamady M et al., "Key challenges in Proteomics and Proteoinformatics", *IEEE Eng. in Med. and Biol. Mag.*, IEEE, USA, pp.34-40, May/June 2005.

[2] Fairbrother et al., "RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons", *Nucleic Acids Res.*, Oxford Univ. Press, London, pp.187-190, 32 WS issue, 2004,.

[3] Zhang M.Q.,"Statistical features of human exons and their flanking regions", *Human Molecular Genetics*, Oxford Univ. Press, Oxford, pp.919-932,7, 5, 1998.

[4] Werner T., Fessele S., Maier H. and Nelson P.J., "Computer modeling of promoter organization as a tool to study transcriptional co regulation", *FASEB J.*, FASEB, Bethesda. MD. USA, pp.1228-1237, 17, 2003.

[5] Guigo R., Knudsen S., Drake N. and Smith T., "Prediction of Gene Structure", *J. Mol. Biol.*, Academic Press, London, pp.141-157, 226, 1992.

[6] Gershenzon N.I. et al., "Computational Technique for improvement of the position-weight matrices for the DNA/protein binding sites", *Nucleic Acids Res.,* Oxford Univ. Press, London, pp.2290-2301, 33, 7, 2005.

[7] Kulp D.C., *Protein-Coding gene structure prediction using generalized Hidden Markov models*, Ph. D. Dissertation, California, March 2003.

[8] Burge C., Karlin S.,"Prediction of complete Gene Structure in Human Genomic DNA", *J. Mol. Biol.*, Academic Press, London, pp.78-94, 268, 1997.

[9] Wu C.H., McLarty J.W., *Neural Networks and Genome Informatics, Methods in Computational Biology and Biochemistry,* Elsevier, Oxford UK. 2000.

[10] Kohonen T., *Self-organization and Associative Memory, 2nd Edition*, Springer – Verlag, Berlin, 1987.

[11] Haykin S., *Neural networks a comprehensive foundation*, Prentice Hall, USA, 1999.

[12] Haussler D., "Computational genefinding", *Trends Biochem. Sci.*, Elsevier, Amsterdam, pp.12-15, 1998.

[13] Claverie J.-M.,"Computational methods for the identification of genes in vertebrate genomic sequences", *Human Mol. Gen.*, Oxford Univ. Press, London, pp.1735-1744, 6, 10, 1977.

[14] Mathé C., Sagot M.-F., Shiex T. and Rouzé P., "Current methods of gene prediction, their strengths and weakness", *Nucleic Acids Res.*, Oxford Univ. Press, London, pp.4103-4117, 30, 19, 2002.

[1] Regards to Keith Miller for reading my English