A GA Approach to the Definition of Regulatory Signals in Genomic Sequences

Giancarlo Mauri, Roberto Mosca, and Giulio Pavesi

Bioinformatics and Natural Computing Group
University of Milano–Bicocca
mauri,roberto.mosca,pavesi@disco.unimib.it

Abstract. One of the main challenges in modern biology and genome research is to understand the complex mechanisms that regulate gene expression. Being able to tell when, why, and how one or more genes are activated could provide information of inestimable value for the understanding of the mechanisms of life. The wealth of genomic data now available opens new opportunities to researchers. We present how a method based on genetic algorithms has been applied to the characterization of two regulatory signals in DNA sequences, that help the cellular apparatus to locate the beginning of a gene along the genome, and to start its transcription. The signals have been derived from the analysis of a large number of genomic sequences. Comparisons with related work show that our method presents different improvements, both from the computational viewpoint, and in the biological relevance of the results obtained.

1 Introduction

One of the main challenges in modern biology in general, and in the analysis of genome data in particular, is to understand the complex mechanisms that regulate the expression (i.e. the activation) of the genes of a given organism. The expression of a gene starts when the corresponding region in the double–stranded DNA sequence is *transcribed* into a single stranded RNA sequence, that later on is translated into the protein encoded by the gene (see Fig. 1). At any given time, not all the genes present in the genome of a given organism are expressed, but only a subset of them: this accounts for example for cell differentiation, that is, the genes that are active in a neural cell are different from those active, say, in a muscle cell. Moreover, genetic diseases are often caused by alterations occurring not within the genes themselves, but in the apparatus governing their activation, thus leading to anomalous expression levels. Transcription is initiated when one or more dedicated molecules called *transcription factors* (TFs) (that are proteins in turn encoded by some genes in the genome) bind to the DNA region adjacent to the gene (region called *promoter* of the gene), causing the double–strand to open and thus allowing the transcription of the gene. In some cases, the binding of a TF has the opposite effect, blocking transcription. Each TF recognizes a set of specific targets along the sequence, that is, short nucleotide fragments it can

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 380–391, 2004.

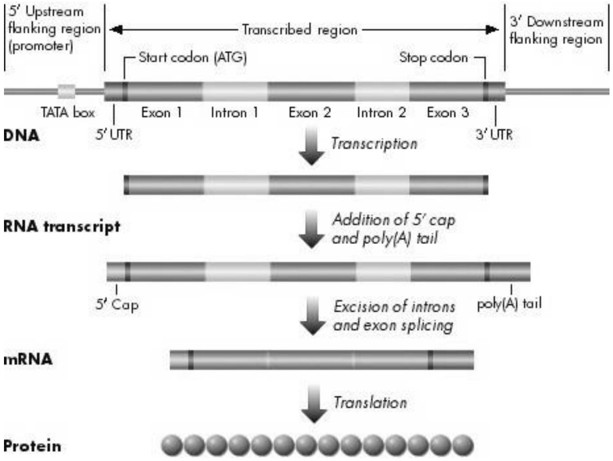

Fig. 1. Gene expression: from DNA to RNA to protein.

bind to, called *binding sites*. Binding sites thus function as regulatory signals in the genome.

Since the experimental characterization and identification of the binding sites of a given TF is a long and painstaking work, the huge amount of genomic data now available to researchers provides a invaluable source of information for shedding further light on this process. If we identify in the promoter of a gene known TF binding sites, we may better understand by whom, and when a gene is activated. Unfortunately, also the computational description and discovery of the binding sites of a given TF is far from being an easy task. The main difficulty lies in the very fact that each TF does not recognize a single binding site, but a set of them, that, although similar, differ in their nucleotide composition. This set is usually referred to as *signal* or *motif*.

1.1 TATA Box and CAP Binding Sites

Usually, each TF influences a relatively small number of organism–specific genes. However, it has been experimentally observed that in virtually every eukaryotic organism a large number of genes present two characteristic signals located in the proximity of the point of transcription initiation (*transcription start site*, TSS). The first one, called *TATA box*, is located along the DNA double helix about 25-30 pairs of nucleotides (base pairs, bp) before (upstream of) the TSS. Its name derives from the fact that when it was discovered the majority of its instances contained the stretch of nucleotides `TATA`. The TATA box is bound by a large complex of some 50 different proteins, including Transcription Factor IID (TFIID) – a complex of the TATA-binding protein (TBP, which is the part of the molecule that binds the TATA box) and 14 other proteins – and Transcription Factor IIB (TFIIB). The CAP signal (also called Initiator, or Inr) instead straddles the TSS, and often cooperates with the TATA box in starting the transcription of the gene. There is significant experimental evidence that also this signal is bound by Transcription Factor IID (TFIID) [1]. While most of the regulatory signals are spread at different positions in genome regions adjacent to genes, both the TATA and CAP motifs have instead a very precise location, perhaps with the purpose of directing the transcription apparatus to the right spot along the DNA sequence, signalling where exactly the gene begins and transcription has to start. While largely present, these two signals are not ubiquitous: in every organism, some genes contain both the signals, some either one, and some neither of them.

Giving a precise characterization of these two signals, as well as a classification of genes according to which signal they are regulated by, could thus provide substantial information on a basic mechanism for gene expression, present virtually in every organism.

2 Describing Binding Sites

The computational characterization of regulatory signals in genomic sequences is usually composed by two different steps. First, we need a method to describe a signal, that is, to represent the set of valid binding sites for a given TF. One possible way is to describe them with a *frequency matrix*, that is built as follows. Let B_{S} be a set of DNA fragments (all having the same length m) that a transcription factor is known to bind. Since all the fragments have the same length we can align them obtaining m columns. Let us consider the first column, containing the first nucleotide of each fragment. We can count the number of times each nucleotide is present in the column, and compute its frequency. The same operation can be performed for the other columns. In this way, a $4 \times m$ matrix P can be built, where the element $\mathrm{P}(i, j)$ is the frequency of nucleotide i ($i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$) in the j–th column of the alignment:

$$\mathrm{P}(i, j) = \frac{n_{i,j}}{N} \tag{1}$$

where $n_{i,j}$ is the number of times nucleotide i is found in the j–th position of the fragments considered, and N is the overall number of binding sites used. Given a set of promoter sequences of genes known (or suspected) to be regulated by the same transcription factor, the problem of finding its binding sites can be defined as the problem of finding the best frequency matrix. The basic idea is to find the matrix whose frequencies differ most from those that would be obtained by putting together random fragments from the sequences. Different measures for this task have been proposed so far, with some success [2,3].

Then, the frequency matrix can be used to determine whether a given DNA fragment can be considered a candidate binding site for the corresponding TF. In fact, given a frequency matrix P describing a signal, and a generic fragment S^{f} of length m the probability of the fragment to be a binding site for the corresponding TF can be estimated by

$$\mathrm{P}_{\mathrm{match}}(S^{\mathrm{f}}) = \prod_{j=1}^{m} \mathrm{P}(S^{\mathrm{f}}_j, j) \tag{2}$$

where S^{f}_j is the j-th nucleotide of S^{f}. Fragment S^{f} is a said to be a binding site when $\mathrm{P}_{\mathrm{match}}(S^{\mathrm{f}})$ is greater than the probability with which S^{f} appears in the genome:

$$\mathrm{P}_{\mathrm{bg}}(S^{\mathrm{f}}) = \prod_{j=1}^{m} b_{S^{\mathrm{f}}_j} \tag{3}$$

where b_i is the frequency with which nucleotide i appears in the genome. The b_i values can be estimated for example by considering the frequency of each nucleotide in the sequence examined, or, since these data are now available, in the whole genome the sequence considered belongs to.

Stated in another way, a fragment S^{f} can be suspected to be a binding site if

$$\frac{\mathrm{P}_{\mathrm{match}}(S^{\mathrm{f}})}{\mathrm{P}_{\mathrm{bg}}(S^{\mathrm{f}})} = \prod_{j=1}^{m} \frac{\mathrm{P}(S^{\mathrm{f}}_j, j)}{b_{S^{\mathrm{f}}_j}} > 1 \Leftrightarrow \log \frac{\mathrm{P}_{\mathrm{match}}(S^{\mathrm{f}})}{\mathrm{P}_{\mathrm{bg}}(S^{\mathrm{f}})} = \sum_{j=1}^{m} \log \frac{\mathrm{P}(S^{\mathrm{f}}_j, j)}{b_{S^{\mathrm{f}}_j}} > 0 \tag{4}$$

so to have negative score for fragments that are not bound by the TF. That is, we evaluate the probability of a sequence fragment to be bound by the TF described by matrix P by checking whether it fits the description of the matrix ($\mathrm{P}_{\mathrm{match}}(S^{\mathrm{f}})$), and by comparing this result with the expected frequency ($\mathrm{P}_{\mathrm{bg}}(S^{\mathrm{f}})$) of the fragment in the genome.

Given a sequence S of arbitrary length $L > m$ we can say that the signal characterized by the frequency matrix P occurs at position t in the sequence if the fragment of length m starting at position t yields a positive value in the above equation.

Nearly all the methods proposed so far for the discovery of signals in genomic sequences are position–independent, that is, do not make any assumption on the location of binding sites along the input sequences. Moreover, they require all (or most of) the sequences studied to contain an instance of a binding site. In our case, instead, the position of the binding sites is known in advance, making the problem somewhat easier. This advantage, however, is balanced by the fact that while general case methods work on a set of pre–selected sequences, most of which supposedly share binding sites for the same (unknown) TF, in the case of the TATA and CAP signals we cannot choose the sequences to examine beforehand, but we have to work on virtually every gene promoter sequence available. Thus, the problem can be recast as choosing a subset of the promoters, and using fragments located in their TATA box and CAP positions to build a frequency matrix. Also, we have to define a suitable function to evaluate the best matrix, that is, the one providing the best partition between sequences containing the signals and those that do not contain it.

Matrices describing the TATA and CAP signals have been first characterized by Bucher in a seminal work in the late '80s [4], and are still used by researchers today. The method used started from two initial matrices, that were optimized using a local search technique. The datasets analyzed, however, composed by the sequences available at the time, were relatively small. Moreover, the fact that the method started from an initial matrix derived from human inspection of the data, inevitably skewed the results according to the starting point (basically, it had been observed that the TATA box contained the fragment `TATAA`, while the CAP motif started with `CA`). Thus, in this work our aim was to see whether the results obtained with a much larger dataset were consistent with the previous ones, and at the same time avoiding any preliminary constraint for the matrix.

In the following, we first introduce a formalization of the problem. Then, we present the genetic algorithm we used for the optimization of the scoring function used to evaluate candidate matrices. Finally, we compare our results with previous characterizations of the same signals.

3 The Problem

First of all, few variables must be introduced for the formalization of the problem:

1. $D_{\mathrm{S}} = \left\{S^1, S^2, \ldots, S^n\right\}$ is the dataset composed by n sequences S^i of length L. Every sequence starts at position p_0 and ends at positon p_{f}, measured with respect to the TSS. Usually $p_0 < 0$, and $p_{\mathrm{f}} > 0$, that is, each sequence encompasses the TSS of a different gene.
2. $\bar{s}$ is the position of the signal with respect to the TSS (for example -2 for the CAP signal).

Since we already know the position $\bar{s}$ of the binding site under investigation, the problem can be defined as finding a *partition* of the sequences in two subsets. Let us suppose that a particular signal of length m starts at position $\bar{s}$. Consider the set of fragments $B = \left\{S^i_{\bar{s}} S^i_{\bar{s}+1} \ldots S^i_{\bar{s}+m-1} | S^i \in D_{\mathrm{S}}\right\}$ consisting of all the fragments of length m starting at position $\bar{s}$ in all the sequences belonging to the dataset. In general not all of them will be binding sites. Let us suppose that just a subset $\bar{S}$ of the sequences in D_{S} contains the signal. Then, considering $B_{\bar{\mathrm{S}}} = \left\{S^i_{\bar{s}} S^i_{\bar{s}+1} \ldots S^i_{\bar{s}+m-1} \in B | S^i \in \bar{S}\right\}$ it is possible to build a frequency matrix from the subset $B_{\bar{\mathrm{S}}}$ of fragments. The fragments used for this purpose are all the substrings of length m starting at position $\bar{s}$ in the sequences belonging to the subset $\bar{S}$.

Once a matrix has been built we need to define a score value for it. Different approaches have been introduced so far, including information content, MAP (Maximum A posteriori Probability) score and other approaches related to the sensitivity and specificity of the matrix [2,3,5,6,7]. Here, instead, we defined the scoring function in order to reflect also the fact that the signals we are trying to describe are position specific. Thus, they should not appear elsewhere along the sequence, so not to confuse the TF that binds them. In other words, we want the probability of the fragments in the signal position to be a binding site to be higher than the probability associated with all the fragments of the sequences in other positions. We associate positive scores to the fragments appearing in the correct position in the sequences selected:

$$POS(\bar{S}) = \frac{\sum_{S^i \in D_{\mathrm{S}}} S_{\mathrm{P}}(S^i, \bar{s})}{n} \tag{5}$$

where $S_{\mathrm{P}}(S^i, \bar{s})$ is the score of the fragment of length m starting at position $\bar{s}$ in sequence S^i defined in (4) as

$$S_{\mathrm{P}}(S^i, \bar{s}) = \sum_{j=1}^{m} \log \frac{\mathrm{P}(S^i_{\bar{s}+j-1}, j)}{b_{S^i_{\bar{s}+j-1}}} \ . \tag{6}$$

Conversely, we associate a negative score with the fragments appearing in the other positions, that is, we want as less instances of the signal as possible to appear in wrong positions in the sequences selected:

$$NEG(\bar{S}) = \frac{\sum_{S^i \in D_{\mathrm{S}}} \sum_{j=p_0, j \neq \bar{s}}^{p_o+L-m} S_{\mathrm{P}}(S^i, j)}{n\,(L-m)} \ . \tag{7}$$

Thus, the score associated with subset $\bar{S}$ and the corresponding frequency matrix is given by the difference between (5) and (7)

$$S_{\mathrm{P}}(\bar{S}) = POS(\bar{S}) - NEG(\bar{S}) \ . \tag{8}$$

According to this score measure, the best matrix will be the one providing the largest difference between the occurrences of the signal in the position selected ($\bar{s}$) and its occurrence elsewhere in the sequences. The greater is the difference between the two terms and the more selective we can consider the matrix describing the binding site in the position under consideration. The goal is to find the subset S^* leading to the maximum score.

$$S^* = \arg \max_{\bar{S} \in \mathcal{P}(D_{\mathrm{S}})} \left\{ S_{\mathrm{P}_{\bar{S}}}(\bar{S}) \right\} . \tag{9}$$

where $\mathrm{P}_{\bar{S}}$ is the frequency matrix calculated from the subset $\bar{S}$ and $\mathcal{P}(D_{\mathrm{S}})$ is the power-set of D_{S}.

4 The Genetic Algorithm

For the solution of the problem we employed a genetic algorithm. For this purpose, two things must be provided: a method to encode an instance of the problem and a fitness function for the genome itself. A very simple solution is to use a binary string genome whose length equals the number n of sequences in the dataset. Given a genome string g the variable g_i indicates the i-th bit in the string. If $g_i = 1$ then the i-th sequence S^i in the dataset is included in the subset $\bar{S}$ of the positive sequences, otherwise it is not included. The fitness evaluation for each genome is done in the following way:

1 Given a genome g, derive the frequency matrix $\mathrm{P}_{\bar{S}}$ from the subset of sequences $\bar{S} = \left\{ S^i \in D_{\mathrm{S}} | g_i = 1 \right\}$.

2 Compute the score $S_{\mathrm{P}_{\bar{S}}}(\bar{S})$. Negative scores are truncated to 0.

For the implementation of the GA we used the Galib library [8]. We employed one point crossovers with probability p_{c} to every couple of individuals during the evolution step. The mutation operator flips one bit in a genome with a probability p_{m}. Parent genomes are selected with a roulette wheel scheme, then mating and mutation are applied. The offspring genomes completely replace the parent population. The fitness of each individual was obtained from the score value with a linear scaling system as described in [9]. Different numbers of evolution steps and termination criteria have been tried. Usually no improvements on the fitness of the best individual were obtained after 20000 steps.

On the best genome output by the algorithm we also applied a local optimization procedure. For each sequence in the dataset, this procedure tries to perform either of the following actions:

1. if the sequence was selected by the GA for the matrix computation, it tries to exclude it;
2. if the sequence was not selected by the GA, it tries to include it.

If the change increases the score, then it is accepted, otherwise it is rejected and a new sequence is processed. This step is repeated until no further improvement can be made to the score.

Bucher's approach, even with some differences in the scoring function, performed, instead, just this procedure, starting from fragments that began with `CA` for the CAP signal and `TATAA` for the TATA box.

5 Results

Our method has been applied to three different datasets retrieved from two databases of sequences freely accessible on the web. The first one is the Eukaryotic Promoter Database (`http://www.epd.isb-sib.ch/`, [10]), a database of promoter sequences belonging to eukaryotic organisms. Each of these sequences belongs to a different gene, encompassing the exact point of transcription initiation. From the release 74 of this database we retrieved sequences belonging to 2199 genes of vertebrate organisms (EPD Vertebrates) 1796 of which were human sequences (EPD Homo Sapiens). It can be clearly seen how this is just a small subset of all the thousands and thousands of genes now available. However, in most of the cases the exact location along the genome of the TSS is not known. The EPD contains promoter sequences of genes where at least one TSS has been determined experimentally. An analogous species–specific database is the Drosophila Core Promoter Database (DCPD, `http://www-biology.ucsd.edui /labs/Kadonaga/DCPD.htm`, [11]) which contains several promoters belonging to the genome of the fruit fly (*Drosophila melanogaster*). Sequences taken from EPD were 101 nucleotides long starting from position -50 with respect to the TSS. From DCPD we retrieved 205 gene sequences, starting at position -47 and 92 nucleotides long. We performed our analysis on all the EPD promoters of vertebrates, and then on human promoters only. The fruit fly dataset has been used to validate the results, since on these sequences the occurrences of CAP and TATA box in each have been verified experimentally. In all the cases, some sequences contained both signals, some either of them, and some none. On each dataset we ran the genetic algorithm in order to obtain the best matrix for the

Table 1. Parameters of the genetic algorithm used in the experiments.

Parameter	Value
Population	500
Generations	20000
Crossover prob. (p_c)	0.9
Mutation prob. (p_m)	0.04
$\bar{s}$	-2 (CAP) -30 (TATA-box)
Matrix length m	6 (CAP) 8 (TATA-box)

CAP signal (choosing position -2 with length 6) and the TATA box. While the TATA box is usually found in a random position between −36 and −24, we fixed position −30 (matrix of length 8), which has the highest frequency of occurrence. The parameters used in the GA are shown in Table 1. The application of the local optimization procedure after different runs of the GA on the various datasets converged to virtually the same matrix in each case. The signals obtained can be thus trusted to be strong (perhaps global) optima for the scoring function employed.

5.1 The CAP Signal

The CAP signal matrices computed by the genetic algorithm on the sequences of the first dataset (EPD Vertebrates and the Homo Sapiens subset) are shown in Table 2. As shown in the table, the local optimization procedure is able to "clean" the frequencies in the matrix, getting a more conserved distribution of nucleotides. Indeed, after the local optimization we can observe that the CAP signal defined over the EPD datasets has either a C or a T in position −1 (just before the TSS), and either an A or a G in position 0. No A is present in position −2 and no G in position 2. The former characterization on eukaryotes of the CAP signal of Bucher [4], where the matrix length was fixed to 8 nucleotides, is shown in Table 3. It is the result of a refinement of an initial matrix built using fragments containing a C in position −1 and an A in position 0. This matrix is still reported in EPD as the reference matrix for the description of this signal, and for this reason we compared the matrix obtained with our method with this one.

The final matrix maintains this strict constraint of a CA dinucleotide at position −1 (clearly determined by the initial choice), while our matrix shows also the presence of possible CG, TA, TG dinucleotides. The biological feasibility of this result is supported by the fact that, in the case of fruit fly sequences (see Table 4), the matrix shows the possible presence of CA and TA dinucleotides at position −1, a fact that is consistent with the results obtained experimentally by Kutach and Kadonaga in [11].

In Table 5 we show a comparison between the scores of the matrices obtained with our technique on the three different datasets and score of the matrix obtained by Bucher. For every matrix the score is calculated with the method

Table 2. CAP signal frequency matrix obtained with the genetic algorithm on the EPD datasets.

	-2	-1	0	1	2	3	-2	-1	0	1	2	3
	EPD Vertebrates						***EPD Homo Sapiens***					
	Non optimized						*Non optimized*					
A	0.069	0.000	**0.608**	0.139	0.226	0.184	0.017	0.000	**0.623**	0.136	0.259	0.174
C	0.345	**0.685**	0.129	0.223	0.240	0.252	0.356	**0.712**	0.092	0.189	0.234	0.229
G	0.294	0.009	**0.263**	0.301	0.151	0.278	0.327	0.000	**0.285**	0.341	0.121	0.293
T	0.293	**0.305**	0.000	0.337	0.383	0.286	0.300	**0.288**	0.000	0.334	0.387	0.303
	Optimized						*Optimized*					
A	0.001	0.000	**0.738**	0.144	0.304	0.177	0.000	0.000	**0.714**	0.119	0.328	0.167
C	0.366	**0.728**	0.000	0.247	0.253	0.231	0.374	**0.725**	0.000	0.227	0.225	0.229
G	0.310	0.000	**0.262**	0.360	0.000	0.323	0.311	0.000	**0.286**	0.385	0.000	0.332
T	0.323	**0.272**	0.000	0.279	**0.444**	0.270	0.315	**0.275**	0.000	0.270	**0.447**	0.271

Table 3. CAP signal obtained by Bucher in [4].

	-2	-1	0	1	2	3	4	5
A	0.162	0.000	**0.950**	0.086	0.254	0.221	0.149	0.165
C	0.158	**1.000**	0.000	0.267	0.314	0.281	0.281	0.317
G	0.228	0.000	0.000	0.383	0.000	0.241	0.241	0.185
T	0.452	0.000	0.050	0.264	**0.432**	0.330	0.330	0.333

described in section 2. Even if the latter is 2 nt longer than ours, its score is always lower. Moreover, in our description the signal appears in a greater number of sequences, at the same time maintaining a higher specificity for the position considered.

5.2 The TATA-Box Signal

The TATA-box signal has been searched in position -30 with length 8. In fact, this signal usually appears in the range between -36 and -24, but shows a high preference for position -30. In this case the score function has been modified, using for the positive term of (5), the average score over the range $[-36, -24]$ instead of position -30 alone. The negative term was taken to be the average

Table 4. CAP signal obtained with the genetic algorithm on the DCPD dataset.

	-2	-1	0	1	2	3
A	0.185	0.000	**1.000**	0.000	0.000	0.228
C	0.000	**0.772**	0.000	0.130	0.185	0.326
G	0.130	0.000	0.000	**0.565**	0.000	0.000
T	**0.685**	**0.228**	0.000	0.304	**0.815**	0.446

Table 5. Comparison between the score of the frequency matrix obtained by Bucher and the matrices computed using genetic algorithms on different datasets. Column *Seqs.* shows the number of sequences containing the signal.

Signal	Dataset	# of seqs.	Bucher		GA with opt.	
			Score	Seqs.	Score	Seqs
CAP	EPD Vertebrates	2199	3.364	581	5.001	917
	EPD Homo Sapiens	1796	3.469	489	5.447	790
	DCPD	205	7.497	114	9.106	102
TATA-box	EPD Vertebrates	2199	2.044	708	3.111	772
	EPD Homo Sapiens	1796	1.427	406	2.048	495
	DCPD	205	1.724	96	1.996	126

score in all the other positions. The results are shown in Table 6. A comparison of the scores obtained by our matrices and the one defined by Bucher (see Table 7) can be found in Table 5. Here the score is always computed on the range $[-36, -24]$. As the table shows, the matrix we obtained has always the greater score and describes a signal present in a higher number of sequences. By looking at the nucleotide frequencies, we can see that the matrices describe a sequence of `A` and `T` rich positions, with no definite preference for either nucleotide as in previous characterizations. Even if the absence of the usual `TATAA` sequence might look surprising, this result is consistent with the finding that the TBP (the TF part that recognizes the TATA box) recognizes the minor groove of DNA, where protein-DNA interactions are typically influenced by A/T-content, but not by the specific nucleotide sequence [12,13]. To our knowledge, this is the first time where a computational method was able to reproduce this result, without the canonical `TATAA` stretch in the signal. Anyway, further experimental investigation is needed, in order to establish whether the blurring of the `TATAA` motif depends on overlapping of occurrences of `TATAA` fragments in the surrounding positions, or actually describes an effective binding site for the TBP that is not strictly related to the usual consensus sequence.

6 Conclusions

In this paper we presented a method for the characterization of regulatory signals in genomic sequences, and we have shown its application to two important examples, the TATA box and CAP signals. The signals found have proved to be consistent with those described experimentally, as we have shown in the case of fruit fly signals. While more general than descriptions proposed in the past, our frequency matrices seem however to be able to characterize with better specificity the respective signals. The matrices obtained can be used to further investigate the mechanism of transcription regulation. For example, most of the genes usually present more than one possible point of transcription initiation. While on the dataset used for its construction (where the experimentally mapped TSSs had in many cases alternatives in the same sequence) the signal was present in less than half of the sequences, when we applied our CAP matrix to a selected set of gene sequences experimentally known to have a strong preference

Table 6. TATA-box signal obtained with the genetic algorithm on the EPD and DCPD datasets after the local optimization procedure.

	0	1	2	3	4	5	6	7
				EPD Vertebrates				
A	**0.372**	**0.471**	**0.538**	**0.640**	**0.794**	**0.643**	0.551	0.225
C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.163
G	0.135	0.000	0.000	0.000	0.000	0.151	0.274	0.465
T	**0.492**	**0.529**	**0.462**	**0.360**	**0.206**	0.206	0.175	0.148
				EPD Homo Sapiens				
A	**0.362**	**0.436**	**0.521**	**0.612**	**0.793**	**0.649**	**0.553**	0.239
C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.133
G	0.154	0.000	0.000	0.000	0.000	0.149	0.287	0.489
T	**0.484**	**0.564**	**0.479**	**0.388**	**0.207**	0.202	0.160	0.138
				DCPD				
A	**0.600**	**0.425**	**0.725**	**0.725**	**0.775**	**0.550**	0.325	0.250
C	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.400
G	0.000	0.000	0.000	0.000	0.000	0.325	0.425	0.350
T	**0.400**	**0.575**	**0.275**	**0.275**	**0.225**	0.125	0.000	0.000

Table 7. TATA-box signal obtained by Bucher in [4].

	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
A	0.157	0.041	0.905	0.008	0.910	0.689	0.925	0.571	0.398	0.144	0.213	0.211	0.211	0.175	0.198
C	0.373	0.118	0.000	0.026	0.000	0.000	0.008	0.005	0.113	0.347	0.378	0.326	0.303	0.275	0.260
G	0.391	0.046	0.005	0.005	0.013	0.000	0.051	0.113	0.404	0.386	0.329	0.329	0.329	0.357	0.360
T	0.080	0.794	0.090	0.961	0.077	0.311	0.015	0.311	0.085	0.123	0.080	0.134	0.157	0.193	0.183

for a single TSS, the percentage of sequences having the signal in the correct position rose to about 70%. Nowadays genomic data are often flanked by transcriptome analysis projects describing for each gene how many TSSs have been detected, the frequency with which each is used, as well as their precise location [14]. Therefore, an interesting study would be to further investigate possible correlations between the presence of TATA and CAP signals and the most frequently used TSSs of a gene. From the computational point of view, the main advantage of this method is the fact that, differently from previous approaches to the same problem, it does not make any prior assumption about the signal to be characterized, and also takes advantage of the specific localization of the signals considered. Moreover, in order to apply our method it is not necessary to select in advance a set of sequences containing the signal. In fact, the method finds the best partition of the dataset between sequences containing and non containing the signal. This distinction is done by computing the frequency matrix that gives the best score in the predefined signal position while penalizing all the other positions.

References

1. Bellorini, M., Dantonel, J.C., Yoon, J.B., Roeder, R.G., Tora, L., Mantovani, R.: The major histocompatibility complex class II EA promoter requires TFIID binding to an initiator sequence. Mol Cell Biol. **16** (1996) 503–512
2. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics **16** (2000) 16–23
3. Pavesi, G., Mauri, G., Pesole, G.: Methods for pattern discovery in unaligned biological sequences. Briefings in Bioinformatics **2** (2001) 417–430
4. Bucher, P.: Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. **212** (1990) 563–78
5. Hertz, G.Z., Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics **15** (1999) 563–77
6. Bailey, T., Elkan, C.: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning **21** (1995) 51–80
7. Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., Wooton, J.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262** (1993) 208–214
8. Wall, M.: (http://lancet.mit.edu/ga/)
9. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Massachusetts (1989)
10. Praz, V., Périer, R., Bonnard, C., Bucher, P.: The eukaryotic promoter database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. **30** (2002) 322–324
11. Kutach, A., Kadonaga, J.: The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol. Cell Biol. **20** (2000) 4754–64
12. Kim, J.L., Nikolov, D.B., Burley, S.I.: Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature **365** (1993) 520–527
13. Lo, K., Smale, S.T.: Generality of a functional initiator consensus sequence. Gene **182** (1996) 13–22
14. Okazaki, Y., Furuno, M., Kasukawa, T., et al: Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature **420** (2000) 563–573