10.2478/v10170-010-0028-1

Archives of Control Sciences
Volume 20(LVI), 2010
No. 4, *pages 491–501*

Position weight matrix model as a tool for the study of regulatory elements distribution across the DNA sequence

ROMAN JAKSIK and JOANNA RZESZOWSKA-WOLNY

Ab initio methods of DNA regulatory sequence region prediction known as transcription factor binding sites (TFBS) are a very big challenge to modern bioinformatics. Although the currently available methods are not perfect they are fairly reliable and can be used to search for new potential protein-DNA interaction sites. The biggest problem of *ab initio* approaches is the very high false positive rate of predicted sites which results mainly from the fact that TFBS are very short and highly degenerate. Because of that they can occur by chance every few hundred bases making the task of computational prediction extremely difficult if one aims to reduce the high false positive rate keeping highest possible sensitivity to predict biologically meaningful sequence regions. In this work we present a new application that can be used to predict TFBS regions in very large datasets based on position weight matrix models (PWM's) using one of the most popular prediction methods.

The presented application was used to predict the concentration of TFBS in a set of nearly 2.2 thousand unique sequences of human gene promoter regions. The study revealed that the concentration of TFBS further than 1kbp from the transcription initiation site is constant but it decreases rapidly while getting closer to the transcription initiation site. The decreasing TFBS concentration in the vicinity of genes might result from evolutionary selection which keeps only sites responsible for interactions with proteins being part of a specific regulatory mechanism leading to cells survival.

Key words: transcription factors, TFBS, regulation of gene expression, regulatory sequence elements, DNA, position-weight matrix, PWM

1. Introduction

Since the beginning of the Human Genome Project, which goal aimed to characterize the entire sequence of human DNA, the amount of methods used for pattern recognition and prediction of functional DNA sequence elements grew rapidly. DNA stores

The Authors are with Institute of Automatic Control, Silesian University of Technology, Akademicka str. 16, 44-100 Gliwice, Poland. E-mails: roman.jaksik@polsl.pl, joanna.rzeszowska@polsl.pl.

Parts of this work were presented at the National Conference on Application of Mathematics in Biology and Medicine in Krynica-Zdrój, Poland, 2010. This work was supported by the Polish Ministry of Science and High Education, Grant number N N514 411936.

Received 18.11.2010. Revised 13.12.2010.

an enormous amount of information used in the development and functioning of an entire organism. It works as a database of schematics needed to create proteins and ways of regulating their concentration in living cells. The size of such database ranges from hundreds of thousands nucleotides in bacteria to even hundreds of billions in vertebrates [1].

The exact knowledge about structure and role of each DNA fragment can bring huge benefits in many fields of science, but although the DNA structure was discovered nearly 60 years ago and 99% of human DNA sequence is known since the year 2003 our knowledge about its structure is still sparse.

One of the most interesting aspects of the genome is the self regulating process of transcription which copies the information stored in a specific DNA fragment known as gene to RNA molecule used in protein production process. Control of the gene expression processes involves highly sophisticated combinatorial interactions between proteins known as transcription factors and regulatory sequences in the genome represented by a specific combination of nucleotides. Transcription factors (TF's) are proteins, which by binding to specific DNA sequences control the transcription of adjacent genes [2], [3]. This function can be performed alone or in a complex with other proteins, by promoting or blocking the recruitment of RNA polymerase to specific genes [4]-[6].

Transcription factors are critical for the transcription process making sure that all of the genes maintain appropriate expression level depending on the changing requirements of the organism. They are found in all living organisms and the number of them increases with genome size - larger genomes tend to have more transcription factors per gene. It is estimated that there are approximately 2600 different proteins in the human genome that contain DNA-binding domains [7] and approximately 10% of genes in the genome code for transcription factors allowing unique regulation of each gene in the human genome [8].

DNA binding proteins interact with specific sequence patterns in promoter or enhancer regions of the DNA located in various parts of the genome. Enhancer regions have been reported upstream and downstream of genes, in 5'-UTR's, within introns, in 3'-UTR [9] and even within the coding sequence [10]. Depending on the transcription factor, the transcription of the adjacent gene is either up- or down-regulated which is done by the use of various regulation mechanisms [11]. These mechanisms include:

- stabilization or blocking the binding of RNA polymerase to DNA,
- histone acetyltransferase/deacetylase activity – weakens/strengthens the association of DNA with histones which makes the DNA more or less accessible to transcription and thereby controlling the amount of transcribed mRNA's [12],
- recruitment of coactivator or corepressor proteins to the transcription factor DNA complex [13].

The stabilization mechanisms and interaction sites can be quite different in various organisms, cell lines and even within a single cell, furthermore regulatory elements don't

just bind to one sequence but are capable of binding to closely related sequences, each with a different strength of interaction. One of the examples is the TATA binding protein (TBP) with TATATAA binding site [14]. It was proven that the TBP transcription factor can also bind similar sequences such as TATATATA, TATAAATA or TATATAAA [15]. This makes the analysis process very difficult, since it is hard to determine if the regulatory motif variant is responsible for protein interactions or occurs by chance and doesn't have any influence on the transcript stability regulation.

Chemically, transcription factors interact with their binding sites on the DNA by using a combination of electrostatic and Van der Waals forces which explains why transcription factors can bind not only to specific sequence fragments. However, not all bases in the transcription factor binding site may actually interact with proteins, making some of the interactions weaker than others. Because they can bind a set of related sequences and these sequences tend to be short we can expect that potential binding sites can occur by chance with frequency depending on the specificity level requirements of a given recognition motif. It is unlikely, that they can occur by chance in unwanted places, however if that happens other constraints, such as DNA accessibility in the cell or availability of cofactors may also help dictate where a transcription factor will actually bind.

2. The basics of PWM

Since regulation elements such as transcription factors do not bind only to specific sequence motifs but also to many similar sequences they cannot be presented as a simple DNA sequence like for example recognition sites for restriction enzymes. Restriction enzymes can cut the DNA in specific places determined by motifs such as GAATTC for the enzyme EcoRI. Single nucleotide substitution in the recognition site causes the enzyme to cut the sequence less by several orders of magnitude.

TFBS are much more tolerant to changes allowing multiple substitutions without losing their functionality. This causes two main problems in the DNA binding sites analysis. First is that we need to develop a representation motif based on a set of experimentally derived sequences that could be used to predict additional binding sites. Second problem is to discover the location of specific sites in a given sequence with highest possible sensitivity of the algorithm, to maximize the amount of detected functional sites, keeping lowest possible specificity reducing the false-positive identification rate.

Most popular representation of TFBS involves position-weight matrices (PWM's) which are simple mathematical objects with limited variability used to capture the information about local sequence patterns characteristic of a given function. PWM's are created based on finite number of experimentally derived motifs proven to be responsible for certain process like TF binding.

As an example let's consider one of the most popular TFBS known as TATA-box. Fig. 1a shows 6 closely related motifs responsible for TBP (TATA-box binding protein) interactions. PWM matrix is created by adding the occurrence number of each possible

nucleotide at each position creating an n by m table (Fig. 1b) where n is the length of the motifs and m is a constant number of unique bases in a nucleotide sequence.

(a)

GTATAAAAAGCGG
CTATAAAAGGCCC
GTATAAAGGGGCG
GTATATAAGCGCG
CTATAAAGGGGCC
GTATAAAGGCGGG

(b)

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0	0	6	0	6	5	6	3	1	0	0	0	0
C	2	0	0	0	0	0	0	0	0	2	2	4	2
G	4	0	0	0	0	0	0	3	5	4	4	2	4
T	0	6	0	6	0	1	0	0	0	0	0	0	0

Figure 1. A set of 6 experimentally derived TATA-box sequences (a) and respective PWM matrix (b).

Another form is to present the motif as a consensus sequence according to IUPAC nucleotide code, omitting specificity of the probability matrix. In such form the considered TATA-box motif would be presented as: STATAWARRSSSS, where $S = G$ or C; $W = A$ or T; $R = A$ or G. Sequence in such form could be then used to search for new sites with specific number of allowed mismatches, although this approach has many weaknesses as presented in [16].

PWM's are the most common way to represent TFBS patterns. There are few approaches used to create PWM's, from very simple ones like the one presented above to more complex involving three weight computation scheme [17] or neural networks [18], where the weights of created network correspond to the weights in the matrix. The biggest challenge is still the way of discovering new sites, based on the created PWMs, that would not only be statistically significant but also biologically meaningful.

One of the easiest methods is the one used by ConSite implementation [19]. Moving by a single nucleotide along the target DNA sequence it calculates each time the occurrence score by summing the respective linear or log scaled values from the rows of PWM for each nucleotide in the columns (Fig. 2).

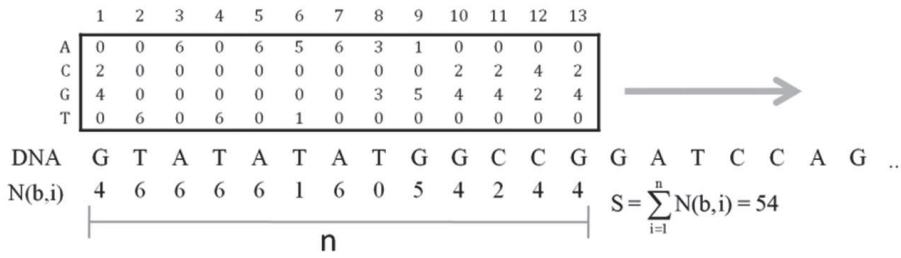

Figure 2. Example of a simple PWM score calculation method for the TATA-box sequence motif.

The higher the score is the more conserved is the analyzed sequence but since motifs can have various length and number of sequences used to create PWM can also change in a large range making each motif unique, scores S are normalized, by simply calculating the percentage of it comparing to the maximum possible score value, according to [20]:

$$S_{norm} = \frac{S - S_{min}}{S_{max-S_{min}}} \cdot 100 \tag{1}$$

where S_{min} and S_{max} are the smallest and highest possible scores that can be achieved depending on the specificity of the PWM matrix.

The score calculated for each position in the target sequence is compared to the user specified threshold (typically 80%) leaving only those sites which are characterized by large score values. The biggest advantage of this method is its simplicity making genome scale analysis of TFBS possible without the need of enormous computational resources.

Although the presented simple weight matrix model can work quite well for various datasets, some mathematical problems have to be faced before calculating the final score value. PWM's are created based on a sequences of limited variability and because of that we often do not observe each base at each position in the set of analyzed motifs resulting in zeros present in the PWM. This causes obvious problems when transforming the table to log scale. One possible solution is to transform the PWM scores according to notation proposed in [21]:

$$W_{b,i} = \log \frac{P_m(b,i)}{P_b(b)} \tag{2}$$

where $P_b(b)$ is the background probability of base b (in most cases $P_b(b) = 0.25$ for $b = 1,\ldots,4$) and $P_m(b,i)$ is the corrected probability of occurrence for base b at position i in the motif of length m, and is calculated by:

$$P_m(b,i) = \frac{N(b,i)}{n} + \varepsilon \tag{3}$$

where $N(b,i)$ are the occurrence counts of base b in position i from the PWM matrix and ε is a smoothing parameter (usually $\varepsilon = 0.01$) which prevents the logarithm problem.

The score for each target site would be then given by a sum of all $W_(b,i)$ values for each base in the analyzed sequence:

$$S = \sum_{i=1}^{n} W_{b,i}. \tag{4}$$

Another obvious problem arises when looking at Eq. (2). $P_b(b)$ value is assigned based on assumption that the DNA sequence is random with the probability of each base occurrence equal to 0.25. This goes quite well when taking into account only small sequence fragments but for the entire genome such assumption is obviously false. Genomes of some species can show large global differences in G and C nucleotide content comparing to A and T like for example 72% of $G+C$ in the genome of *Streptomyces coelicolor* bacteria. Additionally, the differences can sometimes occur also locally. According to isochors genome organization theory there are large parts of DNA sequence in most of the vertebrates that differ significantly in GC content from their surroundings

[22]. It is obvious then that *GC* rich regions will be overrepresented by *GC*-rich motifs while *AT*-rich will not occur by chance as often.

Also, there is a problem concerning score threshold that one should use to predict sites, keeping appropriate balance between precision and sensitivity when comparing the prediction results to experimentally derived sites. Methods of assessing the statistical significance of PWM matches can be a big challenge as well. Based on the distribution of all possible distinct similarity scores some methods were proposed to calculate p-value which describes the probability of a background model achieving a score higher or equal to the observed value [23], [24].

Other more precise methods of TFBS discovery are presented in literature but because of much larger computational complexity they are mostly oriented towards small dataset analysis. The methods include Bayesian networks [25], permuted Markov models [26] or non-parametric models [27] but not always were proven to show significant improvements over the simple weight matrix models while requiring much more computational resources or additional experimental data which are typically unavailable.

3. Global prediction of TFBS using NuceleaSeq application

Many implementations of TFBS search algorithms are available and some of the example tools are: ConSite [19], TFBS [28], Paster [29], Match [30], rVista [31] and Mapper [32] but they focus mostly on the analysis of relatively short specific sequence fragments and therefore are not applicable to global genome analysis either because of the complexity of the search algorithm which requires enormous computational resources or because of the results presentation form. Because of that, a new implementation was made, being a part of NucleoSeq application, based on TFBS detection methods presented in this article. The application is oriented mostly towards the analysis of extremely large data sets, either provided directly by the user or downloaded automatically from the internet based on various gene accession numbers.

Unlike most of the mentioned applications NucleoSeq uses resources of a local computer therefore it doesn't depend on accessibility, stability and performance of remote server. It is very easy to use and can provide results in a form of precise report of each TFBS occurrence count or genomic location, for each individual sequence and also as an overall summary for the entire sequence set. Such approach allows to create TFBS distribution maps in a very rapid and easy way.

NucleoSeq can be freely downloaded from our website at:
`www.bioinformatics.aei.polsl.pl`.

The application was used to search for the occurrence of all 75 known human TFBS derived from the Jaspar database [33] in a set of 21818 unique gene promoter regions defined as a sequence fragments 1 to 5000 base pairs (bp) upstream from the transcription start site (TSS) of known human Reference Sequence transcripts derived from the USCS hg19 assembly of the human genome. The same analysis was performed for a

set of artificial sequences created based on the original USCS dataset by shuffling the nucleotide order.

While calculating the TFBS occurrence 3 different approaches were used:

1. Simple score - which is a sum of respective PWM counts in linear scale normalized according to equation (1).

2. Log-odd score - most popular approach described by equations (1)-(4) with simplified background probabilities of each base occurrence equal to 0.25. item Log-odd score with variable background calculated independently for each analyzed 5000 bp long sequence fragment based on its individual nucleotide composition

Similarity cutoff was set to 80% for all three methods.

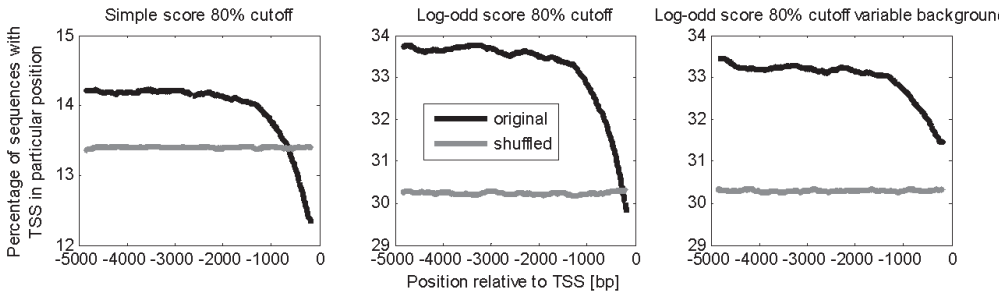

Figure 3. Distribution of all human TFBS from the Jaspar database across 5kbp sequence regions upstream from the transcription initiation site (moving average with 300bp window).

The overall amount of found motifs further than 1000bp from TSS is significantly higher than in the same sequences with randomly displaced nucleotides and gets lower while getting closer to the transcription start site (Fig. 3). Type of the selected method doesn't change significantly the shape of the distribution but as expected has a huge impact on the overall level of detected sites and the difference between random set of sequences. The overall amount of detected sites seems to be extremely high suggesting that TFBS occur every few bases. In reality TFBS motifs have a relatively small variability which leads to a very high number of counts overlapping each other and forming large TFBS clusters.

Method based on variable background probability scores shows much less sites, especially close to TSS, which results from large variations in *G* and *C* nucleotides count between analyzed sequences. Additionally the average *GC* percentage increases significantly in the vicinity of TSS which has a huge impact on TFBS concentration. This shows that even selecting individual background for each sequence fails in such situations since within a single sequence the nucleotide distribution is highly inconsistent with the average amount.

4. Concluding remarks

The presented application was used to predict the concentration of TFBS in a set of nearly 2.2 thousand unique sequences of human gene promoter regions leading to a discovery that the concentration of TFBS further by 1kbp from the transcription initiation site is constant and higher than expected but it decreases rapidly while getting closer to the transcription initiation site. TFBS were expected to increase in count in those regions since they are known to be the basic mechanism of transcription initiation. The decreasing TFBS concentration might result from evolutionary selection which keeps only sites responsible for interactions with proteins being a part of a specific regulatory mechanism.

Many scientists are sceptical to the computational methods of TFBS discovery which are in many cases inconsistent with experimental approaches like ChIP-Sequencing or ChIP-on-chip microarrays. It is a big concern when dealing with individual genes regulated by specific TF since the actual binding of a protein depends not only on the nucleotide order of specific DNA fragment but additionally on the accessibility of the site and concentration of specific TF proteins. Many recognition sites overlap each other forcing TF to compete over the selected region favoring those proteins which are over expressed due to various factors including cell type, developmental stage and environmental conditions. Fig. 3 shows that when focusing on a global distribution analysis we can assume that even if the rate of false positive or negative counts is high resulting from random motif occurrence they will be equally distributed across the sequences and therefore the shape of the distribution would not be significantly affected.

Transcription factors are the main regulatory elements of complex information processing mechanism present in the genome, conserved trough all living organisms, therefore methods of *ab initio* transcription factor binding sites (TFBS) prediction that would be biologically meaningful are very important and set a very big challenge to modern bioinformatics. Although the currently available algorithms are not perfect they are fairly reliable and can be used to search for new potential DNA binding sites in genomes of various organisms leading to discoveries of previously unknown regulatory interactions. The biggest problem of *ab initio* approaches is the very high false positive rate of predicted sites since many sequence regions are inaccessible to proteins and therefore cannot interact with it. Another big challenge is to predict TFBS clusters as they often interact with proteins cooperatively. In order to take the full advantage of genomic sequences and be able to determine the regulatory network features based on the sequence alone there is still much to be done in the field of TFBS discovery. But even now large scale computational methods can provide useful indication for further experimental work which normally could not be performed in such extent because of enormous costs and lack of fast and efficient experimental methods.

References

[1] T.R. GREGORY, ET AL.: Eukaryotic genome size databases. *Nucleic Acids Res.*, **35** (2007), D332-D338.

[2] D.S. LATCHMAN: Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29** (1997), 1305-1312.

[3] M. KARIN: Too many transcription factors: positive and negative interactions. *New Biol.*, **2** (1990), 126-131.

[4] R.G. ROEDER: The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **21** (1996), 327-335.

[5] D.B. NIKOLOV and S.K. BURLEY: RNA polymerase II transcription initiation: a structural view. *Proc. Natl. Acad. Sci. USA*, **94** (1997), 15-22.

[6] T.I. LEE and R.A. YOUNG: Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, **34** (2000), 77-137.

[7] M.M. BABU, ET AL.: Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14** (2004), 283-291.

[8] A.H. BRIVANLOU and J.E. DARNELL, JR.: Signal transduction and the control of gene expression. *Science*, **295** (2002), 813-818.

[9] M. LEVINE and R. TJIAN: Transcription regulation and animal diversity. *Nature*, **424**, (2003), 147-151.

[10] K.K. BARTHEL and X. LIU: A transcriptional enhancer from the coding region of ADAMTS5. *PLoS One*, **3** (2008), e2184.

[11] G. GILL: Regulation of the initiation of eukaryotic transcription. *Essays Biochem.*, **37** (2001), 33-43.

[12] G.J. NARLIKAR, H.Y. FAN and R.E. KINGSTON: Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, **108** (2002), 475-487.

[13] L. XU, C.K. GLASS and M.G. ROSENFELD: Coactivator and corepressor complexes in nuclear receptor function. *Curr. Opin. Genet. Dev.*, **9** (1999), 140-147.

[14] J.M. WONG and E. BATEMAN: TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs. *Nucleic Acids Res.*, **22** (1994), 1890-1896.

[15] F. MUKUMOTO, ET AL.: DNA sequence requirement of a TATA element-binding protein from Arabidopsis for transcription in vitro. *Plant Mol. Biol.*, **23** (1993), 995-1003.

[16] W.H. DAY and F.R. MCMORRIS: Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.*, **20**(5), (1992), 1093-1099.

[17] J.M. CLAVERIE and S. AUDIC: The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12** (1996), 431-439.

[18] G.D. STORMO, T.D. SCHNEIDER and L.M. GOLD: Characterization of translational initiation sites in E. coli. *Nucleic Acids Res.*, **10** (1982), 2971-2996.

[19] A. SANDELIN, W.W. WASSERMAN and B. LENHARD: ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, **32** (2004), W249-W252.

[20] P. BUCHER: Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212** (1990), 563-578.

[21] K. MASUDA, ET AL.: Androgen receptor binding sites identified by a GREF_GATA model. *J. Mol. Biol.*, **353** (2005), 763-771.

[22] G. BERNARDI: Isochores and the evolutionary genomics of vertebrates. *Gene*, **241** (2000), 3-17.

[23] H. TOUZET and J.S. VARRÉ: Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for Molecular Biology*, **2** (2007).

[24] J. ZHANG, B. JIANG, M. LI, J. TROMP, X. ZHANG and M.Q. ZHANG: Computing exact P-values for DNA motifs. *Bioinformatics*, **23** (2007), 531-537.

[25] Y. BARASH, ET AL.: Modeling dependencies in protein-DNA binding sites, Proceedings of the seventh annual international conference on Research in computational molecular biology, Berlin 2003, pp. 28-37.

[26] X. ZHAO, H. HUANG and T.P. SPEED: Finding short DNA motifs using permuted Markov models. *J. Comput. Biol.*, **12** (2005), 894-906.

[27] O.D. KING and F.P. ROTH: A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **31** (2003), e116.

[28] B LENHARD and W.W. WASSERMAN: TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18** (2002), 1135-1136.

[29] G.Z. HERTZ, G.W. HARTZELL, 3RD and G.D. STORMO: Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6** (1990), 81-92.

[30] A.E. KEL, ET AL.: MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31** (2003), 3576-3579.

[31] G.G. LOOTS and I. OVCHARENKO: rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32** (2004), W217-W221.

[32] V.D. MARINESCU, I.S. KOHANE and A. RIVA: MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6** (2005), 79.

[33] A. SANDELIN, ET AL.: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32** (2004), D91-D94.