Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

DNA Motifs Detection Algorithms in Long Sequences

Alin G. VOINA, Petre G. POP, Mircea F. VAIDA
Communications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
alin.voina@com.utcluj.ro

Abstract— **The identification of DNA motifs remains an active challenge for the researchers in the bioinformatics domain. A considerable effort in this area was concentrated on understanding the evolution of the genome by identifying the DNA binding sites for transcription factors. The evolution in genome sequencing has led to the appearance of numerous computational methods for finding the short DNA segments, known as motifs. In this study, we present some of the computational methods that exist and try to evaluate their performance in case of long sequences.**

Keywords - Computational methods, motif finding algorithms, binding sites, bioinformatics, motifs localization

I. Introduction

In recent years, a considerable number of algorithms have been designed for identifying novel regulatory elements in DNA sequences. Most of the algorithms were developed with the scope to detect the regulatory elements by taking into consideration the regulatory regions of several co-regulated genes that belong to a single genome [1]. The algorithms identify the regulatory regions and search for overrepresented motifs which are classified as potential candidates for regulatory elements. Some of the algorithms use phylogenetic footprinting to identify well conserved sites among regulatory regions. Most of the algorithms perform over multiple sequences or an entire gene. The gene is the fundamental unit of inherited information in deoxyribonucleic acid (DNA) and is defined as a section of base sequences that are used as a template for the copying process called transcription. The process of regulating gene expression is done with the help of transcription factors by activating or inhibiting the process of transcription. An important challenge in molecular biology is to completely understand the mechanism that regulates gene expression. A major task in this challenge is to identify the DNA binding sites for transcription factors. Computational methods are expected to offer promising solutions and as consequence, researchers have invested considerable efforts into these methods.

The detection of regulatory elements problem may be announced in the following form: considering a group of N sequences, look for a pattern M of length l which is found frequently. If the pattern M of length l occurs in each sequence from the group of N sequences, then a simple enumeration of the l letters of the pattern M gives the regulatory element. The main issue when we are dealing with the analysis of DNA sequences is that the regulatory elements by which we search may have mutations of nucleotides.

In the analysis of the mechanism that regulates gene expression, sequence motifs have become very important. A DNA motif can be defined as a short, recurring pattern in DNA that is presumed to have some biological function (often they represent binding sites for transcription factors -TF) [2]. A part of the motifs contribute to complex processes that occur at the RNA level, including ribosome binding, mRNA processing and transcription termination [3]. Motifs are relatively short – they have a length between five and twenty base pairs (bp) and can be localized in different genes or even within the same gene. Besides this classification based on length, there are two special types of motifs that are recognized: palindromic motifs and space dyad (gapped) motifs [4]. We called a motif palindromic if it matches its complementary base sequence read backwards (for example 'TCTCGCGAGA' it's a palindromic motif). Motifs that are formed from two sites of short length, well conserved and usually separated by a spacer are called space dyad (gapped) motifs. Because the transcription factor (TF) usually binds as a dimer, the gap is often located in the middle of the motif. Typically, the positions where TF binds to the DNA are well conserved and have a length between three and five base pairs.

In the past, methods like footprinting and gel-shift or reporter construct assays [5] were used for determining binding sites. Nowadays, computational methods are strongly involved in determining overrepresented DNA patterns in a sequence or set o sequences. A motif is overrepresented if it's encountered more often into the analyzed sequence than one would expected by chance [4]. Most of the motifs finding algorithms have great results in the case of lower organisms (including yeast) but their performance is relatively poor in the case of higher organisms. Some recent algorithms, which use phylogenetic footprinting, proved to be more efficient in motif detection for genomic sequences [5].

II. Motif Detection Algorithms

In the recent decade, the algorithms that search for DNA motifs have known a spectacular growth and nowadays we can count more than sixty elaborated methods. Most of these approaches rely on probabilistic models or phylogenetic footprinting.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

Motif detection algorithms can be classified into three major classes [4]:

(1) algorithms that use promoter sequences from co-regulated genes of a single genome;
(2) algorithms that use phylogenetic footprinting;
(3) a combination of (1) and (2).

In earlier literature, motif detection algorithms were classified in two major groups [4]:

- word- based (string-based) methods based on exhaustive enumeration, i.e., counting and comparing oligonucleotide frequencies;
- probabilistic models based on the maximum likelihood principle; the Bayesian inference.

The first group (word-based methods) is more suitable for finding short motifs such as those encountered in eukaryotic genomes. String-based enumerative methods can be relatively fast especially when are implemented with the help of some optimized data structures (e.g. building a suffix tree of the sequences and are very useful when searching for identical instances of a motif). In this study we compare seven of the most frequently used tools: AlignACE, MEME, Improbizer, Weeder, YMF, Scope and LocalMotif. In Table I, we can find a short description for each of these algorithms.

TABLE I. OPERATION PRINCIPLES FOR ANALYZED TOOLS

Analyzed tool	**Principle of functionality**	**Observations**
AlignACE	Gibbs sampling strategy	Distinct motifs are found by using an iterative masking procedure.
MEME	Searches for statistically significantly motifs in the input dataset	It performs expectation maximization (EM) from starting points derived from each subsequence occurring in the input dataset
Improbizer	Weight matrices of DNA motifs are determined using Expectation Maximization	As a particularity, Improbizer uses a Gaussian model for motif localization, so that motifs that are placed in similar positions are more likely to be determined.
Weeder	Consensus-based method	Evaluations are made according to the number of appearance in sequences and how well it's conserved in each sequence. Consensus-based algorithm it's providing a weight matrix which is used to select the best instances of each motif.
YMF	Looking for motifs that have the greatest z-score.	Motifs are reported as sequences over IUPAC alphabet.
SCOPE	Ensemble learning method which combines the outputs of three component algorithms.	Combines methods for the discovery of short non-degenerate motifs, short degenerate motifs and long highly degenerate motifs. As a particularity SCOPE requires as inputs only the sequence to be analyzed and a species selection [7].
LocalMotif	Based on a novel scoring function called spatial confinement score.	The approach successfully discovers biologically relevant motifs and their intervals of localization in scenarios where the motifs cannot be discovered by general motif finding tools. It is especially useful for discovering multiple co-localized motifs in a set of regulatory sequences, such as those identified by ChIP-Seq [8].

III. LOCALISED MOTIFS DETECTION IN LONG SEQUENCES

The discovery of motifs in long regulatory sequences proves to be a current requirement for vertebrate promoters [8], especially in ChIP experiments. Some recent studies [10, 11] revealed the fact that in the case of long sequences, random patterns might become at least as prominent as the real motif and the motif finding algorithms will report false positives that overshadow the real motif. For most motif discovery algorithms, the resources needed; time and memory requirements, increased proportionally with the length of the sequence analysed.

In literature, it is recognized that the binding sites usually occur within the sequence in a specific position relative to a biological landmark [8]. Many Transcription Factor Binding Sites –TFBS are located relative to TSS to allow TFs to anchor at specific positions with respect to each other and the TSS [12]. In these specific situations, the detection of the motif can be done by searching in an appropriate interval after the sequence is aligned relative to an anchor point. One of the big advantages of the localization is that it removes the regions that are not containing any motif and decrease the probability of reporting false positive motifs.

A possible solution is to subdivide long sequences into short overlapping subsequences which have the same length and analyze each subsequence with a motif finding algorithm. Some problems can appear using this approach:

- in most cases we don't know *a priori* the region where motifs are localized
- in the case of a large number of motifs reported within different intervals it is relatively difficult to identify the motifs that are most relevant for the entire analyzed sequence
- the length of the sequences must be chosen carefully (if it is too short then the motif may not appear prominently and if it is too long, the motif may be overshadowed)

- the division of the analysed sequence into subsequences should be automated, otherwise it will be time-consuming and prone to error.

In this study, we have chosen not to split the sequence into subsequences, to not disadvantage any of the algorithms. The analysis was made on the entire sequence so that we evaluate the performance of the algorithms, on long sequences as we find them in genome databases.

IV. Experiments And Results

One of the biggest challenges in our experiments was to get proper datasets such as to not disadvantage any of the applications that we have used for performing the tests. Tompa et al. [13] identified several possibilities for choosing datasets:

- use real sequences that contain real annotated Transcription Factor Binding Sites (TFB) – the main disadvantage of this method could be the fact that there can exist unannotated binding sites and the tools that predict these will be unfairly penalized;
- use synthetic DNA sequences where we could implant at random positions instances of known motifs – the main disadvantage of this approach is that we can favor some tools over others because of the stochastic process that nature uses.

To overcome the main drawbacks of the above possibilities, we used TRANSFAC database, from which we extracted real TF (Transcription Factors). From the TRANSFAC database, we chose the TF that also had a consensus sequence defined. The binding sites for transcription factors are planted at known positions and orientations.

Each TF gives rise to a dataset of sequences which can have one of three types as a background sequence (real promoter sequences, generic promoter sequence, sequences generated using a Markov chain). By using each type as a background sequence we tried to overcome the drawbacks described above.

The tests were performed on long sequences for three species: *Drosophila Melanogaster*, *Homo Sapiens* and *Saccaromyches Cerevisiae*. All of the tools were set to report motifs of a variable length between 6 and 10 bp (this is because not all of the analyzed tools allowed us to specify a fixed length for the motif).

In this study we've take into consideration the top 10 motifs reported by each of the tested tool. To obtain an overview of the performance of each motif detection tool we've run the chosen datasets on each application.

All tested applications were used without modifying any parts of the code and all the tests were run directly on their dedicated web pages or by downloading the executable file locally.

In the tables bellow we've summarized the results obtained for each of the dataset.

TABLE II. RESULTS OBTAINED IN CASE OF *DROSOPHILA MELANOGASTER*

Application	Common motifs reported	Matching TRANSFAC site	TF	Detection accuracy
LocalMotif	TCGAAGGG	TCGAAGGGATTAG	T00456 Kr	10%
	CTCGGGAG	*	*	
AlignAce	AGAGAG	AGAGAGAG	T00301 GAGA	50%
	AGCGAG	*	*	
	CATGGC	CATGGCAGC	T00626 NIP	
	GAGAGA	AGAGAGAG	T00301 GAGA	
	AGTGAG	CGAGTGAGTG	T00918 Zeste	
	GAGCGA	*	*	
	GAGTGA	CGAGTGAGTG	T00918 Zeste	
Improbizer	ACGGTC	*	- *	0%
	TGCGCCA	*	- *	
	TGCGAT	*	- *	
MEME	CCCCAC	ATCACCcCAC	T01559 SREB	40%
	TCCCGC	TGTCCCGC	T01542 E2F-1	
	AGGCAA	*	*	
	TTCGCC	*	*	
	CTCGCC	*	*	
	GAGGAG	TGGgAGGAGC	T00754 Sp1	
	AAAATC	CAGAAaAATC	T00196 Dl	
YMF	AGAGAG	GAGAGAGaG	T00301 GAG	60%
	GAGTGA	CGAGTGAGTG	T00918 Zeste	
	AGCGAG	*	*	
	CATGGC	CATGGCAGC	T00626 NIP	
	GAGAGA	AGAGAGAG	T00301 GAG	
	AGTGAG	cGAGTGAGTG	T00918 Zeste	
	GAGCGA	*	*	
	CACTCA	CACTCA	T00321 GCN4	
Scope	CGCTC	CGCTCCC	T00301 GAG	50%
	AGCACC	TCAGCACCG	T00272 Eve	
	CATGGC	CATGGCAGC	T00626 NIP	
	GGCGA	TGGCGA	T00275 F-AC	
Weeder	CGCTCG	*	*	20%
	CGAACC	*	*	
	CGGTTA	*	*	
	TTAGCG	TTAGCGC	T00221 E2F	
	GAGTGA	cGAGTGAGTG	T00918 Zeste	
	TCGCTC	*	*	

In Table II, we have introduced only the common motifs that were reported by at least two applications, in the case of the *Drosophila Melanogaster* dataset. The number of common motifs varied between two, in the case of LocalMotif application, and eight, in the case of the YMF tool. Also, we observed that some of the common motifs reported were real transcription factor binding sites annotated in TRANSFAC database (Improbizer was the only one that didn't report any motif with a correspondence in transcription factors that we found in the TRANSFAC database).

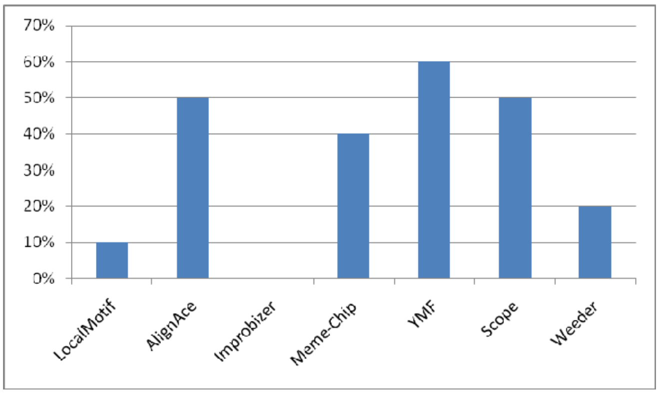

Figure 1. Detection accuracy of the analyzed tools in case of *Drosophila Melanogaster* dataset

From the performance point of view, as we can observe in Fig.1, we can confirm that the most accurate results were reported by YMF (Yeast Motif Finder) because 6 motifs from the top 10 taken into consideration, were identified as real transcription factors .

TABLE III. RESULTS OBTAINED IN CASE OF *HOMO SAPIENS*

Application	**Common motifs reported**	**Matching TRANSFAC site**	**TF**	**Detection accuracy**
LocalMotif	GCGGCGCG	*	*	20%
	GGCGCCTC	*	*	
	CGCTCGCG	*	*	
	TGCCGGCA	TTTCGAAA	T00108 C/EB	
	TTTCGAAA	TTTCGAAA	T00108 C/EB	
MEME	CCTCCC	CCCTCCC	T00711	80%
	ACCCCT	ACCACCCCTC	T00759 Sp1	
	CCTGCC	GCCCTGCCCC	T00759 Sp1	
	CACCTC	CCACCTCT	T00625 AREB6	
	CCCGCC	CCCCGCC	T00759 Sp1	
	GGCCCC	GGCCCC	T00759 Sp1	
	CTCCAC	*	*	
	CTCCTC	CCCTCCTC	T00105 C/EB	
	CACCTC	TCACCTCT	T00625 AREB	
	CAGCTC	*	*	
YMF	AAAAAA	TAAAAAA	T00794 TBP	100%
	AATAAA	AAATAAA	T00798 TBP	
	AAAATA	AAAAAAAtAA	T00395 Hb	
	ATGAAT	ATGAATG	T00691 POU1	
	AAATAA	AAATAAA	T00798 TBP	
	AAACAA	AAACAAA	T00371 HNF	
	ATTTTA	ATTTTA	T00691 POU1	
	ACAAAA	AACAAAA	T02878 TCF	
	ATAAAT	AATAAATA	T00821 TFIID	
	AAAAAT	AAAAAAAtAA	T00395 Hb	
Weeder	CCTATC	GCCTAtCAAT	T00306 GATA	40%
	TCAACG	TCAACGG	T00305 GATA	
	CTATCC	CTATCC	T00305 GATA	
	GGTAAG	ACAGGTAAG	T00625 AREB	
Scope	TGAATC	TGAATCA	T00122	20%
	GTACCAGG	*	*	
	GGATTGAC	*	*	
	ATTCCC	ATTCCC	T01469 Ik-1	
Improbizer	TCAGAGTA	*	*	0%
AlignAce	GAGGCGGG	GAGGCGGGGC	T00758 Sp1	10%

In Table III, we introduced the common motifs reported by the analysed applications, in case of the *Homo Sapiens* dataset. We must specify that in this case the length of the analyzed sequence was over 36000 bp. What we observed was the fact that the tools reported with success short motifs (as real motifs annotated also in TRANSFAC database) but once we increased the length, the motifs were identified as false positives.

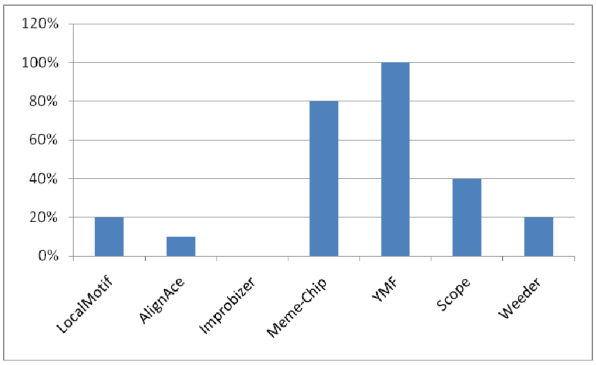

Figure 2. Detection accuracy of the analyzed tools in case of *Homo Sapiens* dataset

From Fig.2, we can also observe that, in this case, YMF was the most accurate and all of the top 10 motifs were identified as being part of real transcription factors reported in TRANSFAC database for the analyzed sequence. We also observed that MEME had an accuracy of detection relatively close to YMF. The detection accuracy in the case of LocalMotif, AlignAce and Weeder was under 20%.

In Table IV, we gathered the common motifs reported by all seven tools in the case of a DNA sequence that belong to *Saccaromyches Cerevisiae* species. The length of the tested sequence was approximately 16000 base pairs.

TABLE IV. RESULTS OBTAINED IN CASE OF *SACCAROMYCHES CEREVISIAE*

Application	**Common motifs reported**	**Matching TRANSFAC site**	**TF**	**Detection accuracy**
MEME	CAGCGG	*	*	60%
	CCGAGC	CTaATCCGAGC	T00322 GCR1	
	TGCCTG	TGCCTGG	T00528	
	CCTGCA	GCCTGCAGGC	T00035 AP	
	CGACCC	*	*	
	GCGGGT	*	*	
	GGAGCC	GGAGCC	T01944 NF	
	AACGTC	*	*	
	AAAAAA	AAAAAAAtAA	T00395 Hb	
	ACAAAC	ACAACAAACA	T04169 FOX	
YMF	AAAAAA	AAAAAAAtAA	T00395 Hb	60%
	CGGCTA	TCGGCGGCtA	T01247 UME	
	GACCCC	TGACCCC	T01331 RXR	
	CGCCGA	TAGCCGCCGA	T01247 UME	
	AACGCG	*	*	
	CTAGCC	*	*	
	GCTAGA	*	*	
	CCCGCG	CCCGCGC	T00034 AP-2	
	AAACGC	AAACGC	T00137 c-Myb	
Weeder	CCCGCG	CCCGCGC	T00034 AP-2	50%
	GCGGAC	*	*	
	CGACCC	*	*	
	CTAGCC	*	*	
	CCCGAC	*	*	
	TAGCCG	TAGCCGCCGA	T01247 UME	
	AGCCGC	AGCCGCC	T02658 EBP	
	GCCCCC	GCCCCCTCCCC	T00759 Sp1	
	CGCCGA	GCCCCCTCCCC	T00759 Sp1	
LocalMotif	ATCCGAGT	*	*	0%
AlignAce	TCAGCGG	*	*	0%
	GAAGGG	*	*	
Scope	CCGCCC	CCGCCC	T00270 ETF	30%
	CCAGC	CCAGCCA	T00368 HNF-1	
	GCGGGAA	*	*	
	ATTCC	ATTCCC	T01469 Ik-1	
	AGAACCG	*	*	
	CCGCACA	*	*	
Improbizer	CCCGCATC	*	*	0%

In this case, the number of false positives was significantly higher. If we look at the results reported by LocalMotif, AlignAce and Improbizer we can observe that from the top 10 motifs taken into consideration, just one was also reported by another application. Likewise, none of the motifs reported were indentified as being part of a real transcription factor binding site.

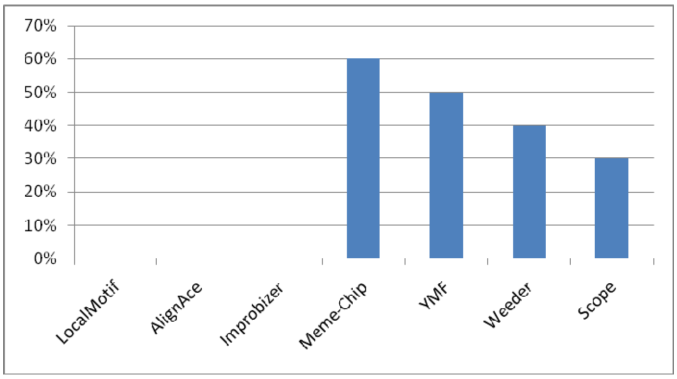

Figure 3. Detection accuracy of the analyzed tools in case of human genes dataset

As a performance particularity (see Fig.3) we should also mention that in this case MEME and YMF proved to be the most accurate on detecting real motifs.

CONCLUSIONS

In this investigation, we looked at some of the most frequently used algorithms for motif detection and presented their performance, especially in the case of long sequences.

Although the efforts put in to exploring computational biology and genome sequencing is relatively significant, the prediction of regulatory elements remains a challenging task for biologists.

At the beginning, the motif search algorithms were based on co-regulated genes and the search process was mainly focused on overrepresented motifs. Nowadays, computational methods have a special place and a considerable effort has been made in designing algorithms that make use of computational methods.

Because there are a considerable number of methods/algorithms that are specialized in motif identification, for a user it would be very useful to have some guidelines in choosing a tool to perform an analysis of a DNA sequence. The main drawback in offering guidelines for choosing a specific method or an algorithm is the diversity of parameters and settings that each algorithm have in particular. However, the evaluation of their performance is relatively difficult. This is primarily due to the fact that the mechanism that regulates gene expression is not yet fully understood and more than that, we don't have a complete model or a predefined standard so that we can measure the effectiveness of each algorithm. In the tests done in this study (against long DNA sequences) we should take into consideration that the detection accuracy calculated for each algorithm is only indicative and might be prone to errors because the applications were configured by human choices of parameters.

Most of the algorithms had good results for low organisms and especially when configured to report short motifs (usually between six and eight base pairs). More recent algorithms, that integrates the overrepresentation of motifs and their conservation between species, proved to be more efficient in the case of both lower and higher organisms, including the human genome, as shown here. The prediction of motifs in long sequences or over an entire genome still remains a challenge for biologists primarily due to the complexity of regulatory genomics. However, from our research we have concluded that YMF and MEME proved to be the most accurate in detecting motifs in long sequences because a large majority of the reported motifs were annotated as binding sites for real transcription factors.

As a particularity of this assessment, we observed that some of the tools perform better on some specific datasets and concluded that is better to use multiple tools as a complementary source of information. Therefore, it's difficult to nominate an algorithm or a method as the best one for analyzing long sequences and we indicate to take into consideration the top ten motifs reported by the method or algorithm.

ACKNOWLEDGMENT

This paper was supported by the project "Doctoral studies in engineering sciences for developing the knowledge based society-SIDOC" contract no. POSDRU/88/1.5/S/60078, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] Mathieu Blanchette, Martin Tompa, "Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting", *Genome Res*, 2002, pp. 739-748.

[2] Patrik D'haeseleer, "Whar are DNA motifs?", *Nature Biotechnology 24,* 423-425 (2006)

[3] Down TA, Hubbard TJ., NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence, *Nucleic Acids Res*. 2005 Mar 10;33(5):1445-53.

[4] Modan K Das, Ho-Kwok Dai, A survey of DNA motif finding algorithms*, BMC Bioinformatics*, 2007

[5] Down TA, Hubbard TJ., NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence, *Nucleic Acids Res*. 2005 Mar 10;33(5):1445-53.

[6] Ao, W., Gaudet, J., Kent, W.J., Muttumu, S. & Mango S.E, Environmentally Induced foregut remodelling by PHA-4/FoxA and DAF-12/NHR, *Science* 305, 1743-1746 (2004)

[7] Jonathan M. Carlson, Arijit Chakravarty, Charles e. DeZiel, Robert H. Gross, SCOPE: a web server for practical de novo motif discovery, *Nucl. Acids Res,* (2007)

[8] Vipin Narang, Ankush Mital, Wing-Kin Sung, "Localized Motif discovery in gene regulatory sequences", Bioinformatics, vol.26, no.9, 2010

[9] Roth FP, Hughes JD, Estep PW, Church GM, Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantization, *Nature Biotechnology* 1998, 16:939-945.

[10] Keich, U., and Pevzner, P.A. (2002a) "Finding motifs in the twilight zone." *Bioinformatics,* 18(10), 1374-1381.

[11] Buhler, J., and Tompa, M. "Finding motifs using random projections.", J Comput Biol, 9(2), 225-242.

[12] Smale, S.T. and Kadonaga, J.T –"The RNA polymerase II core promoter", Annu Rev Biochem, (2003), 449-479

[13] Martin Tompa, Nan Li, et. Al "Assessing computaional tools for the discovery of transcription factor binding sites", Nature Biotechnology, Volume 23, 2005, 137-145