International Journal of Computer Applications (0975 8887)
Volume 94 - No. 11, May 2014

Computational Approaches for Transcription Factor Binding Prediction

Smitha C S
MTech Scholar
College of Engineering Trivandrum
Kerala

Saritha R
Assistant Professor
College of Engineering Trivandrum
Kerala

ABSTRACT

The complexity of any organism depends on cellular behavior which is formed from the interaction between proteins and DNA. The accurate identification of DNA protein interaction is still a mirage. So computational approaches are used for the prediction of DNA protein interaction and the prediction results can be experimentally verified. In this paper a detailed survey of some of the prediction approaches in both prokaryotes and eukaryotes have been made. The prediction results may be useful for identification of putative binding sites.

General Terms:

Genome, Gene, Proteins, Transcription Factor, Binding Site

Keywords:

DNA, Binding Features, Position Specific Scoring Matrix, Consensus, Binding affinity

1. INTRODUCTION

Genes are the fundamental unit of hereditary information. The genome of an organism carries the details regarding traits to be exhibited. A portion of the genome called gene is the basic unit of carrier. The genome is packed in chromosomes within the nucleus of a cell. The cells play a major role in metabolic activities of both prokaryotes and eukaryotes. The genetic data in chromosomes occur as Deoxyribo Nucleic Acid (DNA). DNA is a double stranded helical structure comprising of four nucleotide bases namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T).

Gene expression is the driving factor behind cellular activities. It includes the DNA splicing to form mRNA which is called as transcription and the synthesis of proteins from the RNA known as translation. Some of the generated proteins may enter the nucleus and interact with specific portions of DNA. These proteins are called Transcription Factors (TF) and the DNA interaction regions are called Transcription Factor Binding Sites (TFBS) or in short binding sites. TFs control which genes are turned on or off in the genome. This paper gives a detailed survey about the prediction of binding sites in DNA. The prediction can be based on the nucleotide sequence of DNA, amino acid sequence of TF or on the protein DNA interaction residues.

2. TFBS PREDICTION

The genome consists of protein coding regions and noncoding regions. The protein coding regions can be obtained by any ORF finding tools like ORFFinder [1]. The noncoding genome contains information for the peptide generation. The peptide is a 3D structure which acts as a backbone for the function and interaction of proteins. The functional aspects are sequence based whereas the interaction is both sequence and structure based.

The noncoding DNA do not get translated into proteins. They form the majority of DNA which in turn result in the formation of tRNA, rRNA, miRNA etc. The major types of noncoding DNA sequences are functional RNA, cis-regulatory elements, introns, pseudogenes, transposons and telomers. Functional RNA is used for the formation of ribosomal RNA and micro RNA. Cis-regulatory elements may act as trans factors or promoters of nearby genes and trans factors control the distant genes. These elements may control a group of genes. Promoters facilitate transcription and are included in the noncoding portions of gene.

Noncoding genomes contain the information about where the TFs bind to DNA. This is based on the binding affinity [2] of the DNA sequences. Experimental techniques like Chromatin Immuno Precipitation (ChIP), ChIP-seq [3] can be used to get information regarding the binding site predictions. Although the experimental techniques give accurate results, it is very costly and time consuming. So the computational techniques are used and then the experimental validation can be done.

2.1 Promoters

Most of the binding sites are located in the promoter region that is ahead of protein coding part of the genome. DNA sequences within promoters can act as cis regulatory elements for the same TF regulating nearby genes. Promoter DNA sequences can also be trans regulatory elements for distant genes which may act as activator or repressor [4]. The prediction can be made easier by identification of promoter sequences.

In eukaryotic organisms the features involved in promoters include CAAT box, TATA box, CpG Island and CAP box. These are the conserved eukaryotic promoter elements [5]. In plants the CAAT box may be replaced by AGGA box. CAAT box represents a consensus sequence that is at a proximal distance of -80 base pairs from the binding start site. The starting point of binding sites may

be taken as +1. TATA box is much more closer than the CAAT box which may also be formed by the consensus that is located at a distance of 25 base pairs from the binding start site. Another striking element of TATA box is that it may be surrounded by sequences rich in CG content. CpG Island refers to a sequence rich in C and G found in multiple copies. CAP site is the starting point of transcription denoted as +1 and it is also called as transcription initiation sequence. The presence of CAAT box, TATA box, CpG Island also indicate existence of binding sites in close proximity.

In prokaryotes there are some small variations in the promoters. There is a promoter element called operon that starts the transcription of multiple genes adjacent to it. Here a single transcribed mRNA may be translated to several proteins. There are two control modes for the working of DNA - positive control mode and negative control mode. In positive control mode the binding may take place and in negative control mode there may not be any binding. Search space for binding sites cites can be enhanced by the knowledge of control modes. The binding prediction attempt starts with the gene prediction [6].

2.2 Prokaryotic Binding Site Prediction

Prokaryotes are single cellular organisms where the cellular complications are less. Prediction of binding sites is based on the location of introns and exons. Introns are noncoding sequences and exons are protein coding regions. Prokaryotes have no intervening introns and exons [7]. This makes the binding prediction much easier. Initially the noise associated with the genomic data is removed. Next is the feature extraction using probability based Hidden Markov Model [8].

2.3 Eukaryotic Binding Site Prediction

Eukaryotes have interleaving introns and exons. This makes the prediction much more complicated. The binding sites may be some 6-60 base pairs in length which depends upon the location of promoters. The enzyme RNA Polymerase may interact with DNA based on the enhancer TF [9] or repressor. The Transcription Start Site may be located some 10 base pairs upstream of the gene. Transcription regulatory network [10] can be created based on the binding sites which give details of the existing binding sites and easier prediction of unknown binding sites.

A variety of tools have been developed for the binding site prediction. TFSEARCH [11], MatInd, MatInspector, MATRIX SEARCH, SIGNAL SCAN, MATCH [12] are some of the prominent tools. The input can be given in FASTA format or plain text form. Using TFSEARCH one can model the Hunchback TF (M00022) for the Hb TFBS. But the details for TF Nanog is not available using TFSEARCH which result in invalid prediction. MATCH is a commonly used tool for searching the TFBS. It is primarily based on the similarity score which include the matric similarity score and core similarity score. A threshold cut-off can be given to the scores to evaluate the content of features that are active. The cut-off is used to minimize False Negative and False Positive rates that result in accurate prediction of binding sites.

3. MODELS

The binding site prediction is usually based on a set of genes that are co-regulated which form the cis-regulatory modules. Binding sites are usually based on DNA or RNA motif identification [13].

3.1 Features Used for Binding Site Prediction

Binding affinity can be used as a strategy for prediction. It is based on the sequence features which include the physicochemical properties [14] and evolutionary information. Both depend on the nucleotide sequence data. A total of 38 physico-chemical properties can be integrated with the sequence features. The feature space can be extremely large. The dataset can be obtained from ACTIVITY database [15]. The evolutionary information can be represented as Position Specific Scoring Matrix.

3.2 Phylogenetic Footprinting

There is a close similarity between the evolutionary aspects in DNA and RNA sequences. This is the driving factor behind phylogenetic foot printing. It is a technique of finding the binding sites with the evolutionary information. The noncoding regions of DNA are used for finding the binding sites using the homologous and orthologous information. Orthology refers to the retain of same functional aspects in the course of evolution. There is sequence conservation between the species which acts as a biological pipeline. TargetOrtho is a common tool for finding the phylogenetic details. Footprint [16] is a computer program for finding the phylogenetic footprint details. From the knowledge of binding sites of a particular organism, prediction of binding sites of all evolutionarily related organisms can be made.

3.3 Stochastic Models

Pi-calculus mechanism can be used to model the sequence information exchange between different species. The transition from one state to another can be used to model the sequence information. The transitions include change from chromosome to selected portions of DNA, proteins and DNA-binding residues. It can also be used to model the binding affinity of protein DNA interactions. One of the most commonly used stochastic models is the Position Specific Scoring Matrix, in short PSSM.

Structural and sequential features can be used to model the sample sequence for binding site prediction. Sequential features include the DNA sequence nucleotides, consensus prepared from nucleotides, the helical features of DNA, high CG content, CpG Island and PSSM hit values. The structural features include bending of DNA, TF binding on major groove of DNA, solvent accessibility and surface area available for binding. More the features collected, more accurate will be the resulting prediction and the features can be represented using Bayesian networks [17].

4. DNA PROTEIN INTERACTION RESIDUE

Protein DNA interaction forms a major role in the genetic engineering. Binding residues is a 3D structure with the details that can be obtained from the Protein Data Bank (PDB). The amino acid nucleotide sequence pairs may determine the functional characteristics. Sequence features can be extracted from the amino acid datasets like PROSITE, Pfam database. The interaction may result in the formation of motifs by which one can apply the motif matching as well as motif discovery techniques. The feature extraction can be based on the binding properties like side chain pKa value, hydrophobicity index and thermodynamic properties with the structural and sequential parameters. k-mers can be selected with the association rule mining approach for the binding evaluation. Agarwal et. al suggested an Apriori algorithm for the evaluation based on the binding residues. Binding is based on the basic principle that two atoms are said to be interacting if the distance between them

is less than 3.5 A. k-mers of binding residues [18] can be selected for the comparison with varying values for k. The main limitation of interaction residue is the lack of all the linking counterparts of protein nucleotide interaction.

5. PSSM BASED METHODS

The most common approach for binding site prediction uses the Position Specific Scoring Matrix or in short PSSM technique. This is one of the most common strategies to represent the binding sites accurately. PSSM is a matrix representation of data which is based on the independence between nucleotides. This may add to the disadvantage for sequence extraction. The matrix preparation is based on the frequency distribution of nucleotide bases at each position in the sequence corresponding to the DNA or amino acid [19]. For nucleotide bases 4 rows have to be considered corresponding to the bases and for proteins, 20 rows have to be considered corresponding to 20 amino acids for each sequence data. For evaluation, the log based score can be prepared considering the background distribution also. The main limitation of PSSM is that it is unable to handle the physico-chemical properties of the sequence distribution.

6. ALIGNMENT BASED METHODS

The sequence can be aligned locally or globally. PSI-BLAST or BLAST can be used as a tool for alignment. To form the local approximation, the techniques used are Greedy optimization, Expectation Maximization and Gibbs Sampling. Supervised learning methods use a set of known binding sites from the experimental data. This may be taken as the positive set for training. The negative set can be taken as any of the non-binding sites selected randomly. The learning based methods have the disadvantage of high dimensional and noisy data.

Based on the alignment, one can prepare a consensus and search for the similarity in the subsequences based on a sliding window. This considers the variability data rather than similarity for regulatory regions. Majority of the machine learning approaches take a set of TFBS sequences as input and generate a matrix representation for this. This act as a binding model which is used to scan the DNA sequences to find new binding sites. The model can also be used to find unknown putative binding sites in a sequence for a given TF.

7. DNA BINDING DYNAMICS FOR BINDING CHARACTERISTICS

The binding dynamics refer to the changes occurring to DNA structure as a result of the protein DNA interaction. Studies have been made using the Rap1-Myc protein proposed by Colin et. al [20]. The binding turnover was detected using TATA binding feature. Studies revealed a strong correlation between binding sites and histone acetyltransferase enrichment. Binding stabilization can be made with the nucleosome instability. The histone H3 can be used as the active promoter. The binding sites can be long or short DNA sequences and the binding behavior is encoded within the DNA code. Binding dynamics was influenced by the histone turnover and residence time of transcription factors. They are preferentially A or T rich with the variability being enhanced by the difference in affinity to the RNA Polymerase. Binding affinity is based on the availability of DNA accessible from surface. Binding dynamics determine the functional consequences with a predictable switch between inactive and active TF states. The signal measurement is based on background and current values of bases.

8. DNA PROTEIN BINDING SITES

A lot of approaches have been developed for DNA sequencing. Proteins as well as enzymes may interact with the DNA subsequences to form the binding residue. Binding residues may differ in composition due to the variation in binding affinity for RNA Polymerase which result in the formation of different patterns of binding sites. Binding sites can be used to extract the information content and binding energy which can be estimated by the frequency distribution of dinucleotide or mononucleotides. Sequences rich in specific bases (eg. AT rich) can also be evaluated to give the average binding energy which is also referred as the relative entropy. Entropy can be used as a measure for determining the binding sites. All the probability estimation is based on the assumption that each nucleotide position has independent role in calculating the binding affinity.

9. ALGORITHMS FOR PREDICTING TFBS

Mainly there are four classes of algorithms for predicting the binding sites. They are enumerative algorithms, phylogenetic foot printing, iterative algorithms and content based algorithms. Enumerative algorithms use a background of base pair for searching motifs against the background. Phylogenetic foot printing use the homolog and ortholog data to find the binding motifs. Iterative algorithms use the Expectation Maximization to define the probability distribution of binding sites. Content based algorithms use the segmentation technique to divide a long sequence into subsequences for processing as regular expressions. To reduce the error rate, a set of twelve algorithms are used. This increases the accuracy by reducing error rate of predictions. The result of algorithms can be represented as feature vectors. Representative features include both extrinsic and intrinsic features. Extrinsic features are used for signaling pathways for a phenotype. Intrinsic features are used for identification of the functioning of cells which also include the working of TFs. The details about prediction can be taken from TRANSFAC [21] and JASPAR [22] database which provide information about TFs, associated genes and probable binding sites. Some of the common TFs are Oct4, Sox2, c-Myc, Klf4, Nanog, Esorb, Zfx, STAT3.

A commonly used webserver for the prediction of DNA binding residues is BindN. For any given sequence, alignment can be done locally or globally. This may reduce the sample size for processing. Local alignment is done using Smith Waterman algorithm that is implemented by taking small subsequences for comparison. Global alignment can be done by the Needleman Wunsch algorithm which takes the sequence as a whole for comparison. The alignment can be done based on nucleotide sequences or amino acid sequences.

The processing of sequences are done by accessing as codons. The start codon is the first codon of mRNA. The start codon commonly used is AUG, called as Methionine. The start codon is usually preceded by a non-translated region of 5' upstream of the sequence. UAG, commonly known as amber, UGA which is umber/opal and UAA also called as ochre are the commonly used stop codons. The corresponding stop codons in DNA are TGA, TAA and TAG. The start codon marks the starting point of translation. The expectation maximization can be used when there is some missing data for evaluation. BLAST is a major dynamic programming tool used for the sequence comparison. The webserver will automatically generate PSSM of a given input nucleotide or amino acid sequence against a reference data. The direction of transcription is defined by the downstream end. To search for motifs, find the upstream sequences

Table 1. Computational Binding Site Prediction Programs.

Approach	Organism	Technique	Program
PSSM	Prokaryotes and Eukaryotes	Motif matching	Content based algorithm
Phylogenetic footprinting	Prokaryotes and Eukaryotes	Sequence Conservation	Homolog modelling
Stochastic approaches	Eukaryotes	Sequence patterns	Clustering based on correlation
Feature based	Eukaryotes	Feature motif model	Enumeration algorithm
Probability based	Eukaryotes and Prokaryotes	Markov model	Iteration algorithm
Biological knowledge with physico chemical features	Eukaryotes	Filter based enumeration	Entropy estimation technique
Association rule mining	Eukaryotes	Branch and Bound	Apriori algorithm
Binding affinity	Eukaryotes	Sequence patterns	PRIMA algorithm
Protein DNA binding residue	Eukaryotes	k-mer evaluation	TF conservation

between bound and unbound regions by checking both positive and negative samples.

PRIMA algorithm is used for identifying the TFs with overrepresented binding sites in a set of promoters. The algorithm is integrated along with Expander software with input as a set of correlated genes, background set of genes and PSSM of known binding sites. Output is a set of TFs and their probability values. The first step in the algorithm is to generate a random input sequence of amino acids or nucleotides. The second step is to set a threshold score for the PSSM. The third step is to find the likelihood score. The fourth step is to find the hits of binding sites that pass the threshold value. For this, scan the target sequences and background model in search of hits. The fifth step is to find enrichment score which is done by checking whether the number of hits in target set is higher than the expected value. The distribution of hits is obtained by comparison with background value.

Table 1 shows a comparative outlook of the approaches used. PSSM is the easiest method of extracting binding site details. But it has the limitation of independence between nucleotides. Phylogenetic foot printing collects the evolutionary details, but lacks the structural parameters. This can be overruled by adding the feature details which result in probability based estimate of binding sites. The increased number of features can be limited by filters. Binding affinity and rule based mining can be applied in binding residues for the prediction of binding sites.

10. CONCLUSION

The inherent details of any organism are obtained from gene expression. The basic building blocks of living things are formed by proteins which are formed by the transcription and translation. Transcription factors make a major role in determining how the proteins are formed and metabolic activities are controlled. Transcription factor binding sites are determined by the DNA locations of where proteins interact with DNA sequences. Most of the binding sites are located in promoter regions upstream of protein coding portion. Prediction of binding sites is easier in prokaryotes than eukaryotes. Binding site prediction can be done based on feature extraction, phylogenetic foot printing and stochastic models. DNA protein interaction residues form a major role in the binding prediction by taking the features from known binding residues and data repository from amino acid databases. Binding affinity values can be used for differential calculation to find the log likelihood score. PSSM is the most commonly used technique to represent the binding interaction based on the sequence features whereas the binding affinity use structural features for prediction. To reduce the search space for prediction, alignment of sequences can be done.

11. REFERENCES

[1] Patel S, Panchal H, Anjaria K, *DNA Sequence Analysis by ORF FINDER & GENOMATIX tool: Bioinformatics Analysis of some Tree Species of Leguminosae Family*, Bioinformatics and Biomedicine Workshop (BIBMW), IEEE International Confference 2012; doi:10.1109/BIBMW.2012.6470265.

[2] Franco Zorrilla J M, Solano R, *High Throughput Analysis of Protein-DNA Binding Affinity*, Methods in Molecular Biology, NCBI PubMed 2014, 1062:697-709; doi:10.1007/978-1-62703-58-4_36.

[3] Timothy L Bailey and Philip Machanick, *Inferring Direct DNA Binding from ChIP-Seq*,Nucleic Acids Research 2012, doi: 10.1093/nar/gks433.

[4] Lodish H, Berk A, Zipursky S L, *Eukaryotic Transcription Activators and Repressors*, Book:Molecular Cell Biology 2000.

[5] Anders Gorm Pedersen, Pierre Baldi, Yves Chauvin, Soren Brunak, *The Biology of Eukaryotic Promoter Prediction - A Review*, Computers and Chemistry 1991, 191-207, PERGAMON.

[6] Israa M. Al-Twaiki, Hassan Mathkur, Ameur Touir, M S Aksoy and Alaaeldin Hafez, *Computational Approaches for Gene Prediction : A Comparative Survey*, Sci. Int. (Lahore), 23(2), 83-90, 2001, ISSN:1013-5316.

[7] M Belfort, M E Reaban, T Coetzee and J Z Dalgaard, *Prokaryotic Intons and Inteins : A Panoply of Form and Function*,NCBI PMC177114 Journal of Bacteriology 1995, 3897-3903.

[8] Anthony Mathelier, Wyeth W Wasserman, *The Next Generation of Transcription Factor Binding Site Prediction*, PLOS Computational Biology 2013, doi:10.1371/journal.pcbi.1003214.

[9] Chen C Y, Morris Q, Mitchell J A, *Enhancer Identification in Mouse Embryonic Stem Cells using Integrative Modeling of Chromatin and Genomic Features*, BMC Genomics 2012, doi:10.1186/1471-2164-13-152.

[10] Babu M M, Lucscombe N M, Aravind L, Gerstein M, Teichmann S A, *Structure and Evolution of Transcriptional regulatory Network*, PubMed 2004, 283-291.

[11] Tatsuhiko Tsunoda and Toshihisa Takgi, *Estimating Transcription Factor Bindability on DNA*, BIOINFORMATICS 1999, 622-630.

[12] A E Kel, E Go Bling, I Reuter, E Cheremushkin, O V Kel-Margoulis and E Wingender, *MATCH, A Tool for Searching Transcription Factor Binding Sites in DNA Sequences*, Nucleic Acids Research 2003, Vol.31, No. 3.

[13] Modan K Das and Ho Kwok Dai, *A survey of DNA Motif Finding Algorithms*, BMC Bioinformatics 2007.

[14] Mark Maienschein Cline, Aaron R Dinner, William S Hlavacek, Fangping Mu *Improved Predictions of Transcription Factor Binding Sites using Physicochemical Features of DNA*, Nucleic Acids Research 2012.

[15] Julia V Ponomarenko, Dagmara P Furman, Antoly S Frolov, Nikolay L Podkolodny, Galina V Orlova, Mikhail P Ponomarenko, Nikolay A Kolchanov, Akinoi Sarai, *ACTIVITY : A Database on DNA/RNA Sites Activity Adapted to Apply Sequence Activity Relationships from one System to Another*, IEEE 2006.

[16] Alvaro Sebastian and Bruno Contreras-Moreira, *footprintDB: a database of transcription factors with annotated cis elements and binding interfaces*, Bioinformatics Advance Access, November, 2013.

[17] Swetlana Nikolajewa, Rainer Pudimat, Michael Hiller, Matthias Platzer, Rolf Backofen, *BioBayesNet : A Web Server for Feature Extraction and Bayesian Network Modeling of Biological Sequence Data*, Nucleic Acids Research 2007, PMC 1933181, doi:10.1093/nar/qkm292.

[18] Liangjiang Wang, Mary Qu Yang and Jack Y Yang, *Prediction of DNA-binding residues from protein sequence information using random forests*, BMC Genomics 2009.

[19] Shandar Ahmed and Akinoi Sarai, *PSSM based Prediction of DNA Binding Sites in Proteins*, BMC Bioinformatics 2005.

[20] Colin R Lickwar, Florian Meuller, Sean E Hanlon, James G McNally, Jason D Lieb, *Genome-wide proteinDNA binding dynamics suggest a molecular clutch for transcription factor function*, Nature 484, 2012, 251-255, doi:10.1038/nature10985.

[21] E Wingender, P Dietze, H Karas, R Knuppel, *TRANSFAC : A Database on Transcription Factors and their DNA Binding Sites*, Nucleic Acids Research 1996, Vol. 24, No: 1, 238-241.

[22] Dominique Vlieghe, Albin Sandelin, Rieter J De Bleser, Kris Vleminickx, Wyeth W Wasserman, Frans van Roy, Boris Lenhard, *A New Generation of JASPAR, the Open Access Repository for Transcription Factor Binding Site Profiles*, Nucleic Acids Research 2006 Vol 34, doi:10.1093/nar/gkj115.