3

In Silico Identification of Regulatory Elements in Promoters

Vikrant Nain[1], Shakti Sahi[1] and Polumetla Ananda Kumar[2]
[1]*Gautam Buddha University, Greater Noida*
[2]*National Research Centre on Plant Biotechnology, New Delhi*
India

1. Introduction

In multi-cellular organisms development from zygote to adult and adaptation to different environmental stresses occur as cells acquire specialized roles by synthesizing proteins necessary for each task. In eukaryotes the most commonly used mechanism for maintaining cellular protein environment is transcriptional regulation of gene expression, by recruiting required transcription factors at promoter regions. Owing to the importance of transcriptional regulation, one of the main goals in the post-genomic era is to predict gene expression regulation on the basis of presence of transcription factor (TF) binding sites in the promoter regions. Genome wide knowledge of TF binding sites would be useful to build transcriptional regulatory networks model that result in cell specific differentiation. In eukaryotic genomes only a fraction (< 5%) of total genome codes for functional proteins or RNA, while remaining DNA sequences consist of non-coding regulatory sequences, other regions and sequences still with unknown functions.

Since the discovery of trans-acting factors in gene regulation by Jacob and Monads in lac operon of *E. coli*, scientists had an interest in finding new transcription factors, their specific recognition and binding sequences. In DNAse footprinting (or DNase protection assay); transcription factor bound regions are protected from DNAse digestion, creating a "footprint" in a sequencing gel. This methodology has resulted in identification of hundreds of regulatory sequences. However, limitation of this methodology is that it requires the TF and promoter sequence (100-300 bp) in purified form. Our knowledge of known transcription factors is limited and recognition and binding sites are scattered over the complete genome. Therefore, in spite of high degree of accuracy in prediction of TF binding site, this methodology is not suitable for genome wide or across the genomes scanning.

Detection of TF binding sites through phylogenetic footprinting is gradually becoming popular. It is based on the fact that random mutations are not easily accepted in functional sequences, while they continuously keep on tinkering non functional sequences. Many comparative genomics studies have revealed that during course of evolution regulatory elements remain conserved while the non-coding DNA sequences keep on mutating. With an ever increasing number of complete genome sequence from multiple organisms and mRNA profiling through microarray and deep sequencing technologies, wealth of gene expression data is being generated. This data can be used for identification of regulatory elements through intra and inter species comparative genomics. However, the identification of TF binding sites in promoters still remains one of the major challenges in bioinformatics due to following reasons:

1. Very short (5-15 nt) size of regulatory motifs that also differ in their number of occurrence and position on DNA strands with respect to transcription start site. This wide distribution of short TF binding sites makes their identification with commonly used sequence alignment programmes challenging.
2. A very high degree of conservation between two closely related species generally shows no clear signature of highly conserved motifs.
3. Absence of significant similarities between highly diverse species hinders the alignment of functional sequences.
4. Sometimes, functional conservation of gene expression is not sufficient to assure the evolutionary preservation of corresponding cis-regulatory elements (Pennacchio and Rubin, 2001).
5. Transcription factors binding sites are often degenerate.

In order to overcome these challenges, in the last few years novel approaches have been developed that integrate comparative, structural, and functional genomics with the computational algorithms. Such interdisciplinary efforts have increased the sensitivity of computational programs to find composite regulatory elements.

Here, we review different computational approaches for identification of regulatory elements in promoter region with seed specific legumin gene promoter analysis as an example. Based on the type of DNA sequence information the motif finding algorithms are classified into three major classes: (1) methods that use promoter sequences from co regulated genes from a single genome, (2) methods that use orthologous promoter sequences of a single gene from multiple species, also known as phylogenetic footprinting and (3) methods that use promoter sequences of co regulated genes as well as phylogenetic footprinting (Das and Dai, 2007).

2. Representation of DNA motifs

In order to discover motifs of unknown transcription factors, models to represent motifs are essential (Stormo, 2000).There are three models which are generally used to describe a motif and its binding sites:

1. string representation (Buhler and Tompa, 2002)
2. matrix representation (Bailey and Elkan, 1994) and
3. representation with nucleotide dependency (Chin and Leung, 2008)

2.1 String representation

String representation is the basic representation using string of symbols or nucleotides A, C, G and T of length-l to describe a motif. Wildcard symbols are introduced into the string to represent choice from a subset of symbols at a particular position. The International Union of Pure and Applied Chemistry (IUPAC) nucleic acid codes (Thakurta and Stormo, 2007) are used to represent the information about degeneracy for example: W = A or T ('Weak' base pairing); S= C or G ('Strong' base pairing); R= A or G (Purine); Y= C or T (Pyrimidine); K= G or T (Keto group on base); M= A or C (Amino group on base); B= C, G, or T; D= A, G, or T ; H= A, C, or T ; V= A, C, or G; N= A, C, G, or T.

2.2 Matrix representation

In matrix representation, motifs of length l are represented by position weight matrices (PWMs) or position specific scoring matrices (PSSMs) of size 4x l. This gives the occurrence probabilities of each of the four nucleotides at a position j. The score of any specific sequence is the sum of the position scores from the weight matrix corresponding to that sequence. Using this representation an entire genome can be scanned by a matrix and the score at every position obtained (Stormo, 2000). Any sequence with score that is higher than the predefined cut-off is a potential new binding site. A consensus sequence is deduced from a multiple alignment of input sequences and then converted into a position weight matrix.

A PWM score is the sum of position-specific scores for each symbol in the substring. The matrix has one row for each symbol of the alphabet, and one column for each position in the pattern. The score assigned by a PWM to a substring $S = \left(S_j\right)_{j=1}^{N}$, is defined as $\sum_{j=1}^{N} m_{sj,j}$ where j represents position in the substring, s_j is the symbol at position j in the substring, and $m_{\alpha,j}$ is the score in row α, column j of the matrix.

Although matrix representation appears superior, the solution space for PWMs and PSSMs, which consists of 4l real numbers is infinite in size, and there are many local optimal matrices, thus, algorithms generally either produce a suboptimal motif matrix or take too long to run when the motif is longer than 10 bp (Francis and Henry, 2008).

2.3 Representation with nucleotide dependency

The interdependence between neighboring nucleotides with similar number of parameters as string and matrix representations is described by Scored Position Specific Pattern (SPSP). A set of length-l binding site patterns can be described by a SPSP representation P, which contains c (c ≤ l) sets of patterns Pi, 1 ≤ i ≤ c, where each set of patterns Pi contains length-li patterns $P_{i,j}$ of symbols A, C, G, and T and $\sum_i l_i = l$. Each length- l_i pattern $P_{i,j}$ is associated with a score $s_{i,j\ \mathrm{that}}$ represents the "closeness" of a pattern to be a binding site. The lower the score, the pattern is more likely a binding site (Henry and Fracis, 2006).

3. Methods of finding TF binding sites in a DNA sequence

3.1 Searching known motifs

Development of databases of complete information on experimentally validated TF binding site is indispensable for promoter sequence analysis. Information about TF binding sites remain scattered in literature. In the last one and half decade phenomenal increase in computational power, cheaper electronic storage with faster communication technologies, have resulted in development of a range of web accessible databases having experimentally validated TF binding sites. These TF binding site databases are not only highly useful for identification of putative TF binding sites in new promoter sequences (Table1), but also are valuable for providing positive dataset required for improvement and validation of new TF binding site prediction algorithms.

3.1.1 TRANSFAC

TRANSFAC is the largest repository of transcription factors binding sites. TRANSFAC (TRANSFAC 7.0, 2005) web accessible database consists of 6,133 factors with 7,915 sites, while professional version (TRANSFAC 2008.3) consists of 11,683 factors with 30,227 sites. TRANSFAC database is composed of six tables SITE, GENE, FACTOR, CELL, CLASS and

MATRIX. GENE table gives a short explanation of the gene where a site (or group of sites) belongs to; FACTOR table describes the proteins binding to these sites. CELL gives brief information about the cellular source of proteins that have been shown to interact with the sites. CLASS contains some background information about the transcription factor classes, while the MATRIX table gives nucleotide distribution matrices for the binding sites of transcription factors. This database is most frequently used as reference for TFB sites as well as for development of new algorithms. However, new users find it difficult to access the database because it requires search terms to be entered manually. There is no criterion to select the organism, desired gene or TF from a list, so web interface is not user friendly. Other web tools such as TF search and Signal Scan overcome this limitation to certain extent.

3.1.2 Signal Scan

Signal Scan finds and lists homologies of published TF binding site signal sequences in the input DNA sequence by using TRANSFAC, TFD and IMD databases. It also allows to select from different classes viz mammal, bird, amphibian, insect, plant, other eukaryotes, prokaryote, virus (TRANSFAC only), insect and yeast (TFD only).

3.1.3 TRRD

The transcription regulatory region database (TRRD) is a database of transcription regulatory regions of the eukaryotic genome. The TRRD database contains three interconnected tables: TRRDGENES (description of the genes as a whole), TRRDSITES (description of the sites), and TRRDBIB (references). The current version, TRRD 3.5, comprises of the description of 427 genes, 607 regulatory units (promoters, enhancers, and silencers), and 2147 transcription factor binding sites. The TRRDGENES database compiles the data on human (185 entries), mouse (126), rat (69), chicken (29), and other genes.

Developmental/Environmental stimulus	Transcription factor binding site	Position	Sequence
Core promoter	TATA Box	-33	tcccTATAaataa
	Cat Box	-49	gCCAAc
Stress responsive	G Box	-66	tgACGgtgt
	ABRE	-76	acaccttctttgACTGtccatccttc
	ABI4	-245	CACCg
Pathogen defense	W Box	-72	cttctTTGAcgtgtcca
	TCA		gAGAAgagaa
Light Response	I box	-302	gATATga
Wound specific	WUN	-348	tAATTacac
	TCA	-646	gAGAAgagaa
Seed Specific	Legumin	-118	tccatacCCATgcaagctgaagaatgtc
	Opaque-2	-348	TAATtacacatatttta
	Prolamine box	-385	TTaaaTGTAAAAgtAa
	AAGAA-motif	-294	agaAAGAa

Table 1. *In silico* analysis of pigeonpea legumin gene promoter for identification of regulatory elements. Database search reveals that it consist of regulatory elements that can direct its activation under different envirnmental conditions and developmental stages.

3.1.4 PlantCARE

PlantCARE is database of plant specific cis-Acting regulatory elements in the promoter regions (Lescot et al., 2002). It generates a sequence analysis output on a dynamic webpage, on which TF binding sites are highlighted in the input sequence. The database can be queried on names of transcription factor (TF) sites, motif sequence, function, species, cell type, gene, TF and literature references. Information regarding TF site, organism, motif position, strand, core similarity, matrix similarity, motif sequence and function are listed whereas the potential sites are mapped on the query sequence.

3.1.5 PLACE

PLACE is another database of plant *cis*-acting regulatory elements extracted from published reports (Higo et al., 1999). It also includes variations in the motifs in different genes or plant species. PLACE also includes non-plant cis-elements data that may have homologues with plant. PLACE database also provides brief description of each motif and links to publications.

3.1.6 RegulonDB

RegulonDB is a comprehensive database of gene regulation and interaction in *E. coli.* It consists of data on almost every aspect of gene regulation such as terminators, promoters, TF binding sites, active and inactive transcription factor conformations, matrices alignments, transcription units, operons, regulatory network interactions, ribosome binding sites (rbs), growth conditions, gene product and small RNAs.

3.1.7 ABS

ABS is a database of known TF binding sites identified in promoters of orthologous vertebrate genes. It has 650 annotated and experimental validated binding sites from 68 transcription factors and 100 orthologous target genes in human, mouse, rat and chicken genome sequences. Although it's a simple and easy-to-use web interface for data retrieval but it does not facilitate either analysis of new promoter sequence or mapping user defined motif in the promoter.

3.1.8 MatInspector

MatInspector identifies cis-acting regulatory elements in nucleotide sequences using library of weight matrices (Cartharius et al., 2005). It is based on novel matrix family concept, optimized thresholds, and comparative analysis that overcome the major limitation of large number of redundant binding sites predicted by other programs. Thus it increases the sensitivity of reducing false positive predictions. MatInspector also allows integration of output with other sequence analysis programs e.g. DiAlignTF, FrameWorker, SequenceShaper, for an in-depth promoter analysis and designing regulatory sequences. MatInspector library contains 634 matrices representing one of the largest libraries available for public searches.

3.1.9 JASPAR

JASPAR is the another open access database that compete with the commercial TF binding site databases such as TRANSFAC (Portales-Casamar et al., 2009). The latest release has a collection of 457 non-redundant, curated profiles. It is a collection of smaller databases, viz JASPAR CORE, JASPAR FAM, JASPAR PHYLOFACTS, JASPAR POLII and others, among which JASPAR CORE is most commonly used. The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally determined transcription factor binding sites for multicellular eukaryotes (Portales-Casamar et al., 2009). The JASPAR database can also be accessed remotely through external application programming interface (API).

3.1.10 Cister: cis-element cluster finder

Cister is based on the technique of posterior decoding, with Hidden Markov model and predicts regulatory regions in DNA sequences by searching for clusters of cis-elements (Frith et al., 2001). The Cister input page consists of 16 common TF sites to define a cluster and additional user defined PWM or TRANSFAC entries can also be entered. For web based analysis maximum input sequence length is 100 kb, however, the program is downloadable for standalone applications and analysis of longer sequences.

3.1.11 MAPPER

MAPPER stands for Multi-genome Analysis of Positions and Patterns of Elements of Regulation It is a platform for the computational identification of TF binding sites in multiple genomes (Marinescu et al., 2005). The MAPPER consists of three modules, the MAPPER database, the Search Engine, and rSNPs and combines TRANSFAC and JASPAR data. However, MAPPER database is limited to TFBSs found only in the promoter of genes from the human, mouse and *D.melanogaster* genomes.

3.1.12 Stubb

Like Cister, Stubb also uses hidden Markov models (HMM) to obtain a statistically significant score for modules (Sinha et al., 2006). STUBB is more suitable for finding modules over genomic scales with small set of transcription factors whose binding sites are known. Stubb differs from MAPPER in that the application of latter is limited to binding sites of a single given motif in an input sequence.

3.1.13 Clover

Clover is another program for identifying functional sites in DNA sequences. It take a set of DNA sequences that share a common function, compares them to a library of sequence motifs (e.g. transcription factor binding patterns), and identifies which, if any, of the motifs are statistically overrepresented in the sequence set (Frith et al., 2004). It requires two input files one for sequences in fasta format and another for sequence motif. Clover provides JASPAR core collection of TF binding sites that can be converted to clover format. Clover is also available as standalone application for windows, Linux as well as Mac operating systems.

3.1.14 RegSite

Regsite consists of plant specific largest repository of transcription factor binding sites. Current RegSite release contains 1816 entries. It is used by transcription start site prediction programs (Sinha et al., 2006).

3.1.15 JPREdictor

JPREdictor is a JAVA based cis-regulatory TF binding site prediction program (Fiedler and Rehmsmeier, 2006). The JPREdictor can use different types of motifs: Sequence Motifs, Regular Expression Motifs, PSPMs as well as PSSMs and the complex motif type (MultiMotifs). This tool can be used for the prediction of cis-regulatory elements on a genome-wide scale.

3.2 Motif finding programs

3.2.1 Phylogenetic footprinting

Comparative DNA sequence analysis shows local difference in mutation rates and reveals a functional site by virtue of its conservation in a background of non-functional sequences. In the phylogenetic equivalent, regulatory elements are protected from random drift across evolutionary time by selection. Orthologous noncoding DNA sequences from multiple species provide a strong base for identification of regulatory elements by Phylogenetic footprinting (Fig. 1) (Rombauts et al., 2003).

The major advantage of phylogenetic footprinting over the single genome is that multigene approach requires data of co regulated genes. While phylogenetic footprinting can identifying regulatory elements present in single gene, that remain conserved during the course of divergence of two species under investigation. With steep increase in available complete genome sequences, across species comparisons for a wide variety of organisms has become possible (Blanchette and Tompa, 2002; Das and Dai, 2007). A multiple sequence alignment algorithm suited for phylogenetic footprinting should be able to indentify small (5-15 bp) sequence in a background of highly diverse sequences.

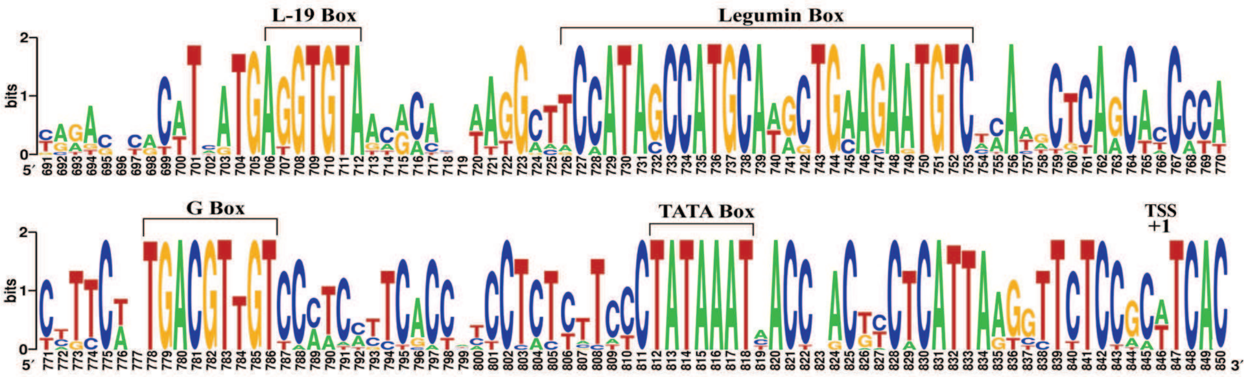


Fig. 1. Identification of new regulatory elements (L-19) in legumin gene promoters by phylogenetic footprinting.

3.2.1.1 Clustal W, LAGAN, AVID

In phylogenetic footprinting primary aim is to construct global multiple alignment of the orthologous promoter sequences and then identify a region conserved across orthologous sequences. Alignment algorithms, such as ClustalW (Thompson et al., 1994), LAGAN (Brudno et al., 2003), AVID (Bray et al., 2003) and Bayes-Block Aligner (Zhu edt al., 1998), have proven useful for phylogenetic footprinting, but the short length of the conserved motif compared to the length of the non-conserved background sequence; and their variable position in a promoter hampers the alignment of conserved motifs. Moreover multiple sequence alignment does not reveal meaningful biological information if the species used for comparison are too closely related. If the species are too distantly related, it is difficult to find an accurate alignment. It requires computational tools that bypass the requirement of sequence alignment completely and have the capabilities to identify short and scattered conserved regions.

3.2.1.2 MEME, Consensus, Gibbs sampler, AlignAce

In cases where multiple alignment algorithms fails, motif finding algorithms such as MEME, Consensus and Gibbs sampler have been used (Fig. 2). The feasibility of using comparative DNA sequence analysis to identify functional sequences in the genome of *S. cerevisiae*, with the goal of identifying regulatory sequences and sequences specifying nonprotein coding RNAs was investigated (Cliften et al., 2001). It was found that most of the DNA sequences of the closely related *Saccharomyces* species aligned to *S.cerevisiae* sequences and known promoter regions were conserved in the alignments. Pattern search algorithms like CONSENSUS (Hertz et al., 1990), Gibbs sampling (Lawrence et al., 1993) and AlignAce (Roth et al., 1998) were useful for identifying known regulatory sequence elements in the promoters, where they are conserved through the most diverged *Saccharomyces* species. Gibbs sampler was used for motif finding using phylogenetic footprinting in proteobacterial genomes (McCue et al., 2001). These programs employ two approaches for motif finding. One approach is to employ a training set of transcription factor binding sites and a scoring scheme to evaluate predictions. The scoring scheme is often based on information theory and the training set is used to empirically determine a score threshold for reporting of the predicted transcription factor binding sites. The second method relies on a rigorous statistical analysis of the predictions, based upon modeled assumptions. The statistical significance of a sequence match to a motif can be accessed through the determination of p-value. P-value is the probability of observing a match with a score as good or better in a randomly generated search space of identical size and nucleotide composition. The smaller the p-value, the lesser the probability that the match is due to chance alone. Since the motif finding algorithms assume the input sequences to be independent, therefore, they are limited by the fact that the data sets containing a mixture of some closely related species will have an unduly high weight in the results of motifs reported.

Multiple genome sequences were compared that are as optimally diverged as possible in *Saccharomyces* genomes. Phylogenetic footprints were searched among the genome sequences of six *Saccharomyces* species using the sequence alignment tool CLUSTAL W and many statistically significant conserved sequence motifs (Cliften et al., 2003) were found.

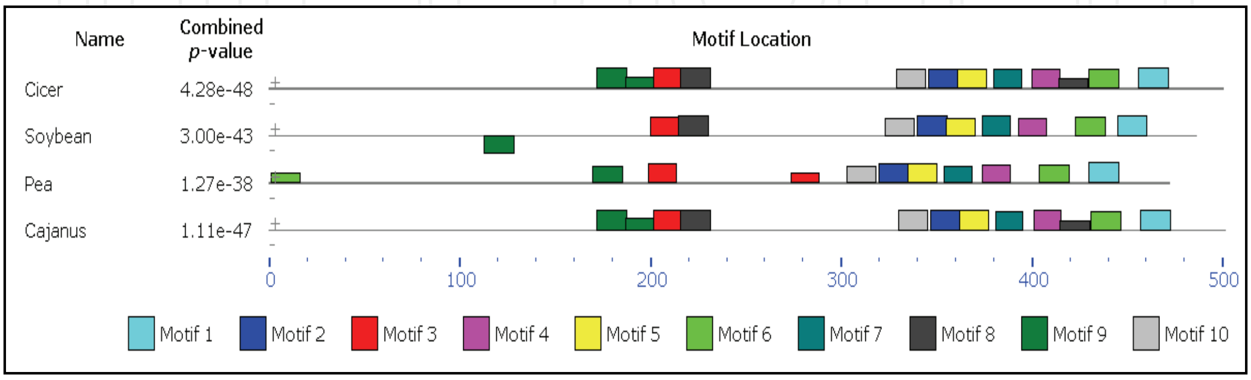


Fig. 2. Combined Block diagram of an MEME output highlighting conserved motifs in promoter regions of legumin seed storage protein genes of four different species.

3.2.1.3 Footprinter

This promising novel algorithm was developed to overcome the limitations imposed by motif finding algorithms. This algorithm identifies the most conserved motifs among the input sequences as measured by a parsimony score on the underlying phylogenetic tree (Blanchette and Tompa, 2002). It uses dynamic programming to find most parsimonious k-mer from each of the input sequences where k is the motif length. In general, the algorithm selects motifs that are characterized by a minimal number of mismatches and are conserved over long evolutionary distances. Furthermore, the motifs should not have undergone independent losses in multiple branches. In other words, the motif should be present in the sequences of subsequent taxa along a branch. The algorithm, based on dynamic programming, proceeds from the leaves of the phylogenetic tree to its root and seeks for motifs of a user-defined length with a minimum number of mismatches. Moreover, the algorithm allows a higher number of mismatches for those sequences that span a greater evolutionary distance. Motifs that are lost along a branch of the tree are assigned an additional cost because it is assumed that multiple independent losses are unlikely in evolution. To compensate for spurious hits, statistical significance is calculated based on a random set of sequences in which no motifs occur.

3.2.1.4 CONREAL

CONREAL (Conserved Regulatory Elements Anchored Alignment Algorithm) is another motif finding algorithm based on phylogenetic footprinting (Berezikov et al., 2005). This algorithm uses potential motifs as represented by positional weight matrices (81 vertebrate matrices form JASPAR database and 546 matrices from TRANSFAC database) to establish anchors between orthologous sequences and to guide promoter sequence alignment. Comparison of the performance of CONREAL with the global alignment programs LAGAN and AVID using a reference data set, shows that CONREAL performs equally well for closely related species like rodents and human, and has a clear added value for aligning promoter elements of more divergent species like human and fish, as it identifies conserved transcription-factor binding sites that are not found by other methods.

3.2.1.5 PHYLONET

The PHYLONET computational approach identifies conserved regulatory motifs directly from whole genome sequences of related species without reliance on additional information was developed by (Wang and Stormo, 2005). The major steps involved are: i) construction of phylogenetic profiles for each promoter , ii) searching through the entire profile space of all the promoters in the genome to identify conserved motifs and the promoters that contain them using algorithm like BLAST, iii) determination of statistical significance of motifs (Karlin and Altschul, 1990). By comparing promoters using phylogenetic profiles (multiple sequence alignments of orthologous promoters) rather than individual sequences, together with the application of modified Karlin– Altschul statistics, they readily distinguished biologically relevant motifs from background noise. When applied to 3524 *Saccharomyces cerevisiae* promoters with *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, and *Saccharomyces bayanus* sequences as references PHYLONET identified 296 statistically significant motifs with a sensitivity of >90% for known transcription factor binding sites. The specificity of the predictions appears very high because most predicted gene clusters have additional supporting evidence, such as enrichment for a specific function, in vivo binding by a known TF, or similar expression patterns.

However, the prediction of additional transcription factor binding sites by comparison of a motif to the promoter regions of an entire genome has its own problems due to the large database size and the relatively small width of a typical transcription factor binding site. There is an increased chance of identification of many sites that match the motif and the variability among the transcription factor binding sites permits differences in the level of regulation, due to the altered intrinsic affinities for the transcription factor (Carmack et al., 2007).

3.2.1.6 Phyloscan

PhyloScan combines evidence from matching sites found in orthologous data from several related species with evidence from multiple sites within an intergenic region.The orthologous sequence data may be multiply aligned, unaligned, or a combination of aligned and unaligned. In aligned data, PhyloScan statistically accounts for the phylogenetic dependence of the species contributing data to the alignment and, in unaligned data; the evidence for sites is combined assuming phylogenetic independence of the species. The statistical significance of the gene predictions is calculated directly, without employing training sets (Carmack et al., 2007). The application of the algorithm to real sequence data from seven Enterobacteriales species identifies novel Crp and PurR transcription factor binding sites, thus providing several new potential sites for these transcription factors.

3.3 Software suites for motif discovery

3.3.1 BEST

BEST is a suite of motif-finding programs that include four motif-finding programs: AlignACE (Roth et al., 1998), BioProspector(Liu et al., 2001), Consensus(Hertz and Stormo, 1999), MEME (Bailey et al., 2006) and the optimization program BioOptimizer (Jensen and Liu, 2004). BEST was compiled on Linux, and thus it can only be run on Linux machines (Che et al., 2005).

3.3.2 Seqmotifs

Seqmotifs is a suite of web based programs to find regulatory motifs in coregulated genes of both prokaryotes and eukaryotes. In this suite BioProspector (Liu et al., 2001) is used for finding regulatory motifs in prokaryote or lower eukaryote sequences while CompareProspector(Liu et al., 2002) is used for higher eukaryotes. Another program Mdscan (Liu et al., 2002) is used for finding protein-DNA interaction sites from ChIP-on-chip targets. These programs analyze a group of sequences of coregulated genes so they may share common regulatory motifs and output a list of putative motifs as position-specific probability matrices, the individual sites used to construct the motifs, and the location of each site on the input sequences. CompareProspector has been used for identification of transcription factors Mef2, My, Srf, and Sp1 motifs from a human-muscle-specific co regulated genes. Additionally in a *C. elegans–C briggsae* comparison, CompareProspector found the PHA-4 motif and the UNC-86 motif.(Liu et al., 2004) Another *C. Elegans* CompareProspector analysis showed that intestine genes have GATA transcription factor binding motif that was latter experimentally validated (Pauli et al., 2006).

3.3.3 RSAT

The Regulatory Sequence Analysis Tools (RSAT) is an integrated online tool to analyze regulatory sequences in co regulated genes (Thomas-Chollier et al., 2008). The only input required is a list of genes of interest; subsequently upstream sequences of desired distance can be retrieved and appended to the list. Further tasks of putative regulatory signals detection, matching positions for the detected signals in the original dataset can then be performed. The suite includes programs for sequence retrieval, pattern discovery, phylogenetic footprint detection, pattern matching, genome scanning and feature map drawing. Random controls can be performed with random gene selections or by generating random sequences according to a variety of background models (Bernoulli, Markov). The results can be displayed graphically highlighting the desired features. As RSAT web services is implemented using SOAP and WSDL, so either Perl, Python or Java scripts can used for developing custom workflows by combining different tools.

3.3.4 TFM

TFM (Transcription Factor Matrices) is a software suite for identifying and analyzing transcription factor binding sites in DNA sequences. It consists of TFM-Explorer, TFM-Scan, TFM-Pvalue and TFM-Cluster.

TFM-Explorer (Transcription Factor Matrix Explorer) proceeds in two steps: scans sequences for detecting all potential TF binding sites, using JASPAR or TRANSFAC and extracts significant transcription factor.

TFM-Scan is a program dedicated to the location of large sets of putative transcription factor binding sites on a DNA sequence. It uses Position Weight Matrices such as those available in the Transfac or JASPAR databases. The program takes as input a set of matrices and a DNA sequence. It computes all occurrences of matrices on the sequence for a given P-value threshold. The algorithm is very fast and allows for large-scale analysis. TFM-Scan is also able to cluster similar matrices and similar occurrences

TFM-Pvalue is a software suite providing tools for computing the score threshold associated to a given P-value and the P-value associated to a given score threshold. It uses Position Weight Matrices such as those available in the Transfac or JASPAR databases. The program takes as input a matrix, the background probabilities for the letters of the DNA alphabet and a score or a P-value.

3.3.5 rVISTA

rVISTA (regulatory VISTA) combines searching the major transcription binding site database TRANSFAC Professional from Biobase with a comparative sequence analysis and this procedure reduced the number of predicted transcription factor binding sites by several orders of magnitude (Loots and Ovcharenko, 2004). It can be used directly or through links in mVISTA, Genome VISTA, or VISTA Browser. Human and mouse sequences are aligned using the global alignment program AVID (Bray et al., 2003).

3.3.6 Mulan

Mulan brings together several novel algorithms: the TBA multi-aligner program for rapid identification of local sequence conservation and the multiTF program for detecting evolutionarily conserved transcription factor binding sites in multiple alignments. In addition, Mulan supports two-way communication with the GALA database; alignments of multiple species dynamically generated in GALA can be viewed in Mulan, and conserved transcription factor binding sites identified with Mulan/multiTF can be integrated and overlaid with extensive genome annotation data using GALA. Local multiple alignments computed by Mulan ensure reliable representation of short- and large-scale genomic rearrangements in distant organisms. Mulan allows for interactive modification of critical conservation parameters to differentially predict conserved regions in comparisons of both closely and distantly related species. The uses and applications of the Mulan tool through multispecies comparisons of the *GATA3* gene locus and the identification of elements that are conserved in a different way in avians than in other genomes, allowing speculation on the evolution of birds.

3.3.7 MotifVoter

MotifVoter is a variance based ensemble method for discovery of binding sites. It uses 10 most commonly used individual basic motif finders as its component (Wijaya et al., 2008). AlignACE (Hughes et al., 2000), MEME (Bailey and Elkan, 1994; Bailey et al., 2006), ANNSpec, Mitra, BioProspector, MotifSampler, Improbizer, SPACE, MDScan and Weeder. All programs can be selected individually or collectively. Though the existing ensemble methods overall perform better than stand-alone motif finders, the improvement gained is not substantial. These methods do not fully exploit the information obtained from the results of individual finders, resulting in minor improvement in sensitivity and poor precision.

3.3.8 ConSite

ConSite is a, web-based tool for finding cis-regulatory elements in genomic sequences (Sandelin et al., 2004). Two genomic sequences submitted for analysis are aligned by ORCA method. Alternatively, prealigned sequences can be submitted in ClustalW, MSF (GCG), Fasta or Pair wise BLAST format. For analysis Transcription factors can be selected on the basis of species, name, domain or user defined matrix (raw counts matrix or position weight matrix). Predictions are based on the integration of binding site prediction generated with high-quality transcription factor models and cross-species comparison filtering (phylogenetic footprinting). ConSite (Sandelin et al., 2004) is based on the JASPAR database (Portales-Casamar et al., 2009). By incorporating evolutionary constraints, selectivity is increased by an order of magnitude as compared to single sequence analysis. ConSite offers several unique features, including an interactive expert system for retrieving orthologous regulatory sequences.

3.3.9 OPOSSUM

OPOSSUM identifies statistically over-represented, conserved TFBSs in the promoters of co-expressed genes (Ho Sui et al., 2005). OPOSSUM integrates a precomputed database of predicted, conserved TFBSs, derived from phylogenetic footprinting and TFBS detection algorithms, with statistical methods for calculating overrepresentation. The background data set was compiled by identifying all strict one-to-one human/mouse orthologues from the Ensemble database. These orthologues were then aligned using ORCA, a pair-wise DNA alignment program. The conserved non-coding regions were identified. The conserved regions which fell within 5000 nucleotides upstream and downstream of the transcription start site (TSS) were then scanned for TF sites using the position weight matrices (PWMs) from the JASPAR database (Portales-Casamar et al., 2009). These TF sites were stored in the OPOSSUM database and comprise the background set.

3.3.10 TOUCAN2

TOUCAN 2 is an operating system independent, open source, JAVA based workbench for regulatory sequence analysis (Aerts et al., 2004). It can be used for detection of significant transcription factor binding sites from comparative genomics data or for detection of combinations of binding sites in sets of co expressed/co regulated genes. It tightly integrates with Ensemble and EMBL for retrieval of sequences data. TOUCAN provides options to align sequences with different specialized algorithms *viz* AVID (Bray et al., 2003), LAGAN (Brudno et al., 2003), or BLASTZ. MotifScanner algorithm is used to search occurrence of sites of transcription factors by using libraries of position weight matrices from TRANSFAC 6 (Matys et al., 2003), JASPAR, PLANTCARE (Lescot et al., 2002), SCPD and others. Motif Sampler can be used for detection of over-represented motifs. More significantly TOUCAN provides an option to select cis-regulatory modules using the ModuleSearch In essence TOUCAN 2 provides one of the best integration of different algorithms for identification of cis regulatory elements.

3.3.11 WebMOTIFS

WebMOTIFS web server combines TAMO and THEME tools for identification of conserved motifs in co regulated genes (Romer et al., 2007). TAMO combines results from four motif discovery programs viz AlignACE, MDscan, MEME, and Weeder, followed by clustering of results (Gordon et al., 2005). Subsequently Bayesian motif analysis of known motifs is done by THEME. Thus it integrates *de novo* motif discovery programs with Bayesian approaches to identify the most significant motifs. However, current version of Web MOTIFS supports motif discovery only for S. *cerevisiae, M. musculus*, and *H. sapiens* genomes.

3.3.12 Pscan

Pscan is a software tool that scans a set of sequences (e.g. promoters) from co-regulated or co-expressed genes with motifs describing the binding specificity of known transcription factors (Zambelli et al, 2009). It assesses which motifs are significantly over or underrepresented, providing thus hints on which transcription factors could be common regulators of the genes studied, together with the location of their candidate binding sites in the sequences. Pscan does not resort to comparisons with orthologous sequences and experimental results show that it compares favorably to other tools for the same task in terms of false positive predictions and computation time.

4. Identification of regulatory elements in legumin promoter

The plant seeds are rich source of almost all essential supplements of diet *viz*. proteins, carbohydrates, and lipids. Seeds not only provide nutrients to germinating seeding but are major source of energy and other cellular building blocks for human and other heterotrophs as well. Consequently, there is a plethora of pathogens attacking plant seeds. Certain pests such as coleopteran insects of the family Bruchidae, have evolved with leguminous plants (Sales et al., 2000). It is believed that these seeds, most of which are not essential for the establishment of the new plant following germination, contribute to the protection and defense of seeds against pathogens and predators.

Genes encoding seed storage proteins, like zein, phaseolin and legumin, were among the first plant genes studied at gene expression level. Hetrologous expression of reporter genes confirmed that there promoters direct the seed specific expression and subsequently these seed storage gene promoters were used for developing transgenic plants for expressing different genes of interest. Earlier we isolated legumin gene promoter from pigeon pea (*Cajanus cajan*) and identified regulatory elements in its promoter region(Jaiswal et al., 2007). Sequence analysis with PLACE, PLANTCARE, MATINSPECTOR and TRANSFC shows that legumin promoter not only consist of regulatory elements for seed specific expression but also have elements that are present in promoters of other genes involved in pathogen defense, abiotic stress, light response and wounding (Table 1). Our pervious study also confirmed that legumin promoter in expressed in the developing seedling as well (Jaiswal et al., 2007). Recent studies have shown that these promoters are expressed in non seed tissues as well (Zakharov et al., 2004) and play a role in bruchid (insect) resistance (Sales et al., 2000).

In such a scenario where seed storage protein performs an additional task of pathogen defense its prompter must have responsive elements to such stresses. In fact legumin promoter consists of transcription factor binding site for wounding, a signal of insect attack and pathogen defense (Table 1). Since promoter sequences are available for legumin promoter from different species it becomes a good system for identification novel regulatory elements in these promoter. Phylogenetic footprinting analysis reveals presentence of another conserved motif 19 base pair downstream to legumin box (Fig. 1), named L-19 box (Jaiswal et al., 2007). Further MEME analysis shows that in addition to four conserved blocks present for TATA box, G box, Legumin box and L-19 box, there are other conserved, non overlapping sequence blocks are present that were not revealed by multiple sequence alignment based phylogenetic footprinting (Fig. 2).

5. Conclusion

With the critical role of cis-regulatory elements in differentiation of specific cells leading to growth, development and survival of an organism, scientists have a great interest in their identification and characterization. However, due to the limited availability of known transcription factors identification of the corresponding regulatory elements through conventional DNA-protein interaction techniques is challenging. Rapid increase in number of complete genome sequences, identification of co-regulated genes through microarray technology with available electronic storage and computational power has put before the scientific community a challenge to integrate these advancing technology and develop computational program to identify these short but functionally important regulatory sequences. Although some progress has been made in this direction and databases like TRANSFC and others store libraries of transcription factor binding sites. However, there are limitations primarily because publicly available libraries are very small and complete datasets are not freely available. Secondly, because of their very small size there is certain degree of redundancy in binding and therefore the chances of false prediction are very high. These limitations have been overcome to some extent by databases like JASPAR that are freely available and have a collection of regulatory elements large enough to compete with the commercially available datasets. Another concern with cis-acting regulatory elements is that the data pool of these functional non coding transcription factor binding sites is very small (a few thousands), compared with the fact that thousand of genes are expressed in a cell at any point of time and every gene is transcribed by a combination of minimum 5-10 transcription factors. Phylogenetic footprinting, has enormous potential in identifying new regulatory elements and completing the gene network map. Although routine sequence alignment programs such as clustalW fail to align short conserved sequences in a background of hyper variable sequences, more sophisticated sequence alignment programs have been specially developed for identification of conserved regulatory elements. These programs such as CONREAL uses available transcription factor binding site data to align the two sequences that decreases the chances of missing a regulatory site considerably. Moreover, other approaches such as MEME altogether abandons the presence of sequence alignment and directly identifies the rare conserved blocks even if they have jumbled up to the complementary strand. With the increasing sophistication and accuracy of motif finding programs and available genome sequences it can be assumed that knowledge of these regulatory sequences will definitely increase (Table 2). Once we have sufficient data it can be used for development of synthetic promoters with desired expression patterns (Fig. 3).

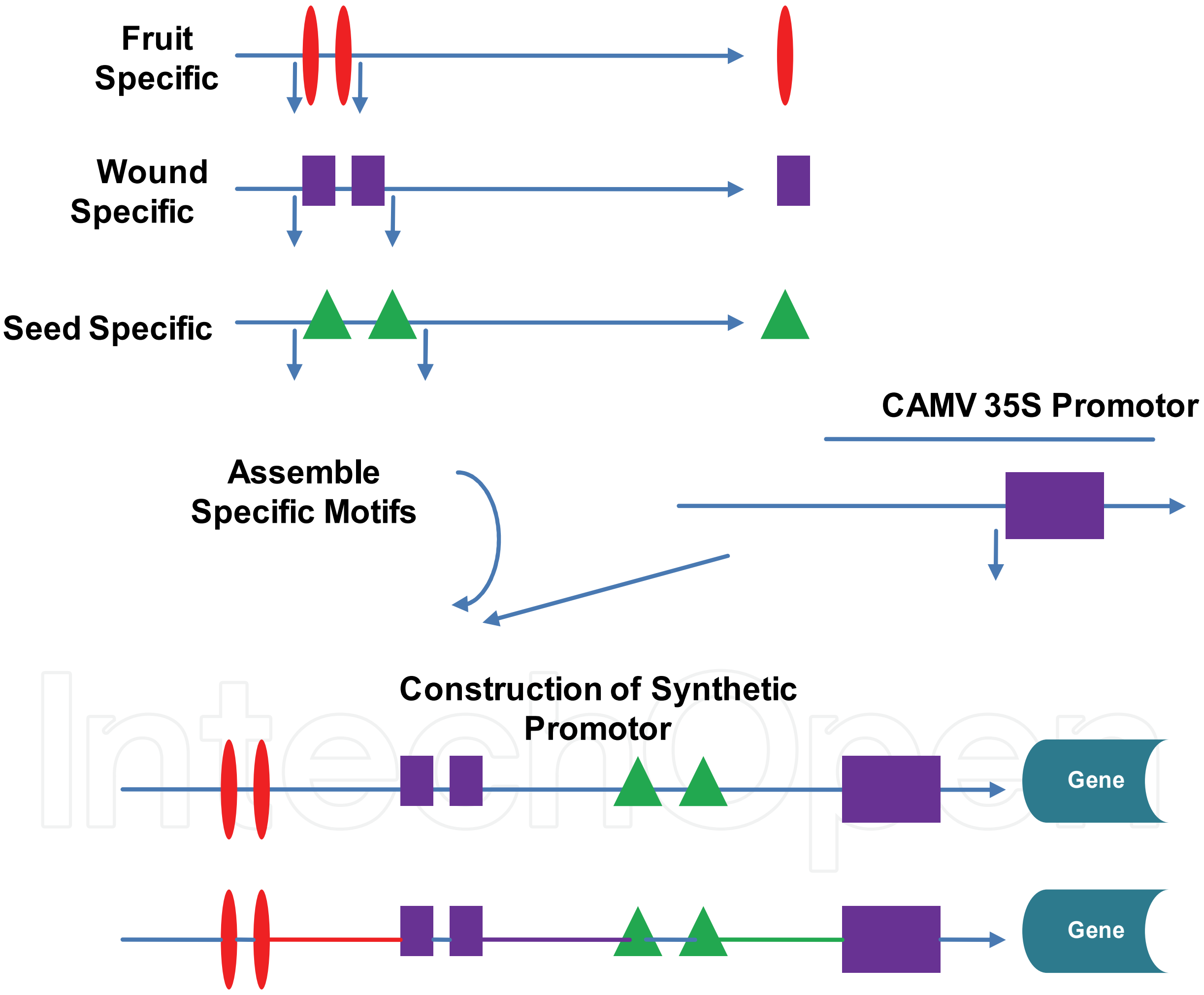


Fig. 3. Future prospects: development of synthetic promoters for expression of gene of interest in desired tissue at defined time and developmental stage. Example: Regulatory elements can be combined from wound, fruit and seed specific promoters and combined with strong CaMV 35S promoter for high level expression of desired gene in all these tissues.

Tools	**Web site**
Sequence Alignment	
Blast-Z	http://www.bx.psu.edu/miller_lab/
Dialign	http://dialign.gobics.de/chaos-dialign-submission
AVID (Mvista)	http://genome.lbl.gov/vista/mvista/submit.shtml
Lagan	http://lagan.stanford.edu/lagan_web/index.shtml
Clustal W	http://www.ebi.ac.uk/Tools/msa/clustalw2/
TF Binding Site search	
Consite	http://www.phylofoot.org/consite
CONREAL	http://conreal.niob.knaw.nl/
PromH	http://www.softberry.com/berry.phtml?topic=promhg&group=programs&subgroup=promoter
Trafac	http://trafac.cchmc.org/trafac/index.jsp
Footprinter	http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinterWeb/FootPrinterInput2.pl
rVISTA	http://rvista.dcode.org/
TFBIND	http://tfbind.hgc.jp/
TESS	http://www.cbil.upenn.edu/cgi-bin/tess/tess
TFSearch	http://www.cbrc.jp/research/db/TFSEARCH.html
Toucan	http://homes.esat.kuleuven.be/~saerts/software/toucan.php
Phyloscan	http://bayesweb.wadsworth.org/cgi-bin/phylo_web.pl
OFTBS	http://www.bioinfo.tsinghua.edu.cn/~zhengjsh/OTFBS/
PROMO	http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
R-Motif	http://bioportal.weizmann.ac.il/~lapidotm/rMotif/html/
Motif Finding	
MEME	http://meme.sdsc.edu/meme4_6_1/intro.html
AlignAce	http://atlas.med.harvard.edu/cgi-bin/alignace.pl
MotifVoter	http://compbio.ddns.comp.nus.edu.sg/~edward/MotifVoter2/
RSAT	http://rsat.scmbb.ulb.ac.be/rsat/
Gibbs Sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html
BioProspector	http://ai.stanford.edu/~xsliu/BioProspector/
MatInspector	http://www.genomatix.de/
Improbizer	http://users.soe.ucsc.edu/~kent/improbizer/improbizer.html
WebMOTIFS	http://fraenkel.mit.edu/webmotifs-tryit.html
Psacn	http://159.149.109.9/pscan/
FootPrinter	http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinterWeb/FootPrinterInput2.pl

Table 2. Regulatory sequences identification programs.

6. References

Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y., and De Moor, B. (2004). TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Research 33, W393-W396.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34, W369-W373.

Berezikov, E., Guryev, V., and Cuppen, E. (2005). CONREAL web server: identification and visualization of conserved transcription factor binding sites. Nucleic Acids Research 33, W447-W450.

Blanchette, M., and Tompa, M. (2002). Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. Genome Research 12, 739-748.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A Global Alignment Program. Genome Research 13, 97-102.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Program, N.C.S., Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Research 13, 721-731.

Buhler, J., and Tompa, M. (2002). Finding motifs using random projections. J Comput Biol 9, 225-242.

Carmack, C.S., McCue, L., Newberg, L., and Lawrence, C. (2007). PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms for Molecular Biology 2, 1.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933-2942.

Che, D., Jensen, S., Cai, L., and Liu, J.S. (2005). BEST: Binding-site Estimation Suite of Tools. Bioinformatics 21, 2909-2911.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003). Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting. Science.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. (2001). Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. Genome Research 11, 1175-1186.

Das, M., and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. BMC Bioinformatics 8, S21.

Fiedler, T., and Rehmsmeier, M. (2006). jPREdictor: a versatile tool for the prediction of cis-regulatory elements. Nucleic Acids Research 34, W546-W550.

Francis, C., and Henry, C.M.L. (2008). DNA Motif Representation with Nucleotide Dependency. IEEE/ACM Trans. Comput. Biol. Bioinformatics 5, 110-119.

Frith, M.C., Hansen, U., and Weng, Z. (2001). Detection of cis -element clusters in higher eukaryotic DNA. Bioinformatics 17, 878-889.

Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical overâ€ representation. Nucleic Acids Research 32, 1372-1381.

Gordon, D.B., Nekludova, L., McCallum, S., and Fraenkel, E. (2005). TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics 21, 3164-3165.

Henry, C.M.L., and Fracis, Y.L.C. (2006). Discovering DNA Motifs with Nucleotide Dependency, Y.L.C. Francis, ed, pp. 70-80.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-577.

Hertz, G.Z., Hartzell, G.W., and Stormo, G.D. (1990). Identification of Consensus Patterns in Unaligned DNA Sequences Known to be Functionally Related. Comput Appl Biosci 6, 81 - 92.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research 27, 297-300.

Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., and Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Research 33, 3154-3164.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of Molecular Biology 296, 1205-1214.

Jaiswal, R., Nain, V., Abdin, M.Z., and Kumar, P.A. (2007). Isolation of pigeon pea (Cajanus cajan L.) legumin gene promoter and identification of conserved regulatory elements using tools of bioinformatics. Indian Journal of experimental Biology 6, 495-503.

Jensen, S.T., and Liu, J.S. (2004). BioOptimizer: a Bayesian scoring function approach to motif discovery. Bioinformatics 20, 1557-1564.

Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences 87, 2264-2268.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262, 208-214.

Lescot, M., DÃ©hais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., RouzÃ, P., and Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Research 30, 325-327.

Liu, X., Brutlag, D.L., and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138.

Liu, X.S., Brutlag, D.L., and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotech 20, 835-839.

Liu, Y., Liu, X.S., Wei, L., Altman, R.B., and Batzoglou, S. (2004). Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics. Genome Research 14, 451-458.

Loots, G.G., and Ovcharenko, I. (2004). rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Research 32, W217-W221.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. (2000). Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons. Science 288, 136-140.

Marinescu, V., Kohane, I., and Riva, A. (2005). MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinformatics 6, 79.

Matys, V., Fricke, E., Geffers, R., GÃ¶ÃŸling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., MÃ¼nch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFACÂ®: transcriptional regulation, from patterns to profiles. Nucleic Acids Research 31, 374-378.

Pauli, F., Liu, Y., Kim, Y.A., Chen, P.-J., and Kim, S.K. (2006). Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. Development 133, 287-295.

Pennacchio, L.A., and Rubin, E.M. (2001). Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2, 100-109.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2009). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., RouzÃ©, P., and Van de Peer, Y. (2003). Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes. Plant Physiology 132, 1162-1176.

Romer, K.A., Kayombya, G.-R., and Fraenkel, E. (2007). WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. Nucleic Acids Research 35, W217-W220.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotech 16, 939-945.

Sales, M.Ì.c.P., Gerhardt, I.R., Grossi-de-SÃ¡, M.F.t., and Xavier-Filho, J. (2000). Do Legume Storage Proteins Play a Role in Defending Seeds against Bruchids? Plant Physiology 124, 515-522.

Sandelin, A., Wasserman, W.W., and Lenhard, B. (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Research 32, W249-W252.

Sinha, S., Liang, Y., and Siggia, E. (2006). Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Research 34, W555-W559.

Stormo, G.D. (2000). DNA Binding Sites: Representation and Discovery. Bioinformatics 16, 16 - 23.

Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R.s., Defrance, M., Vervisch, E., Brohee, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Research 36, W119-W127.

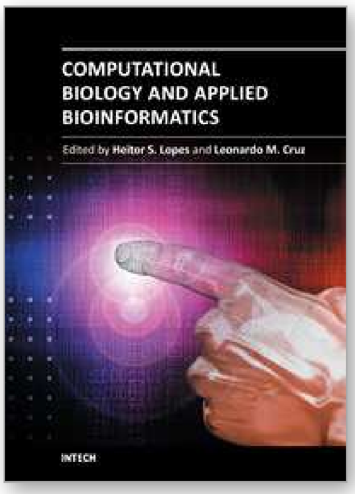
Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673-4680.

Wang, T., and Stormo, G.D. (2005). Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proceedings of the National Academy of Sciences of the United States of America 102, 17400-17405.

Wijaya, E., Yiu, S.-M., Son, N.T., Kanagasabai, R., and Sung, W.-K. (2008). MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. Bioinformatics 24, 2288-2295.

Zakharov, A., Giersberg, M., Hosein, F., Melzer, M., MÃ¼ntz, K., and Saalbach, I. (2004). Seed-specific promoters direct gene expression in non-seed tissue. Journal of Experimental Botany 55, 1463-1471.

Zhu, J., Liu, J.S., and Lawrence, C.E. (1998). Bayesian adaptive sequence alignment algorithms. Bioinformatics 14, 25-39.


COMPUTATIONAL BIOLOGY AND APPLIED BIOINFORMATICS
Edited by Heitor S. Lopes and Leonardo M. Cruz
INTECH

Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
Publisher InTech
Published online 02, September, 2011
Published in print edition September, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821