Chapter 18
In Silico Discovery of DNA Regulatory Sites and Modules

Panayiotis V. Benos

Abstract In this chapter we describe methods commonly applied for pattern representation and some analogies to the physical world of protein–DNA interactions. Next we present the general methodology for *de novo* DNA pattern discovery in a set of promoter sequences with and without prior information and we discuss some well-established algorithms. The common problems of pattern matching, i.e., the prediction of a site based on prior information are discuss along with the contribution of the evolutionary information to the pattern discovery and pattern matching algorithms. Finally, in we introduce the topic of *cis-regulatory modules* (CRMs) and some of the algorithms designed to find them.

Keywords Bioinformatics · Genetics · Genomics · Transcription · DNA regulatory regions

18.1 Introduction

18.1.1 Role of TFs in Regulation of Gene Expression

Systems Biology aims to the understanding of the interactions between the various molecular components in the cell and between the cell and its extra-cellular environment. Cells respond continuously to a constantly changing environment by adjusting the expression levels of their genes. This applies to responses to extracellular stimuli, developmental needs, and differentiation. A fundamental control of the expression of every gene takes place at *transcription*, during which the genomic DNA is copied ("transcribed") into RNA. The *rate of transcription* is influenced by various factors. On a large scale, the status of the chromatin ("opened" or "closed" form) determines whether the genomic DNA is accessible or not to interacting proteins, which has been associated with activation or silencing of large parts of the genome. On a gene-by-gene basis, transcription regulation is achieved for the most part by the presence or absence of *transcription factor* (*TF*) proteins in the nucleus. TFs recognize short DNA "signals" (typically 6–15 bp long) in the vicinity of the genes' *transcription start site* (TSS) and they have the potential to initiate (*activators*) or repress (*repressors*) the transcription of nearby genes. These "DNA signals" are commonly referred to as *transcription factor binding sites* (TFBSs) or more generally as *cis-regulatory elements*.

P.V. Benos
Department of Computational Biology, School of Medicine, 3501 Fifth Avenue, Suite 3064 BST3, Pittsburgh, PA, 15260, USA
e-mail: benos@pitt.edu

S. Krawetz (ed.), *Bioinformatics for Systems Biology*, DOI 10.1007/978-1-59745-440-7_18,
© Humana Press, a part of Springer Science+Business Media, LLC 2009

18.1.1.1 Organization of the Promoter in Prokaryotes and Eukaryotes

The genomic region with the TFBSs that control the expression of a gene is called the *promoter*. The size and the organization of a promoter differ depending on the organism and the gene under study. Protein coding genes have a set of *core promoter elements*, which are characteristic of the domain the organism belongs. Prokaryotic organisms (like bacteria) have two distinct regulatory elements in their core promoters: the *Pribnow box*, which is located about 10 bp *upstream* of the TSS (typical sequence: TATAAT) and a second element located around position −35 (typical sequence: TTGACA). Eukaryotic core promoters are more variable. However, many of them contain a *TATA box* at position −30 from the TSS (typical sequence: TATAAA) and a *CAAT-box* at around 75–100 bases upstream of the TSS (typical sequence: CCAAT).

Besides these core elements, a number of gene-specific DNA elements regulate the appropriate transcription rate of a gene. These are the target sites of the sequence-specific transcription factor proteins. They are located either upstream of the gene's TSS or downstream, in its exons or introns (for the eukaryotic genes). The distance at which these elements can be found varies from some hundreds of bases in bacteria or single cell eukaryotes (e.g., yeast) to 10–20,000 bases in complex eukaryotes (e.g., in the fruitfly *Drosophila melanogaster*). Sometimes the more distant regulatory elements, also known as *enhancers*, are not sufficient to drive transcription but they rather play an assisting role. Mutations on TFBSs have been associated with diseases, like β-thalassemia in humans. Schematic views of typical prokaryotic and eukaryotic core promoters are presented in Fig. 18.1.

18.1.1.2 Protein-DNA interactions

The DNA *cis*-regulatory sites are recognized by the transcription factor proteins *via* the formation of chemical bonds between the amino acids and the DNA. Hydrogen bonds as well as van der Waals interactions and water-mediated bonds have been observed in crystal structures. Most of these interactions, are not sequence-specific, i.e., they are formed between the DNA backbone and the protein. As such they do not contribute much to the binding specificity except through *indirect reading*, which is associated with the "bendability" potential of the DNA and other DNA

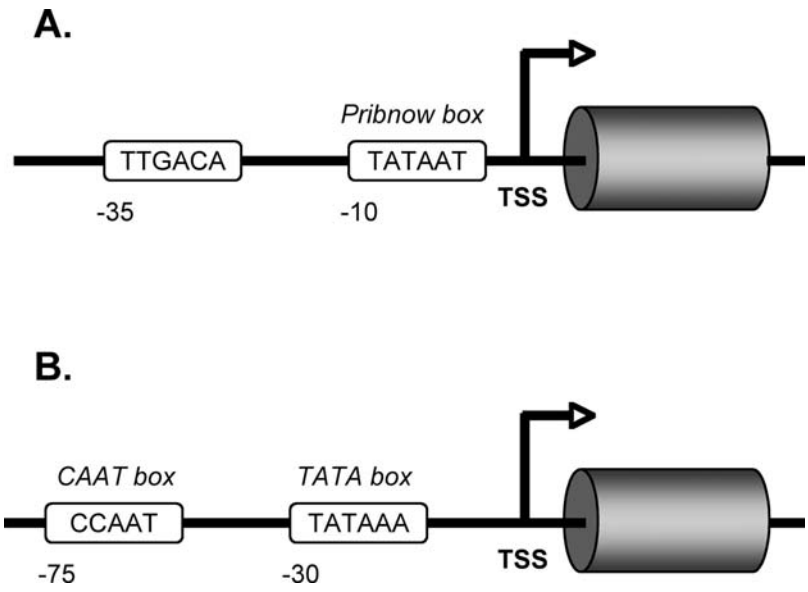

Fig. 18.1 Schematic view of prokaryotic **(A)** and eukaryotic **(B)** core promoters. The elements of the general transcription factors as well as examples of gene-specific binding sites are depicted (Copies of figures including color copies, where applicable, are available in the accompanying CD)

properties. Nevertheless, most of the sequence specificity is obtained *via* hydrogen bonds between the amino acids and bases.

In most cases, the TF proteins use α-helices to attack the DNA major groove. Each helix can "read" only 3–4 bases, which is an insufficient length for unique target recognition in a genome of thousands or millions of bases. Therefore, TFs utilize multiple α-helices to build composite DNA patterns (e.g., proteins of the C_2H_2 zinc-finger family), or a combination of multiple helices and amino acid tails that contact the bases in both the major and the minor grooves (e.g., some homeodomain proteins, like Drosophila *engrailed*). In other cases, proteins with short DNA targets (4–6 bases long) can form homodimers, hence expanding their total target length (e.g., bZIP and bHLH proteins). In addition, eukaryotic organisms are able to inactivate large parts of the genome, by making them inaccessible to TFs through chromatin packaging or *insulator* molecules (like CTCF protein).

In terms of computational searching for DNA patterns in the promoter regions of the genes, two questions are frequently asked: (i) given a set of co-regulated genes (e.g., from microarray experiments), what are the putative DNA binding sites of the (unknown) TFs that are regulated by them? (ii) given a single promoter, does TF *X* regulate the associated gene? A number of methods have been developed to address the first question. The underlying hypothesis of these methods is that all or most of the promoters contain binding sites for the same TF(s). In Section 18.3, we will describe the most important of these strategies and algorithms. For the second question, the straight-forward approach is to scan the promoter of the gene of interest with the binding site pattern of the TFs of interest. In this way, putative sites are predicted based on some user-defined threshold. Typically, this method results in a high number of false positive predictions, which can be reduced by utilizing evolutionary information. The underlying hypothesis here, is that biologically important parts of the promoter (e.g., those containing binding sites of TFs important for that gene's expression) will evolve at a slower rate and thus will tend to be more conserved between species.

However, the hypotheses underlying both these methodologies have been known to be true only to a certain extent. Noise in microarray data and uncertainties associated with the data analyses may result in the inclusion of irrelevant genes in the set of co-regulated genes. This will increase the noise of the method, especially when analyzing promoters of complex eukaryotic organisms, where usually large parts of the genomic DNA upstream and downstream of the TSS are considered. For the second methodology, it has been recently shown that in some organisms TFBSs have a high turnover. In other words, sites that are functional in one organism may become non-functional in another organism and may get replaced by sites in other parts of the promoter region. Despite those known limitations, these general methodologies are used to address these two very important questions.

18.2 DNA Pattern Representation

Typically, each TF recognizes multiple DNA target sites in a sequence-specific manner. The target sites are frequently viewed as variants of a "preferred" (*consensus*) site. Given a set of known binding sites of a TF, the first question is: what is the best way to organize the information so that it becomes useful in the search of other, yet unknown, sites? When the number of known TFBSs is small or their variation is limited, they can be used directly in simple string searches to identify new occurrences. This is the case of the *c-myc* TF. The vast majority of the known *c-myc* targets contain the sequence CACGTG, although another target CATGTG or its reverse-complement (CACATG) has also been reported. With such limited repertoire, one can directly scan a genomic DNA

sequence and seamlessly identify all potential *c-myc* sites. Given that most of the known *c-myc* sites are CACGTG, biologists will naturally have more "trust" in predictions with this hexanucleotide than in the other two forms when it comes to prioritizing their tests.

18.2.1 Representation by Consensus

Most of the TFs have more than three known target forms, in which case a straightforward method to summarize and present the binding preferences is done by calculating their *consensus sequence*. The targets are aligned on the top of each other and for each position the IUPAC code is used to denote bases and base ambiguities (Table 18.1). A regular expression search can then be employed to predict new sites of this TF. Consensus target representation is used frequently in the literature to effortlessly present positional base preferences, but it has limited value when it comes to searching for new sites. In order to better understand why, let's consider the *c-myc* example above. The consensus representation of the three *c-myc* sites would be CAYRTG. This representation will capture all three known sites, but it will also yield sequence CATATG as putative *c-myc* target, which is not known to be true. For TFs with longer and more variable targets, the consensus may quickly deteriorate into a useless pattern as more variable functional sites are discovered and added to the consensus. Consider, for example, the aligned target sequences of Fig. 18.2A. If we knew only the top eight targets, then the consensus pattern would have been GGRHKTYCCC, which would have detected 2.3 putative "hits" on an average on a random DNA sequence about 100,000 bases long (assuming equal background probabilities for all bases). These are the sites generated under a background model, so they represent the amount of "noise" or false positive predictions one might expect. When all known sites of Fig. 18.2A are considered, the consensus pattern becomes RGRNDNYYMH, which will detect 4.4 sites for every 1,000 bases on average. So, the latter pattern will generate ~200 times more false positive sites than the former, although the former will miss more of the (total) known sites of Fig. 18.2A.

Furthermore, the consensus representation provides no insights on the quantitative nature of the binding. The example of the three *c-myc* sites, it neglects the fact that CACGTG is a much more frequent *c-myc* target than all the rest. Clearly, a probabilistic way to represent these patterns should be more powerful.

Table 18.1 IUPAC codes for DNA bases and their combinations

Symbol	Combination	Name
A		Adenosine
C		Cytidine
G		Guanine
T		Thymidine
M	A C	*Amino*
R	A G	*Purine*
W	A T	*Weak*
S	C G	*Strong*
Y	C T	*Pyrimidine*
K	G T	*Keto*
V	A C G	
H	A C T	
D	A G T	
B	C G T	
N	A C G T	

(Copies of tables are available in the accompanying CD.)

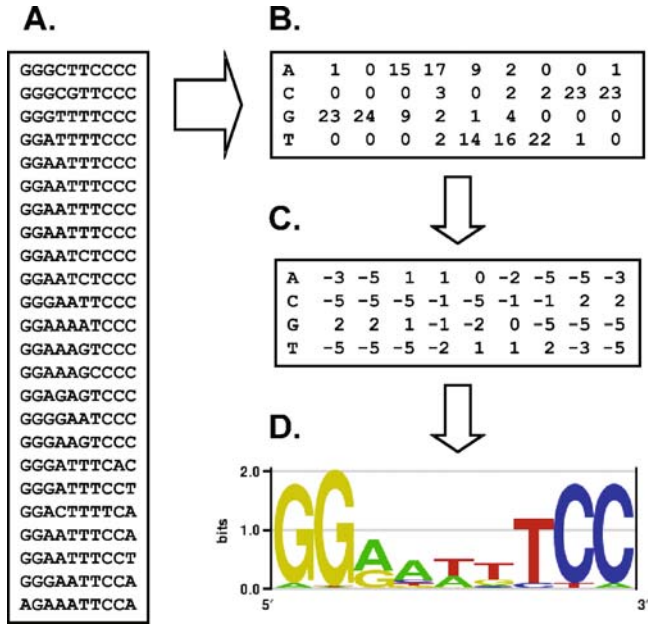

Fig. 18.2 PSSM model representation. A set of aligned target sites **(A)** can be used to generate a position-specific *count matrix* **(B)** The count matrix is later transformed into a log-frequency matrix **(C)** (with or without background correction). Any of the three forms of information can be graphically represented as a LOGO of symbols **(D)** using the appropriate software (Copies of figures including color copies, where applicable, are available in the accompanying CD)

18.2.2 Representation by Weight Matrices

The *Position-Specific Scoring Matrices* (*PSSM*s; also known as *Position Weight Matrices* or *PWM*s) is the most widely used way to capture and represent binding site information. Like consensus sequences, PSSM models are generated from a set of aligned known sites, but each position is represented by a set of four weights that correspond to the likelihood of each base appearing in this position of the target sequence. The construction method of a PSSM model is illustrated in Fig. 18.2. For each position, I, of the alignment, the four-dimensional vector constitutes the log-likelihood of the observed frequency of each base at this position over an expected background frequency (Fig. 18.2C). These weights provide a quantitative measure of how frequently a particular base is observed in the set of known sites as opposed to the background (e.g., in the genome). Positions that are more important for the TF binding tend to have higher log-likelihood ratios. The composite model of all L vectors (L is the length of the pattern) is the $4 \times L$ PSSM model. The average log-likelihood for each position is the *relative entropy*, which is formally defined as:

$$RH(I) = \sum_{b=A}^{T} f(b,I) \cdot \ln \frac{f(b,I)}{P_{ref}(b)} \tag{18.1}$$

where $f(b,I)$ is the estimated frequency of base b at position I of the pattern and $P_{ref}(b)$ is the background frequency of base b (e.g., in the genome). Averaging all the L positions, we obtain the average relative entropy of the motif. A number of motif finding algorithms identify patterns that maximize either the overall log-likelihood of the motif or its relative entropy.

Although they are derived from Shannon information theory, PSSM models have an interesting explanation that is based on the thermodynamic properties of the proteins and DNA [1]. Assuming that a TF, P, interacts freely with its genomic DNA targets (i.e., no protein-protein synergistic or competing actions take place and no change in the protein concentration occurs during the

interaction), then under equilibrium one would expect that the probability that a given target, D, is bound by the TF would be:

$$P(D \cdot P) = \frac{P_{ref}(D) \cdot e^{-H(D,P)/RT}}{\sum_{D_i} P_{ref}(D_i) \cdot e^{-H(D_i,P)/RT}} \tag{18.2}$$

where $P_{ref}(D)$ is the background frequency of D and $H(D, P)$ is the energy of the interaction between the target and the protein. The denominator is the *partition function*, which ensures that the sum of the probabilities over all targets, D_i, will be 1. Based on that, the log-likelihood PSSM score (see above) is equal to the negative energy of the interaction in RT units (plus some constant).

PSSM models, like the consensus patterns, assume that the positions in the DNA target are *independent* in their contribution to the overall TF-DNA specificity. In other words, the observed base frequencies at position I are independent of the frequencies in any other position. In equation 18.2, this translates into: the energy function being further partitioned on individual base-amino acid contacts to simplify their calculation. This is known as the *additivity assumption* and has been highly debated over the years. Studies on individual TFs show that although this assumption is not completely accurate, it is still useful in many cases, in the sense that it produces reliable predictions [2]. From a biologist point of view, this is what matters most. Few models have been proposed that consider higher order of position dependencies [3, 4], but the limited data availability for most TFs makes these models inefficient due to the problem of *overfitting*.

18.2.3 Scoring Sites with a PSSM Model

There are many ways to interpret a PSSM model. The most useful one is perhaps as a probabilistic model of a set of known sites (for a very good review, see [5]). According to that, a frequency matrix is a multinomial distribution that can generate target sites. Hence, we can measure the *probability* that a given site, D_k, was generated by this frequency matrix by multiplying the probabilities of the D_k bases at the corresponding positions. However, multiplying the probabilities can quickly lead to a memory underflow problem. It is usually more convenient to calculate the sum of log-likelihoods instead. When the PSSM model, W, is the log-likelihood ratio of the frequency matrix over the background frequencies, the following score function can be used:

$$S(\vec{D}_k|W) = \sum_{i=1}^{L} \sum_{b=A}^{T} w(b,i) \cdot \delta(D_k(i), b) \tag{18.3}$$

where $w(b, i)$ is the PSSM weight of base b at position I in D_k, and δ is the *Kronecker's delta* function, which is 1 if the i-th base of the D_k is b and zero otherwise. Assuming that the PSSM scores have taken into consideration the background frequencies, a positive composite score will generally mean that the sequence, *seq_x*, is more likely to have been produced by the model that generated the real target sites than by the background model. For example, the highest scoring sequence in the model of Fig. 18.2, is GGAATTTCC (score = 14), whereas one of the lower scoring sequences is TTTCCAAAT (score = −38).

18.2.4 Visual Pattern Representation with LOGOs

We, humans, can much easily comprehend and appreciate biological patterns when they are presented in a graphical rather than in a numeric form. In 1986, Schneider and colleagues developed

a method to graphically represent the DNA and amino acid patterns [6, 7]. The method represents each position of the multiple-sequence alignment as a stack of symbols (bases, amino acids). For DNA sequences, the total height of the stack at position I is equal to:

$$H(I) = 2 + \sum_{b=A}^{T} f(b, I) \cdot \log_2 f(b, I) \tag{18.4}$$

where $f(b, I)$ is the observed frequency of base b at position I. This formula calculates the decrease in *entropy* (from a maximum value of two for DNA sequences in $\log_2$ scale) due to uncertainty. For a column with maximum uncertainty (equally probable bases), $H(I) = 0$. For a column with maximum conservation or zero uncertainty [when one base has $f(b,I) = 1$], $H(I)$ takes its maximum value, i.e.,$H(I) = 2$. Once the stack's height, $H(I)$, has been determined, the height of the individual symbols within the stack is calculated proportionally to its relative frequency, $f(b,I)$. The LOGO of the frequency matrix of the alignment in Fig. 18.2A is presented in Fig. 18.2D as an example. Note that the *relative entropy* in equation 18.1 is the *entropy* in equation 18.4 normalized for the background frequencies.

18.3 De Novo Pattern Discovery

Let's now focus on the most common problem related to the identification of regulatory sequences (for a very good review of various methods, see [5]). Suppose we have a list of genes that are found to be co-regulated. These, for example, can be genes resulting from microarray or other gene expression data analysis. A reasonable assumption is that a large proportion of these co-regulated genes share a set of TFs that regulate them by binding to conserved DNA elements in their corresponding promoter regions. In most cases the identity of these TFs and their DNA binding preferences are unknown so the problem can be formalized as: *given a set of (unaligned) promoter regions identify common DNA elements that are likely to be targets of some TF.*

Several methods have been developed to address this problem. Typically, these methods search for DNA patterns over-represented in these sequences compared to some "background" or "expected" frequency. The methods differ on the objective function they try to optimize, the computational algorithm they use for this optimization and the background model they consider. Below we describe some general principles of these methods.

18.3.1 Finding Patterns with a Greedy Search

One of the first algorithms for finding patterns in unaligned DNA sequences was *Consensus*, developed by Hertz and Stormo in 1990 [8, 9]. Program *Consensus* (not to be confused with the *consensus representation* of binding patterns described before) uses a greedy algorithm in order to identify the set of sequences of a given length that maximize the pattern's *information content*, formally defined as:

$$IC(pattern) = \sum_{I=1}^{L} \sum_{b=A}^{T} f(b, I) \cdot \log_2 f(b, I) \tag{18.5}$$

where $f(b,I)$ is the frequency of base b at position I. For a given pattern length L, the algorithm starts by creating one sequence matrix for each sequence L-mer "word". In the subsequent cycles, this matrix is compared against all the other sequences and ranks the resulting pairwise alignments according to their information content. In order to reduce the search space and time, in each cycle, only a percent of the examined patterns is retained (e.g., 10% matrices with the highest information content). In its present form the algorithm manages to keep the calculation cost very low [$O(N^2)$ in time and $O(N)$ in space for N sequences].

18.3.2 Finding Patterns with Iterative Optimization

One general algorithm for the maximum likelihood parameter estimation in probability mixture models, is the *expectation-maximization* or *EM*. In 1990, Lawrence and Reilly applied it for the first time in biological pattern finding [10], although perhaps the most famous implementation as of today is program *MEME*, developed in 1993 by Bailey and Elkan [11]. Assuming a set of observed quantities (e.g., *binding sites*) determined from a set of *hidden* or *missing data* (e.g., PSSM model of TF-DNA binding), the EM algorithm seeks to identify the PSSM model that best explains the observed quantities. However, since the binding sites are also unknown, the EM algorithm iterates between predicted sites and PSSM models. For each iteration the current PSSM model is used to calculate the *expected likelihood* of all subsequences of length L in the unaligned promoter data-set. The set that *maximizes* this expectation is the new (predicted) set of binding sites, which is used to calculate the new PSSM model to be used in the next cycle. In other words, in each cycle the algorithm selects the parameters (PSSM model) that maximize the expected likelihood of the observed data calculated over the values of the missing data and model parameters. When the PSSM model does not change significantly or when a maximum number of iterations have been reached, the algorithm terminates. Initializations of the PSSM model can be done by performing a one step EM with all possible motif-starting points and then evaluating the results.

One of the disadvantages of the EM algorithms is that they can be trapped in local maxima. Another very popular algorithm, *Gibbs sampling*, is designed to overcome this problem. Gibbs sampling is very similar to EM, but in each cycle the candidate sites are selected randomly with their probability of selection calculated by their match to the current PSSM model. In this way, suboptimal patterns (with respect to the current PSSM model) can enter the next cycle hence help escaping local optima. Initialization of the Gibbs sampling algorithm is typically done by randomly selecting one "word" from every input sequence and by using them to calculate the first PSSM model. The first Gibbs sampling algorithm implementation for DNA and protein sequences was performed in 1993 by Charles Lawrence and colleagues [12], since then many other algorithms have implemented it [13, 14].

18.3.3 Other Pattern Finding Methods

A number of other strategies have been explored in the search for better methods for DNA pattern identification. These include artificial neural networks like perceptrons and self-organizing maps (e.g., programs *ANN-SPEC* [15] and *SOMBRERO* [16]). Also, they include the so called *dictionary-based* or *word enumerating* methods. The basic principle of the dictionary-based methods is that the biologically important patterns will show relatively small variability and there will be significantly more frequent in the test set (e.g., unaligned promoter sequences) than in some background. Thus, enumeration of all "words" and comparison of their frequencies to the expected background can help in the identification of the most statistically significant ones. In this context, the concept of the

"word" can be extended to include patterns with a predetermined variability (e.g., k of L bases to be conserved) or regular expression patterns (e.g., GG[AG]A[TA][TG][TC]CC for the pattern in Fig. 18.2D).

18.4 Pattern Matching

Now we switch to a different kind of problem. Assuming we know the binding preferences of a TF in a form of a weight matrix of length L, how can we predict if and where it binds in a given promoter? One straightforward way is to use the PSSM model to "scan" the promoter region and score each subsequence of length L (both strands) using equation 18.3. Note that since the weights, $w(b,i)$, in equation 18.3 are a function of the logarithms of the base probabilities from the model of the *observed* (known) sites, the above sum is the log-likelihood that the target comes from that set (corrected for the background base frequencies). Having calculated a score for each subsequence of the promoter region, the initial problem now changes into a problem of determining an appropriate threshold that discriminates between true positive and true negative sites. Unfortunately, this is not a trivial problem. The prediction of TF binding sites is notoriously difficult mainly due to a relatively low signal-to-noise ratio. Part of the problem is the incomplete modeling of the binding preferences. Higher order models have been proposed in the past, but with limited usefulness, due to the small number of known sites for most TFs. However, the high false positive rate may also be an inherent property of the TFs themselves. After all, a TF needs to specifically recognize its targets, but the protein-DNA association also needs to be easily reversible.

Having noted this, there are two main methodologies one can follow to set up a threshold. According to the first method, given a set of known target sites and their resulting PSSM model, a score can be calculated for each of these sites based on the model. Then, a score threshold can be defined as the score that would detect the top x percent of the known targets (x is a user-defined cutoff, e.g., 90%, 95% or 100%). This cutoff will correspond to the expected *false negative rate*.

The second method calculates a score according to a user-defined expected *false positive rate* or *false discovery rate* (*FDR*). In order to do so, we first need a model for the background DNA. For the purposes of this test, *background DNA sequence* is a sequence that is not expected to contain binding sites for the TF of interest. On a first approximation, the background can be modeled by the single nucleotide or higher order frequencies of the genome (di-nucleotides, tri-nucleotides, etc.). A number of random background sequences can be generated based on these frequencies. Alternatively, as background, we can consider the coding nucleotide regions or the intronic regions, since these are not expected to contain TF binding sites (although some times the introns –especially the first intron- may contain TF binding sites). The background sequences will then be scanned and each subsequence will be scored according to the PSSM model. Since the background is not expected to have *any* binding sites, all predictions can be considered "false positives". The score of the top xth percentile subsequence is the threshold with x percent FDR. For example, if we have 10 background sequences, each of length 1,000 bp, then the PSSM score of the 10th top scoring subsequence (20th if we scan both strands) will constitute the threshold for FDR = 1%. *MatInspector*, a popular scanning program, uses FDR $= 3\times10^{-4}$ as its default value. We note that the FDR is sometimes referred to as the *p-value* by some researchers.

These two methods for determining the threshold are based on different expectations (false negative *vs*. false positive predictions). So, it will be useful to compare the two. Fig. 18.3 presents graphically the performance of these two methods for four TFs with good PSSM model site representation. For the background, we have used the first order Markov model of 1,000 randomly selected genomic pieces, 5,000 bases long.

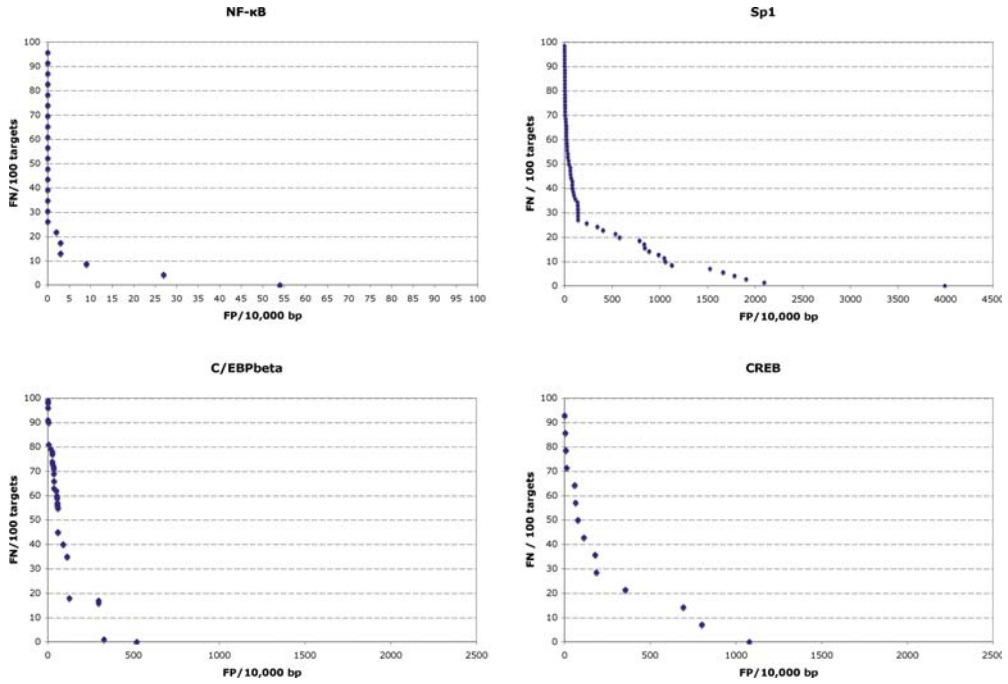

Fig. 18.3 False positive (FP) and false negative (FN) rate comparison for various TFs. The FP to FN rate is very different for the four well-studied mammalian TFs. These plots illustrate the difficulty one has to establish an objective PSSM threshold for predicting new binding sites (Copies of figures including color copies, where applicable, are available in the accompanying CD)

18.5 The Light of Evolution

In order to overcome the problem of the low signal-to-noise ratio on pattern matching algorithms, evolutionary information can be taken into account. The idea is simple–"true" (i.e., biologically relevant) TF binding sites should be conserved between related species. Of course, sites can be lost or replaced and new sites can be generated and the mechanism of this process is still unclear. It has been shown, however, that the functional sites tend to congregate in evolutionary conserved regions. In a recent study, we analyzed the promoters of 513 human genes in 8 other vertebrate species. We compared the promoter percent conservation with the number of known sites found in conserved regions. Table 18.2 shows the results of this analysis (all species compared to human, the

Table 18.2 Promoter and TF binding site conservation

Human vs.	No. ortholog. genes	Detectable sites	Promoter conservation	*Sites in conserved regions*
Chimp	512	1157	94.06%	94.81%
Mouse	506	1146	24.20%	72.34%
Opossum	389	912	6.72%	41.23%
Chicken	189	451	3.21%	21.73%
Tetraodon	166	363	2.50%	12.12%

Average percent conservation of the 5 kb regions upstream of the transcription start site of 1,162 protein coding genes in six vertebrate species. Percent of 513 known human TF binding sites located in the corresponding conserved regions. Conserved regions are defined as a window of at least 50 bp in length with >65% nucleotide identity. All species conservation is measured with respect to the human genes. Data from [17]
(Copies of tables are available in the accompanying CD.)

reference genome in this study). We see that although both promoter sequences and site conservation rates decrease with the evolutionary distance, the sequence conservation does so much faster.

This idea of evolutionary conservation has been explored by various pattern-matching algorithms, in order to reduce the large number of false positive predictions that are almost inevitable in such searches. This concept is known as *phylogenetic footprinting*. *rVista* [18], a popular phylogenetic footprinting method, scans one of two homologous promoters with PSSM models of known TFs and then evaluates the putative sites based on the degree of conservation of the site themselves and the interval in which they are located between the two species under comparison. Sites should be "globally aligned" in the homologous promoters in order to be reported, meaning, they should be located within a specific window length (typically: 21 bp). *ConSite* [19] is another phylogenetic footprinting algorithm. Unlike *rVista*, *ConSite* scans *both* promoter regions and reports those predicted sites that are located in equivalent positions in the conserved regions of the two homologous promoters. Like in *rVista*, conserved intervals are also calculated using a sliding window approach. *FOOTER* [20] is the newest of these methods. Unlike the other two, *FOOTER* uses both the location and the PSSM score to statistically evaluate the potential of a pair of sites to be functional. Also, predicted sites on both the conserved and non-conserved regions are examined.

18.6 *Cis*-Regulatory Modules

Sydney Brenner once said: "complex organisms evolve from simpler ones not by constantly inventing new genes, but by fine-tuning the regulation of existing ones".[1] With the maximum number of genes set to a moderate number of less than 25,000, complex eukaryotic organisms, such as flies and mammals, are expected to have developed complicated mechanisms for gene regulation. This has been shown to be true in many cases. Multiple TFs are co-operating or competing in order for the gene to be properly expressed under a given condition. At the DNA level, the binding sites of competing TFs are frequently overlapping or located in close proximity on the promoter. In this way, when one of the proteins is bound to its target, it inhibits the binding of its competitor. Similarly, the binding sites of co-operating TFs are expected to be located close to each other, so that the proteins can facilitate binding through cross-talk *via* protein-protein interactions on a given promoter. Sometimes, proteins that bind DNA targets, thousands of bases from each other, can "communicate" *via DNA looping* (see Fig. 18.4).

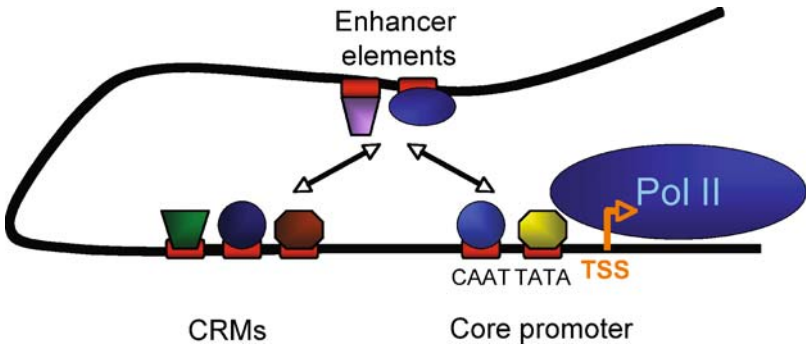

Fig. 18.4 Schematic view of a typical eukaryotic promoter. The elements of the general transcription factors as well as examples of gene-specific binding sites (CRMs) and enhancer elements are presented. Arrows indicate possible communication between the enhancer binding proteins and the CRMs on the core transcription complex. The interactions can be direct or *via* other proteins (Copies of figures including color copies, where applicable, are available in the accompanying CD)

[1] From Sydney Brenner's seminar "*Return of the human genome*", Washington University in St. Louis, St. Louis, Missouri, 2000.

A number of algorithms have been developed to identify co-occurrences of TF binding sites, also known as *cis-regulatory modules* or *CRM*s. Some of these algorithms incorporate a *de novo* motif finding method with CRM identification. *Co-Bind* [21], for example, applies a Gibbs sampling method to identify concurrently *two* motifs located within a window of a user-defined length. Other methods attempt to assign significance on the observed motif co-occurrences. The significance can be assigned by various methods, including *p*-value calculation (e.g., method *MSCAN*), and Monte Carlo simulations (e.g., method *SCORE*).

Despite the success of such algorithms in identifying some of the known CRMs, it is still unclear what constitutes a CRM or how general their observed properties may be. For example, in some cases the order and spacing of the individual TF binding sites is important. This is the case of the well-known SRY module in the promoters of the *major histocompatibility complex* (*MHC*) genes. In this case, the distance between the individual *cis*-regulatory elements as well as the distance of the whole SRY module from the transcription start site of the downstream gene is important. In other cases, the order and/or the spacing of the individual sites as well as the exact location of the CRM does not seem to be very important. This is the case of many enhancers that have been found to work equally well from tenths of thousands of bases away from the transcription start sites and when placed near it. This variability in the CRM properties makes the design of efficient algorithms for CRM detection more difficult.

18.7 On-Line Resources

Although links to on-line resource pages tend to quickly become outdated, we mention here some that we believe will be useful to the readers.

DNA Motif and cis*-regulatory site databases*

- JASPAR (open access database): http://jaspar.genereg.net/
- PLACE (plant): http://www.dna.affrc.go.jp/htdocs/PLACE/
- TRANSFAC (public version): http://www.gene-regulation.com/pub/databases.html

Promoter retrieval

- EPD: http://www.epd.isb-sib.ch/
- dbTSS: http://dbtss.hgc.jp/index.html
- SCPD (yeast): http://rulai.cshl.edu/SCPD/
- TRED promoter database: http://rulai.cshl.edu/TRED/

Motif finders

- AlignACE: http://atlas.med.harvard.edu/
- ANN-SPEC: http://www.cbs.dtu.dk/services/DNAarray/ann-spec.php
- BioProspector: http://robotics.stanford.edu/~xsliu/BioProspector/
- Consensus: http://bifrost.wustl.edu/consensus/html/Html/interface.html
- Gibbs motif sampler: http://bayesweb.wadsworth.org/gibbs/gibbs.html
- MEME: http://meme.nbcr.net/meme/intro.html
- MotifSampler: http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html

Phylogenetic footprinting

- ConSite: http://www.phylofoot.org/consite
- FOOTER: http://www.benoslab.pitt.edu/Footer
- rVista: http://rvista.dcode.org/

Acknowledgments The author would like to thank Patrick Shea, who kindly provided the data for Fig. 18.3. This work was supported by NSF grant MCB0316255, NIH grants RR014214 and NO1 AI-50018 and by a tobacco settlement grant from the Pennsylvania Department of Health. PVB was also supported by NIH grant 1R01LM007994-01, TATRC/DoD USAMRAA Prime Award W81XWH-05-2-0066 and by intramural funds from the Department of Computational Biology, University of Pittsburgh and the University of Pittsburgh Cancer Institute (UPCI).

Glossary and Abbreviations

PSSM	Position-Specific Scoring Matrix
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TSS	Transcription Start Site

References

Protein-DNA interactions

1. Benos PV, Lapedes AS, Stormo GD: Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol* 2002, **323**(4):701–727.
2. Benos PV, Bulyk ML, Stormo GD: Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* 2002, **30**(20):4442–4451.
3. Barash Y, Elidan G, Friedman N, Kaplan T: Modeling Dependencies in Protein-DNA Binding Sites. In: *Seventh Annual International Conference on Computational Molecular Biology (RECOMB): 2003*; 2003.
4. Zhou Q, Liu JS: Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 2004, **20**(6):909–916.

DNA pattern representation

5. Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, **16**(1):16–23.
6. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, **18**(20):6097–6100.
7. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986, **188**(3):415–431.

De novo motif finding

8. Hertz GZ, Hartzell GW, 3rd, Stormo GD: Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* 1990, **6**(2):81–92.
9. Hertz GZ, Stormo GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999, **15**(7–8):563–577.
10. Lawrence CE, Reilly AA: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* 1990, **7**(1):41–51.
11. Bailey TL, Baker ME, Elkan CP: An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. *J Steroid Biochem Mol Biol* 1997, **62**(1):29–44.

12. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, **262**(5131):208–214.
13. Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127–138.
14. Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000, **296**(5):1205–1214.
15. Workman CT, Stormo GD: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000:467–478.
16. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: Transcription factor binding site identification using the self-organizing map. *Bioinformatics* 2005, **21**(9):1807–1814.
17. Mahony, Corcoran, Benos (2007) Genome Biol **8**:R84.

Phylogenetic footprinting in motif detection

18. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* 2002, **12**(5):832–839.
19. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003, **2**(2):13.
20. Corcoran DL, Feingold E, Dominick J, Wright M, Harnaha J, Trucco M, Giannoukakis N, Benos PV: Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. *Genome Res* 2005, **15**(6):840–847.
21. Guha Thakurta D, Stormo GD: Identifying target sites for cooperatively binding factors. *Bioinformatics* 2001, **17**(7):608–621.

Key References

Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, **16**(1):16–23.

Bailey TL, Baker ME, Elkan CP: An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. *J Steroid Biochem Mol Biol* 1997, **62**(1):29–44.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, **262**(5131):208–214.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, **23**(1):137–144.