

Machine Learning Approaches in Promoter Sequence Analysis

N.T. Tung ⁽¹⁾, E. Yang ⁽²⁾, I.P. Androulakis ^(2,*)

⁽¹⁾ BIOMAPS Institute for Quantitative Biology, Rutgers University

⁽²⁾ Department of Biomedical Engineering, Rutgers University

^(*) corresponding author: yannis@rci.rutgers.edu

Abstract

Gene transcription is one of the main biological processes that govern an organism's response to external stimuli. Understanding the mechanism of gene regulation offers an avenue with which to model this response. It has been hypothesized that one of the primary mechanisms for gene regulation is via transcription factor binding in which a protein (transcription factor) binds to certain sequences in the genome. Computationally, researchers hope to identify both the promoter region as well as the sequence motifs which are bound by transcription factors via analysis of genomic sequences. Machine learning methods make the hypothesis that these sequences are drawn from some underlying but unknown patterns of base-pairs, and are attractive due to their ability to process the large amounts of sequence and experimental data present. The ability to isolate these sequences is predicated upon the large amount of binding and expression data from experimental protocols such as Chip-Chip, SELEX, microarray, and GRIP. We review the basic background of promoter structure and then focus on three popular aspects of promoter studies: promoter prediction, promoter analysis, and promoter modeling. The paper will cover widely used computation techniques such as Gibbs sampling, MEME, phylogenetic foot-printing, and position weight matrices in promoter analysis, motif alignment and promoter modules in promoter modeling. Finally, this manuscript will describe the use of

these algorithms in an integrative sense to isolate regulatory modules to rationalize the results of experimental data.

Introduction

Since the discovery of the structure of DNA [7, 8], biology has advanced from a descriptive science to a more quantitative field. The central dogma of molecular biology laid the foundations of modern biology by characterizing the importance of DNA transcription in cellular function. One of the most influential aspects of gene regulation is the structure of the genomic region located upstream of the coding region, known as the 'promoter', as the binding of regulatory proteins in this domain critically affects the transcription to mRNA, the initial step in gene expression that controls effectively all biological processes such as development, proliferation, apoptosis, aging, and differentiation [9]. Consequently, a small change in the process of regulation have significant implications in the cell fate [10].

The completion of the human genome sequencing in 2003 [11] and the numerous related genome projects, high-throughput technologies such as ChIP-chip [12, 13], SELEX [14, 15] make promoter-related studies feasible. With the aim of deciphering the regulatory mechanism of the cell in response to diverse stimuli, the interplay between trans-factors and cis-regulatory elements has become an active area of research [16]. Although experimental techniques can help discover such crucial cis-regulatory elements and their implication in the regulation of gene expression, they can only partially address the complexities associated with gene regulation and transcription. Among the main difficulties is the fact that high-throughput experimental techniques produce enormous amounts of data, making it difficult to analyze with traditional approaches. As a result, computational algorithms, and especially machine learning techniques, have become an essential tool in accelerating the analyses of these experimental studies.

To provide a basic platform of computational methods relevant to promoter studies within the content of sequence analysis, in this brief review we begin by introducing the basic elements of promoter structure and then discuss three main aspects of computational promoter sequence

analysis in the current literature with basic concepts and relevant algorithms. We also discuss available tools and their relevance in current research.

Basic elements of promoter structure

Promoters are DNA sequences located upstream the coding region of each gene towards the 5' endpoint. Combined with other regulatory elements in the upstream region of a gene, these elements in the promoter region interact with transcription factors, recruit RNA polymerases, and then initiate the transcription of a gene.

There are three classes of promoters that are recognized by three corresponding RNA polymerases (Figure 1):

- Class I promoters are made up of two regions, an upstream control element (UCE) and a core promoter. They serve for the regulation of ribosomal RNAs synthesis(5.8S, 18S, and 28S rRNAs).
- Class II promoters are mainly involved in transcribing protein-coding genes which generate pre-mRNAs and almost all small nuclear RNAs (snRNAs). Each member of this class consists of a core promoter, proximal promoter elements and distal regulatory elements.
- Class III promoters have three types: type I and II are internal promoters that regulate the synthesis of 5S rRNAs and tRNAs and interact with sites in the RNA polymerase. Type III promoters are upstream promoters similar to class II promoters and regulates the synthesis of some snRNAs or viral-associated RNAs [17].

Figure 1

Although the process of gene expression is regulated at many levels e.g. genomic level, transcriptional level, RNA processing level, translational level, or post-translation level, promoter regions and regulatory elements are still considered as one of the most important factors [10]. Since proteins in eukaryotes are mostly transcribed by RNA polymerase II, computational promoter studies are mainly focused on protein-coding genes, in this review we will concentrate on the structure of class II promoters (Figure 2a) which are characterized by the core promoter, proximal- and distal- promoter elements [9].

Core promoter is a small stretch sequence about 100bp flanking the transcription start site (TSS) which incorporate a combination of four common components consisting of the TATA box, initiator (Inr), TFIIB recognition element (BRE), and the downstream promoter element (DPE) [18, 19]. This serves for the initiation of the transcription process (Figure 2b). The TATA box, the binding site for TATA-binding protein (TBP), is a TA-rich site at 26-31bp upstream in higher eukaryotes and 40-120bp upstream in yeast [20]. Inr, also called the Transcriptional Start Site (TSS), is the start position located in the core promoter and functions similarly to the TATA box [19]. A comprehensive statistical analysis on a dataset with more than 10,000 human promoter from EPD [21, 22] and DBTSS [23] demonstrated that it is not necessary for all these components to be simultaneously present in the core promoter [24]. Specifically, Inr elements are present in nearly half of the promoters whereas TATA boxes are present in only around 10% of the promoters in the dataset and seem to simultaneously present with the Inr elements. BRE and DPE elements are present about 25% of the time. Furthermore, the presence of DPE is independent of the presence of TATA-box and Inr elements whereas BRE-containing promoters are present in TATA-less promoters. Besides these elements, a number of other motifs in this region e.g. YY1, CAAT, CREB, etc. were also discovered in an analysis on a set of high-quality human core promoters [25], These features became important criteria to scan for approximately promoter regions in promoter prediction (section 3).

Proximal promoter elements are located on the proximal promoter which is defined as the region up to 1Kbp upstream of the core promoter. The presence and importance of these cis-regulatory elements were characterized via a technique called linker-scanning mutagenesis [26] which showed that any mutation at one site in a regulatory element in this region can cause a significant change in transcription levels. Elements in the region between -350 and -40 have the positive effect on the promoter activity whereas those in the region from -350 up to -1000 appear to have a negative regulation on the expression of the gene [10].

Besides cis-regulatory elements, another feature of the proximal promoter region is the appearance of CpG islands which are short stretches of unmethylated DNAs (~200bp) with GC percentage higher than 0.5 (i.e. $p_C + p_G > 0.5$) and an observed/expected CpG ratio greater than 0.6 (i.e. $p_{CpG} > 0.6 \times p_C \times p_G$) [27]. They are frequently located in or near the promoter regions with more than 40% promoters of mammalian genes (about 70% in human promoters) [28] and in fact, their presence implies the existence of the promoter region [29].

Distal regulatory elements are characterized by four regulatory groups (Figure 2c). Enhancers work as cis-regulatory elements near the TSS with the positive effects on promoter activity and in many cases, they both share the same activators [30]. Silencers are bound by repressors to negatively regulate the expression. The third group is insulators which are similar to a wall, preventing the mutual transcriptional effects of regulatory elements between neighbor genes. The last is a combination of different regulatory elements (known as locus control regions (LCRs) which regulate an entire locus or a number of genes [31]. These trans-regulatory elements function in the same way as cis-regulatory elements although they are located far from the TSS and work under the control of trans-acting factors[9].

Figure 2

Computationally, it is hypothesized that the regulatory elements are subsequences in a DNA sequence that exhibit significant conservation. The hypothesis is that functional regions are conserved during evolution compared with non-functional regions even though they are located on non-coding sequences [32, 33] due to the importance they have upon an organism's evolutionary fitness. With the large amounts of current experimental data, machine learning techniques have become an indispensable tool to exploit these conserved regions for understanding the underlying mechanism of regulatory processes. A list of databases relevant to promoter studies is displayed in Table 1.

Table 1

General Problem Definition

Given a dataset $D = \{X_i, Y_i\}_{i=1}^N$, $X_i = \{x_{i1}, x_{i2} \dots x_{id}\}$ where $X = \{X_i\}$ is a set of objects and x_{id} is a numeric value for the corresponding attribute d of object X_i ; $Y = \{Y_i\}$ is a corresponding set of labels associated with each object in the data set. If Y is not known, the problem becomes one of unsupervised learning. Otherwise, if Y is known, either as a binary value {TRUE, FALSE}, discrete attribute or continuous value, the problem becomes a supervised learning problem. Of particular importance to this work are binary classification problems which samples belong to discrete categories. This is because multi-class problems can be reduced to binary classification problem in an all-against-one or one-against-one framework. In the context of promoter analysis, binary problems are relevant since the usual question is to decide whether a segment of DNA sequence is a promoter or whether a given promoter defines or not a regulatory element.

A common framework of current aspects in promoter sequence analysis is presented in Figure 3. Promoter prediction is a process of learning a mapping $f : X \rightarrow Y$ to decide whether a

subsequence DNA is a promoter. X now becomes a set of sequences with each attribute corresponding to a base in the sequence, a K-base pair window extracted from the sequence, or a numeric value if the original form of the sequence is transformed. More generally, this is still a process of learning motifs on DNA sequences to differentiate between promoter regions and other regions on a DNA sequence in general. In promoter analysis, the main problem is to discover transcription factor binding sites (TFBSs) in the promoter sequence or find overrepresented motifs in a given set of sequences. Utilizing TFBS motif models built from experimental collections of TFBSs, TFBS candidates on a new sequence can be scanned. From those matches, promoter sequences now can be reorganized by a list of TFBSs as which it can be considered as a vector of promoter features used for successive analyses.

Figure 3

Promoter prediction

With the completion [11] and detailed analysis [10, 34, 35] of the human genome, the emphasis has now shifted towards annotating gene functionality and understanding the regulatory network behind. Promoter prediction has gradually become a common element of many gene prediction methods as well as a topic of research in its own right (Table 2). Given a subsequence, identifying whether it is a promoter or non-promoter became an important problem in gene discovery. Even when a large number of experimental methods e.g. oligo-capping [36], CAP-trapping [37], expressed sequence tags, cDNA/mRNA are applied, the TSS location cannot be easily determined [38]. Therefore, an alternative is to use computational techniques to support some part the process, reduce the wet-lab labor, and guide back experimental biologists. The problem can be summarized into three systematic steps as follows.

Problem definition

(1a) Given a DNA sequence $s = \{s_l\}_{l=1}^L$, $s_l \in A$, $A = \{A, C, G, T\}$, extract a feature vector $x = \{x_1, x_2, \dots, x_d\}$, $d \in N$, $x_i \in R$ from s .

(1b) Given a dataset $D = \{X_i, Y_i\}_{i=1}^N$, $X_i = \{x_{i1}, x_{i2}, \dots, x_{id}\}$, $Y_i = \begin{cases} +1 & \text{if promoter} \\ -1 & \text{if not} \end{cases}$, $x_{id} \in R, d \in N$,

learn a model M to predict whether a DNA subsequence s that is characterized by feature vector as X_i is a promoter or not.

(1c) Given a promoter prediction model M ; for a new long sequence S (a new chromosome or a new genome sequence), scan all possible promoter regions s .

Experimental biology may need the support of promoter prediction programs (PPPs) for two critical tasks: (i) search for TSSs or alternative TSSs in a sequence and (ii) search for unknown genes in targeted chromosomal segments or in the throughout the chromosome [39]. However, the problem turns to be more complicated since current data suggest that a gene might have multiple promoters, each of which can have different TSSs instead of the traditional idea that one gene has one promoter with one corresponding TSS. Furthermore, in mammalian genomes, transcription can initiate at unusual points around the gene regions, e.g. intergenic regions far from known genes, 3' UTR, coding exons or introns [40]. What makes the process more difficult is the lack of a ground truth from which to evaluate the results of TSS prediction. Consequently, an assessment of the performance of these algorithms was specifically designed to address how good PPPs are when used in genome-scale studies under a common test set with the same method to evaluate [39].

Datasets

DBTSS [23, 41] and EPD[22, 42] are the largest and main sources to provide experimentally determined TSSs of human genes and their corresponding promoter sequences. To get the data for non-promoter sequences, exons/introns and 3' UTR sequences are collected from GenBank and UTRdb [43]. There are two types of PPPs; one tries to predict as accurately as possible the actual position of the TSS for a gene; the other attempts to find approximately the promoter regions near the TSSs. However, the data can be generalized into the same format for learning and testing but the extracted features depend on the strategy of each PPP.

Each promoter sequence of length L contains a TSS, n bp upstream and m bp downstream of the TSS ($L = n + m$) (Figure 4a). There are two ways to evaluate the model depending on whether we study question 1b or 1c. A local test will evaluate the model on the test set which is a part of the extracted dataset and used for problem 1b; if an actual promoter sequence was predicted as a promoter sequence or a sequence containing a TSS, it will be considered as a true positive and so on; the dataset has a clear separation of promoter set and non-promoter set. A global test will evaluate the model on the real genome or chromosomes (problem 1c). At this point, when a promoter region is predicted, the corresponding TSS is inferred; let d be the maximum allowed distance from a real TSS to be considered as a correct hit; for a real TSS, if there is a predicted TSS falls into region $[-d, +d]$ with the reference point at the real TSS location, it will be counted as a TP; otherwise, if there is no predicted TSS in this region, that will be considered as a FN; in the segment $[+d+1, \text{GeneEnd}]$, if there is no predicted TSS in this region, it will be counted as a TN and if not, it will be counted as a FP. In the case of two or more genes overlapping together, refer to Figure 4b [38].

Figure 4

As indicated in the early review of [44], eukaryotic promoter prediction is not an easy task and various algorithms have been proposed for this task. Different programs have developed

different strategies to extract the features of this region but in general, they can be classified into three categories: (1) search by signal, (2) search by CpG island, and [45] search by content [46]. Signals are known as motifs or statistically significant motifs that frequently appear in the promoter region e.g. TATA-box, CAAT-box, etc. (Figure 2). Although they can be used to differentiate between promoter sequences and non-promoter sequences, they are insufficient to get high accuracy [24, 39] by themselves. These are usually incorporated in the predictive model utilizing other features. The second category attempts to identify CpG-related promoter sequences which have different properties from non-CpG related promoters. Search by content utilizes a k-size window slides on the input sequence to extract continuous k-size words as features. In this method the input sequence can be transformed into some other form in some works. This becomes a model that learns short motifs to differentiate promoter sequences and non-promoter sequences e.g. exons, introns, 3' UTR (Table 2). Figure 5 shows a general framework of how a PPP works in predicting whether a DNA subsequence is a promoter or not.

Figure 5

When the data is identified and the features extracted, different techniques present significant difference in performance. The techniques can be quite varied such as artificial neural network (ANN) [47-53], linear or quadratic discrimination function [54, 55], relevance vector machine [56], interpolated Markov model [57-60], support vector machine (SVM) with string kernel [61], relative entropy [46] or even using only rules based on statistics [29, 62]. But the main concept is still that statistically significant motifs of the promoter regions which are exploited by a k-size window sliding on the promoter sequence to extract features for learning models e.g. ANN, SVM (more details in Table 2). [61, 63] represent techniques that utilize these k-size windows in various machine learning algorithms for the extraction of promoter regions (more details in Table 2).

Table 2

However, not all of the methods for promoter identification need to be based upon DNA motifs alone. [64] analyzed the structure of the core promoter in mammalian and plant genomes by a number of physicochemical properties such as DNA bending, denaturation, propeller twist, duplex disrupt energy, A-philicity, etc. and then [64], [65] utilized these to transform the input sequences before extracting features for training the models.

Promoter analysis

Promoter analysis is another very important aspect in promoter sequence analysis aiming towards understanding the mechanisms that drive gene expression and gene regulation. While promoter prediction deals with finding approximately, the promoter regions or TSSs of the genes, promoter analysis assumes that promoter sequences of genes are already known and are thus treated as prior information. It mainly focuses on identifying cis-regulatory elements or transcription factor binding sites (TFBSs). Promoters are one of the keys to understanding the underlying hypotheses of gene expression and regulation which is mainly controlled by transcription factors (TFs), i.e., proteins that bind to promoter regions at specific sites (TFBSs) and regulate the process of transcription initiation.

Since a promoter is considered to be a string of characters A, C, G, and T, features of promoters are almost considered to be conserved patterns in this region compared with other regions on the genomic sequence. Since functional elements are usually conserved, these motifs, considered to be the TFBS, are usually the sites where TFs bind to regulates the transcription process [32, 33]. Therefore, computational techniques look at a TFBS as a sequence motif or a conserved subsequence on the promoter sequence. Given a promoter sequence, one would be then interested in finding all TFBSs and if given a list of promoter

sequences, the problem becomes how to find common TFBSs in those sequences. These approaches basically combine the two perspectives of promoter analysis, i.e. discovery by experimental data and pure computation [66, 67]. The former depends on how the motif is modeled based experimental data and how the promoter sequence is scanned to identify possible matches whereas the latter relies on how the motif is modeled based on a given set of sequences.

With the concept of string pattern discovery, a multitude of methods have been proposed for the second problem in recent years yielding a variety of algorithmic approaches, underlying models, and testing methodologies. Models can be pattern or sequence-driven, whereas patterns can be deterministic or statistical [68]. However recently, the advances in the field have focused towards modeling regulatory modules rather than working with individual motif [69]. Thus an integrated framework working at multiple levels deals with TFBSs starting from single-motif, composite-motif, gene-level, to genome-level models is introduced [69]. Single-motif is the level of handling individual motifs whereas composite-motif works on clusters of TFBSs (also called cis-regulatory modules). At the gene-level the focus is on how several modules act together to regulate a single gene and finally at the genome-level, how several sets of modules work on sets of to control expression. In this review however, we will consider the single-motif level because it functions as a basic kernel for the other methods.

An underlying hypothesis of these computational predictions is that the potential for coregulation between two genes is effectively proportional to the probability that a large number of TF would bind to the promoter region of these genes. Therefore, an important question is how to determine all possible TFs that bind a given promoter region. However, it must be emphasized that the possibility of binding should not be directly translated to coregulation, since binding does imply functions. It defines however a good starting point for further analyses.

Problem definition

We define the following computational problems capturing the aforementioned questions:

(2a) Given a DNA sequence $s = \{s_l\}_{l=1}^L$, $s_l \in A$, $A = \{A, C, G, T\}$ and a motif profile $M = \{m_{ij}\}$, $i = 1..|A|$, $j = 1..m$, $m_{ij} \in [0,1]$ where m is the length of the profile, scan s for all possible matches with the profile.

(2b) Given a set of DNA sequences $S = \{s_i\}_{i=1}^N$, $s_i = \{s_{il}\}_{l=1}^{|s_i|}$, $s_{il} \in A$, search for conserved motifs (patterns) $p = \{p_k\}_{k=1}^K$, $p_k \in A$ that are over-represented in S .

These two general problems are relevant to how a motif is modeled. However, the former is more popular among experimental biologists since they would like to know which cis-regulatory elements exist in a newly determined promoter sequence by scanning for known motif profiles available from databases and tools such as TRANSFAC [70, 71] or Genomatix [72]. The second is more relevant to the computational area and as such a wide range of computational techniques from pattern recognition have been employed over the years [73]. There are two popular methods in biological sequence analysis for motif modeling e.g. position weight matrix (PWM) and hidden Markov model (HMM) [74, 75] but in promoter sequence analysis, PWM is the best-known method since it is simple and effective.

Position weight matrix

PWM (also called position specific scoring matrix or position specific frequency matrix) is a matrix of (log normalized probability) scores with four rows corresponding to four DNA bases and m columns, each of which is a position in the motif. PWM assumes the independence between positions in the motif; thus the fitness score of an oligo with this profile is only the sum of the fitness at each position (Figure 6). The consensus sequence is a preliminary

representation for a motif which is formed by a list of DNA oligos; the consensus character is the highest frequency base in that column but when there is an approximately equivalent frequency between bases in that column, a corresponding IUPAC character (Table 3) is assigned. To construct the motif representation more effectively, a profile M is used by aligning TFBS instances and describing the alignment with the frequency of each DNA base in each column.

Figure 6

Table 3

Based on the concept of PWM, various measures are defined to evaluate an alignment or more importantly to estimate the probability an oligo p fits with that motif. The information content (IC) of a PWM is the most widely used measure to consider how different a given PWM is from a uniform distribution and also to quantify how much the corresponding are similar [76]. The IC of

a PWM with profile M is $IC(M) = \sum_{i=1}^4 \sum_{j=1}^m m_{ij} \log \frac{m_{ij}}{b_i}$ where m_{ij} is the frequency of base i in

position j and b_i is the expected frequency of this base in the background set of DNA sequences that are non-motifs or random sequences. To make a comparison between two alignments with the same number of oligos, the IC can be used directly. When the number of oligos is different,

a maximum a posteriori (MAP) measure is defined $MAP(M) = -nIC(M)$ where n is the number of oligos used to build the profile M. Another important measure is the probability an oligo p belongs to this motif or is a part of the background noise

$$L(p) = \frac{\Pr(p | M)}{\Pr(p | background)} = \prod_{i=1..4, j=1..l} \frac{m_{ij}}{b_i}. \text{ This quantity then becomes a scoring function for}$$

most of the discussion that follows.

Scanning for TFBSs

With the support of high-throughput techniques (e.g. ChiP-chip, SELEX), experimental data are now likely sufficient to support the results of the computational methods. The TF that have been identified experimentally have been associated with a list of binding sites and therefore corresponding motif profiles can be constructed. There are around seven hundreds PWMs for human TFs and more than six thousand TF profiles in total listed in TRANSFAC [70, 71], making it possible to scan for a significant number of potential TFBSs in a new promoter sequence. Every available TF profile M for its corresponding organism is compared with all K -size words of the promoter sequence, if the fitness measure $L(p)$ of some word p is over a threshold, it can be considered as a match or a TFBS in this problem [77-79]. Multi-TFBSs for the same TF can appear in one promoter sequence and likewise there may also be no significant matches for any TF in a promoter region.

However, the TFBS is usually degenerate and short (8-15bp) sequence since the motif only has four states (A, C, G, and T). As a result the binding sites can occur very frequently by chance, leading to a high level of false-positive predictions. Therefore a number of approaches have been identified in order to optimize motif content e.g. Markov chain optimization [80], mixture models [81], or the Staden-Bucher approach [82] while others rearrange the motif model to get better performance e.g. doing multiple local alignment on oligos of the motifs before extracting the motif profile [83], dividing TFs into TF families and build a hierarchical tree for efficient searching [84].

Searching for conserved motifs

However, while the use of PWM allows researchers to identify which transcription factors interact with which genes, the identification of position weight matrices from a set of sequence data remains. DNA motif discovery relies upon multiple sequence alignment (MSA) from global-to local-alignment. Given a set of DNA sequences and a substitution matrix, various

computational techniques were developed to align these sequences e.g. clustalW [85], DIALIGN [86], T-COFFEE [87], ProbCons [88] and then extract conserved motifs. However, the fact that TFBS are arbitrarily located leads to a new and challenging problem from both a computational and biological point of view. Fortunately, a number of techniques from pattern recognition have been applied and proven their effectiveness. The core algorithms, in general, can be classified into two categories: combinatorial and probabilistic (Table 4). A couple of typical algorithms will be presented in the following with emphasis on two important aspects: the search method and the scoring function.

Table 4

The combinatorial category is the starting point of discovering TFBSs or conserved motifs on a set of promoter sequences. It is an exhaustive search with pattern-based scoring for all possible cases of the motifs or of the given set of sequences, and then the search is refined gradually by the greedy heuristic and probabilistic concepts. The first type of exhaustive searches is pattern-driven algorithms (Figure 7) which assume that the motif has length K and each position in the motif has four states (A, C, G, and T). The algorithm then makes a complete search for all 4^K possible patterns to pick out some that satisfy the selected criteria. A score is estimated to indicate the extent to which a sequence contains a pattern and then generalizes the extent to which that pattern belongs to the conserved motif of the given set. The primary drawback with the current algorithm is that the complexity is very high ($O(N 4^K)$), making it impossible to search for long motifs (e.g. more than 10). By changing the collection of patterns needed to search i.e. using K -size patterns instead of 4^K patterns, a significant improvement can be achieved, leading to the sequence-driven (also called sample-driven) algorithm [76]. The exhaustive search can be accelerated by using the suffix tree [89] or some index way but the searching space is still large.

To further reduce the searching space of the algorithm and improve the scoring efficiency, the consensus method was proposed [45]. It is also a type of pattern-based scoring but explores the alignment concept instead of using Hamming distance as exhaustive searches do. The searching space is reduced by applying a greedy heuristic in the area of multiple sequence alignment leading to the agglomerative conserved profiles and greedy selection [45]. At first, all K-size words of the first sequence are treated as a collection of selected alignments; then for each sequence every of its K-size words is aligned with all selected alignments and put into the new alignment collection. These new alignments are scored and compared with a threshold to select the matches. The algorithm is significantly affected by the sequence order and the way alignments are scored. WConsensus [90] is a typical example of the approach.

Figure 7

Probabilistic methods define a type of profile-based scoring to optimize the motif profiles. MEME (Multiple Expectation maximization for Motif Elicitation) [91] and Gibbs sampling [92] are two typical techniques (Figure 8). With the assumption that every sequence has only one oligo for one motif, they start with a randomly selected motif model characterized by a profile of a list of oligos as mentioned above. MEME classifies the remainder of each sequence (i.e. taking away the oligo of that sequence in the motif model) as the background to create the background model. Based on the motif and the background model, all K-size words of the given set are scored and then they are reclassified as either the motif or the background model. The goal of this technique is to maximize the likelihood function of all K-size words with respect to the motif or the background model. The algorithm runs until the likelihood function converges and the motif model is reported. To obtain a different motif, the algorithm is restarted. Gibbs sampling is another approach similar to MEME, however, MEME updates the entire motif profile for every cycle whereas Gibbs only updates a single oligo. This technique considers one randomly

selected sequence at each cycle and scores the profile based on its own information content. All K-size words of the selected sequence are scored over the motif profile; the best fit word will be chosen as a new oligo to replace the current oligo of that sequence in the motif model for the next cycle. The algorithm runs until the score of the motif profile converges.

Figure 8

Enhancing TFBS discovery

In addition to enhancing the accuracy of locating TFBSs, additional information from expression data and orthologous species is utilized defining the so-called multiple data integration [93]. Combining gene expression data and promoter analysis allows one to identify and predict the TFBSs more accurately [94-96] in a manner to the way using information from orthologous genes can refine the set of potential TFBSs [97-99]. Other method focus on how to extract efficiently structured motifs EXMOTIF [100] and how to search utilizing those structured motifs – SMOTIF [101]. Especially, the concept of using ensemble of methods to improve predictions also appears to be promising [102] and a number of implementations have shown improved accuracy BEAM [103], PRISM [104], and SPACER [105].

However, in principle the common problem is still searching for conserved motifs as was discussed above but with different types of input data, we generate alternative interpretation of the output data. Cross-species comparison is a popular strategy to limit the potential regulatory regions and for improving the accuracy of TFBS discovery. In terms of a promoter of a gene, the promoter sequences of orthologous species are obtained and set up a set of DNA sequences for which motif discovery techniques work. The discovered motifs are considered as potential regulatory regions and have high probability to be bound by TFs. This type of analysis reduces significantly the false-positive rates in scanning TFBSs by using PWM, so-called comparative sequence analysis (Figure 9). Besides that, one can look for conserved motifs over a number of

promoter sequences from a set of coexpressed or coregulated genes; although coregulation and coexpression are two different concepts, at present coexpressed genes under a number of conditions are considered as being coregulated, a hypothesis not technically correct.

Figure 9

Promoter models

The most basic way also the most widely used method for representing a promoter is in the form of a sequence of bases or a string of characters (Table 3). When put into computational techniques, the promoter sequence is divided into a list of successive K-size words and then coded to form an input vector. This is so-called 'linear word-map' representation of a promoter sequence. To make the map more flexible, words are assigned a position but can be ordered for computational purposes. When the words are rearranged, the representation becomes a non-linear word-map of the promoter sequence. Many applications used this concept to measure the similarity of promoters but in that case only selected words are used e.g. TFBSs or conserved motifs instead of all extracted words. Blanco [106] defined a non-linear TF-map alignment where the selected words are TFBSs and utilized the concept of sequence alignment to estimate the similarity of two promoters and then extended to multiple TF-map alignment [107]. [108] did the same and used the Jaccard algorithm to estimate the similarity score between two promoter sequences for promoter clustering.

Since TFs does not come alone to bind with TFBSs and regulate the transcriptional process, we need to also consider how these single motifs combine together to regulate the gene expression – level two of promoter analysis. A cluster of TFBSs in a short sequence segment is defined as a cis-regulatory module (CRM); in brief, it is a set of motifs close together and conserved without order). Cis-regulatory modules are also usually conserved across species due to

evolutionary constraints and considered as the segment for which synergistic TFs bind to cooperate the initiation transcription [109]. A number of computational tools have been developed to search for CRMs e.g. CisModule [110], CREME [111], ModuleFinder [112]. Thus instead of finding single motifs, almost all applications move into working on set of motifs; this seems to be more powerful since CRMs are longer than TFBSs, leading to lower the false-positive rate.

In order to extend the range of motif modeling approaches beyond the single motif and extend the approach to the analysis of set of potentially co-regulated genes we need to move towards the concurrent analysis of sets of promoters associated with specific functions. Once a list of motifs is discovered with any of the techniques discussed above; they can all be organized in some way to form the model for promoters of this function. These promoter models could be used to possibly predict new genes associated with this function as suggested in [113] [114] [115]. However, much work is needed to improve the computational efficiency and reduce the excessive number of false positives associated with those methods.

Conclusions

Unraveling the mysteries and complexities of transcriptional regulation is of paramount importance in modern biology. What causes a stem cell to commit to a particular lineage, what makes a cell response to an external perturbation or an organism to a drug is largely determined by the transcriptional machinery that is the control mechanisms that dictate the up- or down-regulation of genes. In that context, the part of the non-coding regions of genes located upstream the transcription start site holds the keys to this mystery. The main theater of controlling transcriptional regulation is the region with transcription factors, proteins that control the transcription of genes, and such its identification, characterization and prediction is of paramount importance. In that respect computational methodologies, and most notably those based on fundamental principles of machine learning, have proven to be critical in deciphering

the complexities of gene regulation. In this short review we discussed a number of general computational issues in an effort to illustrate the implications of machine learning approaches in addressing a fundamental problem in biology.

Acknowledgments

The authors acknowledge support from NSF grant 0519563 and the EPA grant GAD R 832721-010. EY acknowledges a Graduate Training Fellowship through the NSF IGERT Program on Integratively Engineered Biointerfaces, DGE 0333196.

Literature Cited

1. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19**(18):2369-2380.
2. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, **5**:170.
3. Carmack CS, McCue LA, Newberg LA, Lawrence CE: **PhyloScan: identification of transcription factor binding sites using cross-species evidence.** *Algorithms Mol Biol* 2007, **2**:1.
4. Siddharthan R, Siggia ED, van Nimwegen E: **PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.** *PLoS Comput Biol* 2005, **1**(7):e67.
5. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12**(5):739-748.
6. Fang F, Blanchette M: **FootPrinter3: phylogenetic footprinting in partially alignable sequences.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W617-620.
7. Crick F: **On Protein Synthesis.** *Symp Soc Exp Biol* 1958, **XII**:139-163.
8. Crick F: **Central Dogma of Molecular Biology.** *Nature* 1970, **227**:561-563.
9. Maston GA, Evans SK, Green MR: **Transcriptional Regulatory Elements in the Human Genome.** *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
10. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16**(1):1-10.
11. Consortium IHGS: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
12. Blais A, Dynlacht BD: **Devising transcriptional regulatory networks operating during the cell cycle and differentiation using ChIP-on-chip.** *Chromosome Res* 2005, **13**(275-288).
13. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**:2306-2309.
14. Ellington AD, Szostak JW: **In vitro selection of RNA molecules that bind specific ligands.** *Nature* 1990, **346**:818-822.
15. Stoltenburga R, Reinemanna C, Strehlitz B: **SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands.** *Biomolecular Engineering* 2007, **22**(4):381-403.
16. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**(4):276-287.
17. Lewin B: **Gene IX - Promoters and Enhancers.** 2007, **ch.24**:609-635.
18. Butler JEF, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes Dev* 2002, **16**(20):2583-2592.
19. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72**:449-479.
20. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes.** *Annu Rev Genet* 2000, **34**:77-137.
21. Périer RC, Junier T, Bucher P: **The Eukaryotic Promoter Database EPD.** *Nucleic Acids Res* 1997, **26**(1):353-357.
22. Périer RC, Praz V, Junier T, Bonnard C, Bucher P: **The eukaryotic promoter database (EPD).** *Nucleic Acids Res* 2000, **28**(1):302-303.

23. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs**. *Nucleic Acids Res* 2002, **30**(1):328-331.
24. Gershenzon NI, Ioshikhes IP: **Synergy of human Pol II core promoter elements revealed by statistical sequence analysis**. *Bioinformatics* 2005, **21**(8):1295-1300.
25. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z: **Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1**. *Genome Res* 2007, **17**(6):798-806.
26. McKnight SL, Kingsbury R: **Transcriptional control signals of a eukaryotic protein-coding gene**. *Science* 1982, **217**:316-324.
27. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes**. *J Mol Biol* 1987, **196**(2):261-282.
28. Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters**. *Proc Natl Acad Sci USA* 2006, **103**(5):1412-1417.
29. Ioshikhes IP, Zhang MQ: **Large-scale human promoter mapping using CpG islands**. *Nat Genet* 2000, **26**(1):61-63.
30. Blackwood EM, Kadonaga JT: **Going the Distance: A Current View of Enhancer Action**. *Science* 1998, **281**(5373):60 - 63.
31. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G: **Locus control regions**. *Blood* 2002, **100**(9):3077-3086.
32. Bush EC, Lahn BT: **Selective Constraint on Noncoding Regions of Hominid Genomes**. *PLoS Comput Biol* 2005, **1**(7):e73.
33. Jegga AG, Aronow BJ: **Evolutionarily Conserved Noncoding DNA**. *Encyclopedia of Life Sci* 2006:doi:10.1002.
34. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, et al: **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nat Genet* 2006, **38**(6):626-635.
35. ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
36. Maruyama K, Sugano S: **Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides**. *Gene* 1994, **138**(1-2):171-174.
37. Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M et al: **High-efficiency full-length cDNA cloning by biotinylated CAP trapper**. *Genomics* 1996, **37**(3):327-336.
38. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL: **Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment**. *Genome Biol* 2006, **7**(Suppl 1:S3):1-13.
39. Bajic VB, Tan SL, Suzuki Y, Sugano S: **Promoter prediction analysis on the whole human genome**. *Nature Biotech* 2004, **22**(11):1467-1473.
40. FANTOM Consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group): **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**(5740):1559-1563.
41. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006**. *Nucleic Acids Res* 2007:1-5.
42. Schmid CD, Perier R, Praz V, Bucher P: **EPD in its twentieth year: towards complete promoter coverage of selected model organisms**. *Nucleic Acids Res* 2006, **34**(Database issue):D82-D85.

43. Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C: **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002.** *Nucleic Acids Res* 2002, **30**(1):335-340.
44. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition.** *Genome Res* 1997, **7**(9):861-878.
45. 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.
46. Wu S, Xie X, Liew AW, Yan H: **Eukaryotic promoter prediction based on relative entropy and positional information.** *Phys Rev* 2007, **E 75**:041908(041901-041907).
47. Bajic VB, Seah SH: **Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units.** *Genome Res* 2003, **13**(8):1923-1929.
48. Bajic VB, Seah SH: **Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes.** *Nucleic Acids Res* 2003, **31**(13):3560-3563.
49. Bajic VB, Seah SH, Chong A, Krishnan SPT, Koh JLY, Brusic V: **Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates.** *J Mol Graphics & Modelling* 2003, **21**(5):323-332.
50. Knudsen S: **Promoter2.0: for the recognition of PolII promoter sequences.** *Bioinformatics* 1999, **15**:356-361.
51. Li T, Chen C: **PromPredictor: A Hybrid Machine Learning System for Recognition and Location of Transcription Start Sites in Human Genome.** *LNAI* 2005, **3584**:552-563.
52. Reese MG: **Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome.** *Comput Chem* 2001, **26**(1):51-56.
53. Scherf M, Klingenhoff A, Werner T: **Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach.** *J Mol Biol* 2000, **297**(3):599-606.
54. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29**(4):412-417.
55. Solovyev VV, Shahmuradov IA: **PromH: Promoters identification using orthologous genomic sequences.** *Nucleic Acids Res* 2003, **31**(13):3540-3545.
56. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12**(3):458-461.
57. Luo Q, Yang W, Liu P: **Promoter recognition based on the Interpolated Markov Chains optimized via simulated annealing and genetic algorithm** *Pattern Recognition Letters* 2006, **27**(9):1031-1036.
58. Ohler U, Harbeck S, Niemann H, Noth E, Reese MG: **Interpolated markov chains for eukaryotic promoter recognition.** *Bioinformatics* 1999, **15**:362-369.
59. Ohler U, Liao GC, Niemann H, Rubin GM: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3**(12):RESEARCH0087.
60. Ohler U, Stemmer G, Harbeck S, Niemann H: **Stochastic Segment Models of Eukaryotic Promoter Regions.** *Pacific Symp on Biocomp* 2000, **5**:377-388.
61. Sonnenburg S, Zien A, Rätsch G: **ARTS: accurate recognition of transcription starts in human.** *Bioinformatics* 2006, **22**(14):e472-e480.
62. Hannenhalli S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17**:S90-S96.
63. Xie X, Wu S, Lam KM, Yan H: **PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm.** *Bioinformatics* 2006, **22**(22):2722-2728.

64. Florquin K, Saeys Y, Degroeve S, Rouzé P, Van de Peer Y: **Large-scale structural analysis of the core promoter in mammalian and plant genomes.** *Nucleic Acids Res* 2005, **33**(13):4255-4264.
65. Uren P, Cameron-Jones RM, Sale A: **Promoter Prediction Using Physico-Chemical Properties of DNA.** *LNCS* 2006, **4216**:21-31.
66. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5**(1):201.
67. Qiu P: **Recent advances in computational promoter analysis in understanding the transcriptional regulatory network.** *Biochem Biophys Res Commun* 2003, **309**(3):495-501.
68. Brazma A, Jonassen I, Eidhammer I, Gilbert D: **Approaches to the automatic discovery of patterns in biosequences.** *J Comput Biol* 1998, **5**(2):279-305.
69. Sandve GK, Drablos F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct* 2006, **1**:11.
70. Matys V, Fricke E, Geffers R, et al.: **TRANSFAC®: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31**:374-378.
71. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC®: A database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**:238-241.
72. Genomatix DB: <http://www.genomatix.de/>.
73. Pavesi G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W199-203.
74. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis: probabilistic models of proteins and nucleic acids.** *Cambridge University Press* 1998.
75. Yada T, Totoki Y, Ishikawa M, Asai K, Nakai K: **Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences.** *Bioinformatics* 1998, **14**(4):317-325.
76. Pavesi G, Mauri G, Pesole G: **In silico representation and discovery of transcription factor binding sites.** *Brief Bioinform* 2004, **5**(3):217-236.
77. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21**(13):2933-2942.
78. Chekmenev DS, Haid C, Kel AE: **P-Match: transcription factor binding site search by combining patterns and weight matrices.** *Nucleic Acids Res* 2005, **33**:W432-W437.
79. Goessling E, Kel-Margoulis OV, Kel AE, Wingender E: **MATCH™ - a tool for searching transcription factor binding sites in DNA sequences.** *GCB '01* 2001:158-161.
80. Ellrott K, Yang C, Sladek FM, Jiang T: **Identifying transcription factor binding sites through Markov chain optimization.** *Bioinformatics* 2002, **18**:S100-S109.
81. Hannenhalli S, Wang LS: **Enhanced position weight matrices using mixture models.** *Bioinformatics* 2005, **21 Suppl 1**:i204-212.
82. Gershenzon NI, Stormo GD, Ioshikhes IP: **Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites.** *Nucleic Acids Res* 2005, **33**(7):2290-2301.
83. Fu Y, Weng Z: **Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences.** *Genome Inform* 2005, **16**(1):68-72.
84. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *J Mol Biol* 2004, **338**(2):207-215.
85. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.

86. Morgenstern B, Frech K, Dress A, Werner T: **DIALIGN: finding local similarities by multiple sequence alignment.** *Bioinformatics* 1998, **14**(3):290-294.
87. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**(1):205-217.
88. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15**(2):330-340.
89. Marsan L, Sagot MF: **Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.** *J Comput Biol* 2000, **7**(3-4):345-362.
90. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**(7-8):563-577.
91. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
92. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4**(8):1618-1632.
93. Ambesi-Impiombato A, Bansal M, Lio P, di Bernardo D: **Computational framework for the prediction of transcription factor binding sites by multiple data integration.** *BMC Neurosci* 2006, **7 Suppl 1**:S8.
94. Birnbaum K, Benfey PN, Shasha DE: **cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships.** *Genome Res* 2001, **11**(9):1567-1573.
95. Kim SY, Kim Y: **Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data.** *BMC Bioinformatics* 2006, **7**:330.
96. Bae SH, Tang H, Wu J, Xie J, Kim S: **dPattern: transcription factor binding site (TFBS) discovery in human genome using a discriminative pattern analysis.** *Bioinformatics* 2007, **23**(19):2619-2621.
97. Defrance M, Touzet H: **Predicting transcription factor binding sites using local over-representation and comparative genomics.** *BMC Bioinformatics* 2006, **7**:396.
98. Monsieurs P, Thijs G, Fadda AA, De Keersmaecker SC, Vanderleyden J, De Moor B, Marchal K: **More robust detection of motifs in coexpressed genes by using phylogenetic information.** *BMC Bioinformatics* 2006, **7**:160.
99. Zeng E, Narasimhan G: **Enhancing Motif Refinement by Incorporating Comparative Genomics Data.** *LNBI* 2007, **4463**:329-337.
100. Zhang Y, Zaki MJ: **EXMOTIF: efficient structured motif extraction.** *Algorithms Mol Biol* 2006, **1**:21.
101. Zhang Y, Zaki MJ: **SMOTIF: efficient structured pattern and profile motif search.** *Algorithms Mol Biol* 2006, **1**:22.
102. Chakravarty A, Carlson JM, Khetani RS, Gross RH: **A novel ensemble learning method for de novo computational identification of DNA binding sites.** *BMC Bioinformatics* 2007, **8**:249.
103. Carlson JM, Chakravarty A, Gross RH: **BEAM: a beam search algorithm for the identification of cis-regulatory elements in groups of genes.** *J Comput Biol* 2006, **13**(3):686-701.
104. Carlson JM, Chakravarty A, Khetani RS, Gross RH: **Bounded search for de novo identification of degenerate cis-regulatory elements.** *BMC Bioinformatics* 2006, **7**:254.
105. Chakravarty A, Carlson JM, Khetani RS, DeZiel CE, Gross RH: **SPACER: identification of cis-regulatory elements with non-contiguous critical residues.** *Bioinformatics* 2007, **23**(8):1029-1031.

106. Blanco E, Messeguer X, Smith TF, Guigo R: **Transcription factor map alignment of promoter regions**. *PLoS Comput Biol* 2006, **2**(5):e49.
107. Blanco E, Guigo R, Messeguer X: **Multiple non-collinear TF-map alignments of promoter regions**. *BMC Bioinformatics* 2007, **8**:138.
108. Veerla S, Hoglund M: **Analysis of promoter regions of co-expressed genes identified by microarray analysis**. *BMC Bioinformatics* 2006, **7**:384.
109. Zhou BQ, Wong WH: **Coupling hidden Markov models for the discovery of cisRegulatory modules in multiple species**. *Annals of Applied Statistics* 2007, **1**(1):36-65.
110. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling**. *Proc Natl Acad Sci U S A* 2004, **101**(33):12114-12119.
111. Sharan R, Ben-Hur A, Loots GG, Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W253-256.
112. Philippakis AA, He FS, Bulyk ML: **Modulefinder: a tool for computational discovery of cis regulatory modules**. *Pac Symp Biocomput* 2005:519-530.
113. Werner T, Fessele S, Maier H, Nelson PJ: **Computer modeling of promoter organization as a tool to study transcriptional coregulation**. *FASEB J* 2003, **17**(10):1228-1237.
114. Chowdhary R, Ali RA, Albig W, Doenecke D, Bajic VB: **Promoter modeling: the case study of mammalian histone promoters**. *Bioinformatics* 2005, **21**(11):2623-2628.
115. Shelest E, Wingender E: **Construction of predictive promoter models on the example of antibacterial response of human epithelial cells**. *Theor Biol Med Model* 2005, **2**:2.
116. Hüttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype?** *Trends Genet* 2005, **21**(5):289-297.
117. Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison**. *Nucleic Acids Res* 2004, **32**(Web server):W249-W252.
118. Loots GG, Ovcharenko I: **rVISTA 2.0: evolutionary analysis of transcription factor binding sites**. *Nucleic Acids Res* 2004, **32**(Web server):W217-W221.
119. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER--an integrative program suite for microarray data analysis**. *BMC Bioinformatics* 2005, **6**:232.
120. Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B: **TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis**. *Nucleic Acids Res* 2005, **33**(Web server):W393-396.
121. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, **32**(Database):D91-D94.
122. Zhao F, Xuan Z, Liu L, Zhang MQ: **TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies**. *Nucleic Acids Res* 2005, **33**(Database issue):D103-107.
123. Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV: **PlantProm: a database of plant promoter sequences**. *Nucleic Acids Res* 2003, **31**(1):114-117.
124. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database:1999**. *Nucleic Acids Res* 1999, **27**(1):297-300.
125. Zhu J, Zhang MQ: **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae***. *Bioinformatics* 1999, **15**(7-8):607-611.
126. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG: **Transcription Regulatory Regions Database (TRRD): its status in 2002**. *Nucleic Acids Res* 2002, **30**(1):312-317.

127. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V *et al*: **Ensembl 2002: accommodating comparative genomics**. *Nucleic Acids Res* 2003, **31**(1):38-42.
128. Brown RH, Gross SS, Brent MR: **Begin at the beginning: predicting genes with 5' UTRs**. *Genome Res* 2005, **15**(5):742-747.
129. Gross SS, Brent MR: **Using Multiple Alignments to Improve Gene Prediction**. *J Comp Bio* 2006, **13**(2):379-393.
130. Ponger L, Mouchiroud D: **CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences**. *Bioinformatics* 2002, **18**(4):631-633
131. Liu R, States DJ: **Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling**. *Genome Res* 2002, **12**(3):462-469.
132. Nomenclature Committee of the International Union of Biochemistry (NC-IUB): **Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984**. *Proc Natl Acad Sci U S A* 1986, **83**(1):4-8.
133. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al*: **Assessing computational tools for the discovery of transcription factor binding sites**. *Nat Biotechnol* 2005, **23**(1):137-144.
134. 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.
135. Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity**. *Pac Symp Biocomput* 2000:467-478.
136. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling**. *Bioinformatics* 2001, **17**(12):1113-1122.
137. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies**. *J Mol Biol* 1998, **281**(5):827-842.
138. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads**. *Nucleic Acids Res* 2000, **28**(8):1808-1818.
139. Sinha S, Tompa M: **YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation**. *Nucleic Acids Res* 2003, **31**(13):3586-3588.

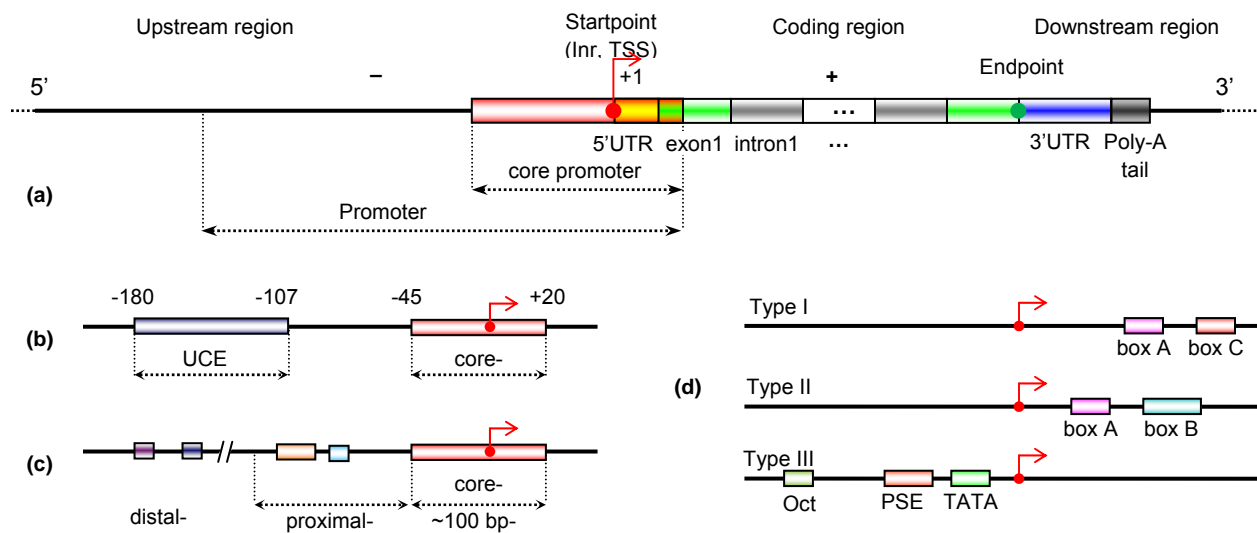


Figure 1: Basic structure of promoter classes [17]. (a) A general structure of an eukaryote gene; the promoter region contains crucial regulatory elements to control the transcription of the gene; the gene is copied to a pre-mRNA from which the RNA Pol-II transcribes into an mRNA; the coding region contains alternatively exons and introns where introns are removed in the transcription process; a gene is marked by an integer 1D-coordinate system without zero point, i.e. TSS is +1 and before is negative; the untranslated regions (UTRs) are particular sections of mRNA; the 5' UTR starts from the TSS and ends just before the start codon (usually AUG), the 3' UTR follows the coding region and ends before the poly-A tail – the sign to stop the transcription. (b)(c) Typical structures of class I promoters and class II promoters, respectively. (d) The typical structure of class III promoters; box A, B, C as well as TATA, PSE, Oct are conserved sequences which are bound by TFs to initialize the transcription process; internal promoters (Type I, II) have short conserved sequences located within the coding region; upstream promoters (Type III) contain short conserved sequences upstream of the start point.

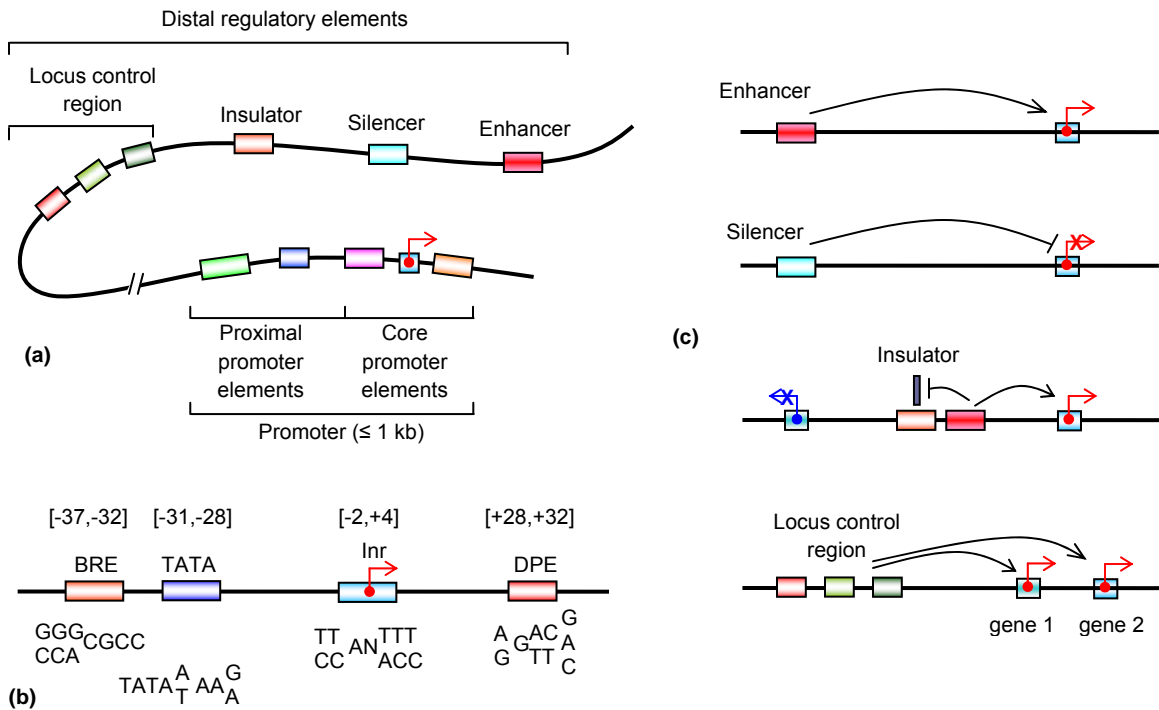


Figure 2: Class II promoter structure and relevant regulatory elements; these are redrawn from ([9, 19]). (a) Typical regulatory elements of a gene including a core promoter, proximal promoter elements and distal regulatory elements; the promoter region which contains a core promoter and proximal promoter elements is usually no longer than 1kb. (b) A detailed structure of a core promoter; the top is the positions of the conserved elements in the core promoter within the gene coordinate system; the bottom is the corresponding consensus sequences (c) Four typical elements of distal regulatory elements and their corresponding effects; enhancers activate whereas silencers repress the transcription; insulators block the gene from being affected by other regulatory elements; A locus control region can affect the transcription of a number of genes.

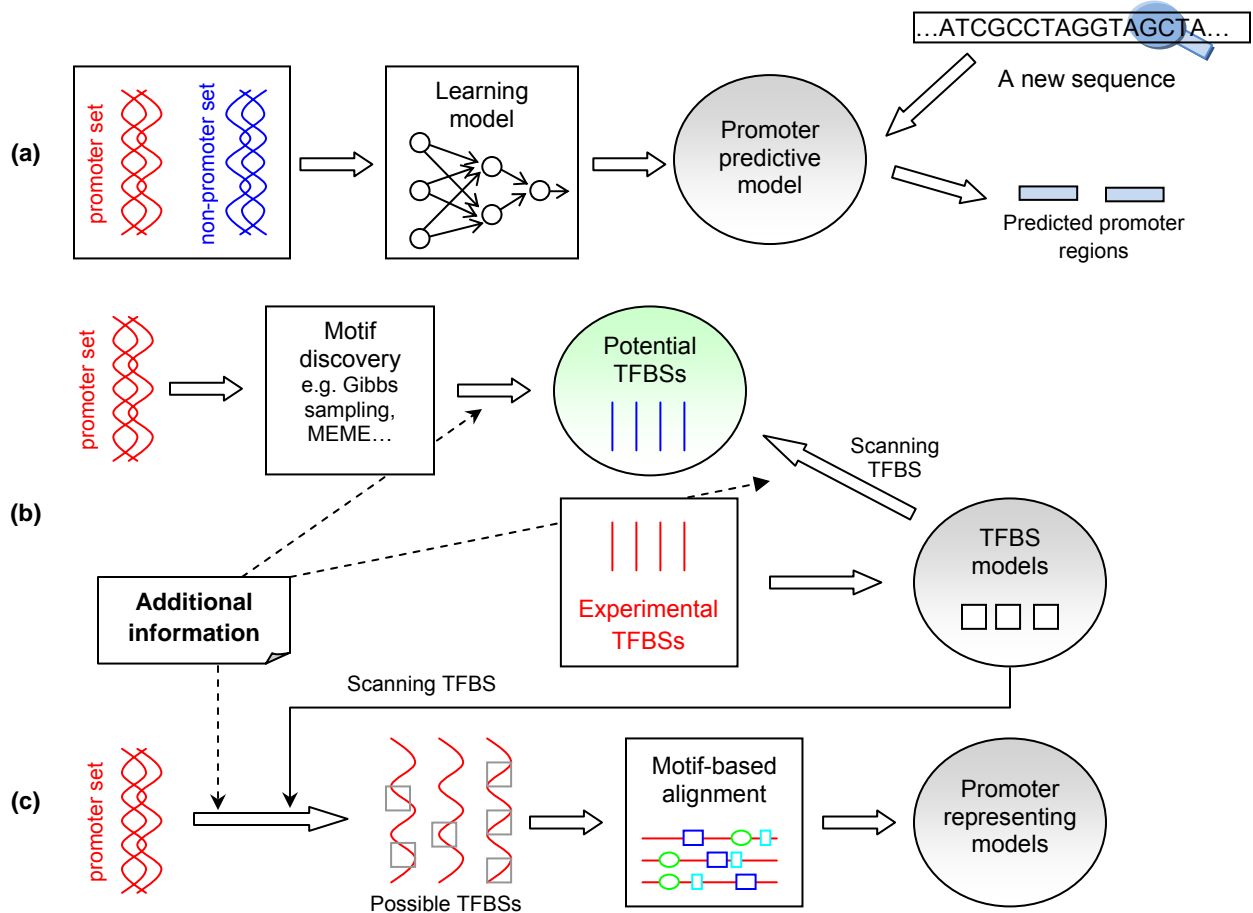


Figure 3: A brief view on three typical aspects in promoter sequence analysis. (a) Promoter prediction can recognize approximately the promoter regions (or TSSs) in a new sequence or differentiate which promoter (or non-promoter) set a sequence belongs to. (b) TFBSs discovery can be searched by purely computational techniques or scanned under the support of experimental data; additional information from gene expression data or cross-species data with orthologous sequences can be utilized to enhance the accuracy; dash-line shows that it can be used or not. (c) A new way to represent the promoter sequence e.g. a list of ordered TFBSs; the promoter sequence can be transformed following some way so that applications can exploit as many as possible its features.

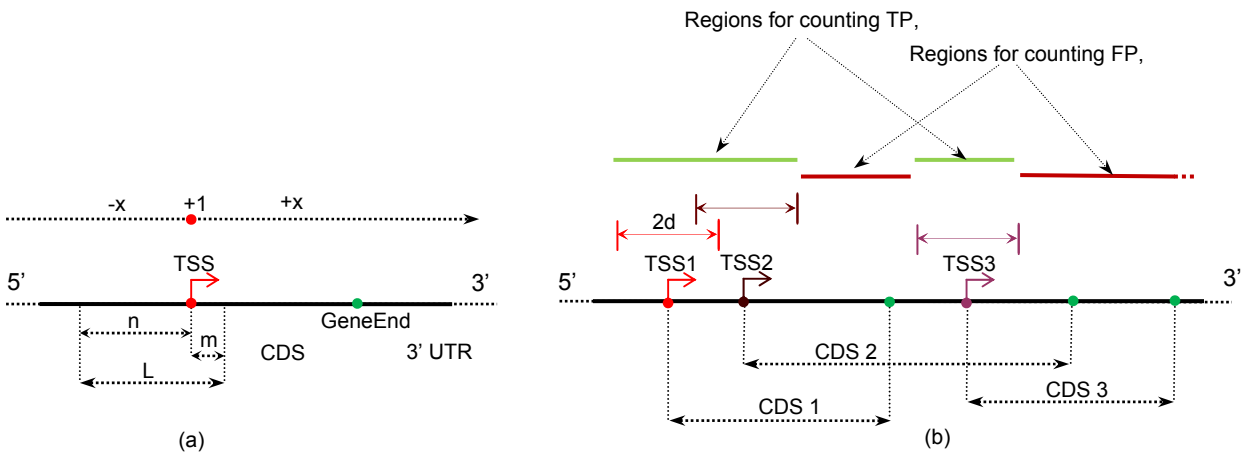


Figure 4: Method to collect data and evaluate PPPs [38]. (a) A local coordinate system of every gene on the genomic sequence, starting at TSS and ending at GeneEnd; the promoter sequence in this aspect is the subsequence DNA, length L (usually from 100 to 500) with n bp upstream and m bp downstream ($n \gg m$). (b) The evaluation method for problem 1c; the TSSs are inferred for estimating the performance instead of using directly the results from the predictive models as that in problem 1b; if d is the maximal allowance for the region to be considered as a hit, $2d$ is the correct region; the orange lines are the regions for counting TP if there is one hit, if not it will be counted as FN; the dark red lines shows the regions to count for FP if there is a hit in there, if not it will be a TN; however, for every line, it is only considered as one TP (FN) or one FP [116] although there are many hits (TP or FP) or no hit (FN or TN) on there; the dash line region of the second dark-red line implies that the region continuously lengthens.

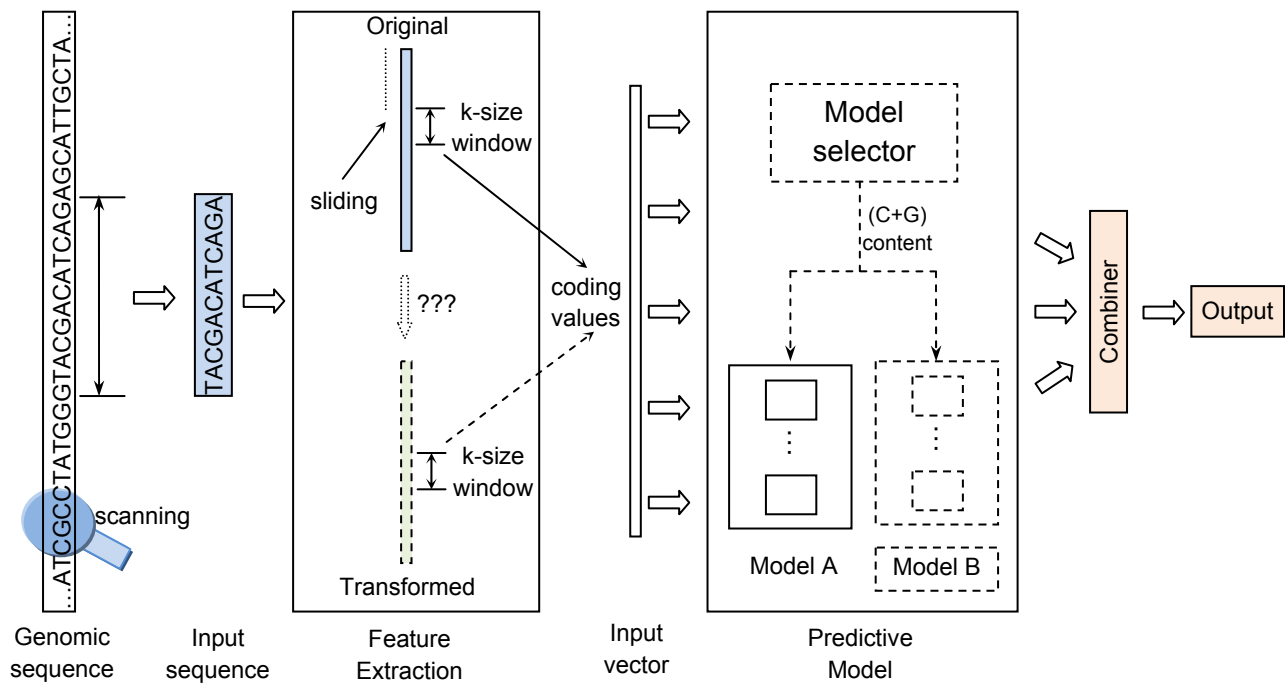


Figure 5: A general framework for the working mechanism of PPPs based on content searching. An input sequence is drawn from the genomic sequence following the criteria of problem 1b or 1c. From the input sequence, whether it is transformed or not depends on different PPPs but then a k-size window slides on it to extract k-size words which are coded into numeric values and put into the predictive model. The model selector can be appeared in some PPPs to characterize if the input sequence is CpG related or non-CpG related since their properties might be different. After that, the input vector is put into model A (or model B) which contains a list of classifiers (usually ANNs). The output is determined based on the results of previous classifiers. Dash arrows and dash frames show that the modules can be applied or not.

Site name	Alignment	Matrix similarity
Chop(1)	ATTGCATCA	(0.992)
Chop(2)	GTTTCACCA	(0.952)
Asn-S(1)	GTTTCATCA	(1.000)
Asn-S(2)	ATTACATCA	(0.974)
Trp-tRNA-S(1)	ATTGCATCA	(0.992)
Trp-tRNA-S(2)	GTTTCCTCA	(0.910)
Trp-tRNA-S(3)	GTTTCCTCA	(0.910)
Neutral	TTTGCATCA	(0.975)
Neutral	ATTTTCATCT	(0.876)
DRAL/FHL2	TTTCCATCA	(0.958)
PAX6	ATTGCACCA	(0.945)
Asp-AT	GTTGCATCA	(0.996)
P5C_reductase(1)	GTTGCATCA	(0.996)
P5C_reductase(2)	GTTTCACCA	(0.952)
alpha-L-iduronidase	ATTTCAACA	(0.936)

Raw motif profile by PWM:									
Pos.	1	2	3	4	5	6	7	8	9
A	6	0	0	1	0	13	1	0	14
C	0	0	0	1	15	2	3	15	0
G	7	0	0	6	0	0	0	0	0
T	2	15	15	7	0	0	11	0	1

Motif profile and consensus sequence:									
Pos.	1	2	3	4	5	6	7	8	9
A	.40	.00	.00	.07	.00	.87	.07	.00	.93
C	.00	.00	.00	.07	1.0	.13	.20	1.0	.00
G	.47	.00	.00	.40	.00	.00	.00	.00	.00
T	.13	1.0	1.0	.46	.00	.00	.73	.00	.07
IUPAC	R	T	T	K	C	A	T	C	A

Figure 6: The motif of the amino acid response element, activating transcription factor 4 (ATF4) binding sites from Genomatix – matrix name V\$AARE.01 [72]. Left is the list of binding sites used to construct the motif model; each sequence is called an oligo or a conserved sequence; oligos can be aligned with gaps to maximize the motif content but in this case, it is a gap-free alignment; therefore, the motif model is also the middle column in the table; matrix similarity is the fitness of the corresponding site with the motif profile after construction. Right is the motif profile which is created by using the PWM method without using pseudo-count and any normalization method; however, sometimes the name PWM can be used to indicate the motif profile; the raw profile (top) which is counted directly from the left alignment is normalized by the total number of binding sites to estimate the motif profile (bottom); the consensus sequence is drawn from the distribution of bases in each column with IUPAC characters from Table 4; the consensus sequence is also called the (conserved) motif or the (conserved) pattern.

<p><u>Pattern-driven algorithm:</u></p> <ul style="list-style-type: none"> - Let W be all 4^K possible patterns from AA...A to TT...T and w_{ik} be an oligo on sequence i^{th} at position k^{th}. - For each $p \in W$ do <ul style="list-style-type: none"> - For $i = 1$ to N do <ul style="list-style-type: none"> - $d(p, s_i) = \min_k \{d(p, w_{ik})\}$ - End For - $d(p, S) = \sum_{i=1}^N d(p, s_i)$ - End For - Report some p^* with $d(p, S) > \delta$ for some δ. 	<p><u>Sample-driven algorithm:</u></p> <ul style="list-style-type: none"> - Let $W = \{w_{ik}, i = 1..N, k = 1..(\ s_i\ - K)\}$ be a collection of all K-size word w_{ik} in S. - For each $p \in W$ do <ul style="list-style-type: none"> - For $i = 1$ to N do <ul style="list-style-type: none"> - $d(p, s_i) = \min_k \{d(p, w_{ik})\}$ - End For - $d(p, S) = \sum_{i=1}^N d(p, s_i)$ - End For - Report some p^* with $d(p, S) > \delta$ for some δ.
<p><u>Consensus algorithm:</u></p> <ul style="list-style-type: none"> - Let W_i be all K-size word of s_i. - Assume each $p \in W_1$ is an alignment and store it to $A_1 = \{A_{1j}\}_{j=1}^{\ W_1\ }$. - For $i = 2$ to N do <ul style="list-style-type: none"> -For each $p \in W_i$ do <ul style="list-style-type: none"> -For $j = 1$ to $\ A_{i-1}\$ do <ul style="list-style-type: none"> -Create alignment of $(p, A_{i-1,j})$ -End For -End For - Select and keep good alignment in the new set of alignments A_i. - End For 	<p><u>Consensus illustration:</u></p> <p>Assume $S = \{\text{ACTGAT}, \text{CTGAAC}, \text{AGATGA}\}$</p> <p>Cycle 1: Keep all K-size words of s_1 in A_1 ACTG CTGA TGAT</p> <p>Cycle 2: Keep best alignments between W_2 and A_1 in A_2 (using gap-free alignment, no more two different bases between two oligos) CTGA TGAT CTGA TGAA</p> <p>Cycle 3: Keep best alignments between W_3 and A_2 CTGA TGAT CTGA TGAA ATGA AGAT</p> <p>Report best alignments in A_3</p>

Figure 7: Algorithms for exhaustive search and consensus methods. At the top are two types of exhaustive search; the algorithms are the same but the input sets are different. At the bottom is a consensus algorithm and illustration. The algorithm can be avoided some preliminary random alignments by keeping all of the alignments in A_i for a number of cycles instead of starting the process of evaluating and selecting from the beginning. However, many improvements were discussed in [90]

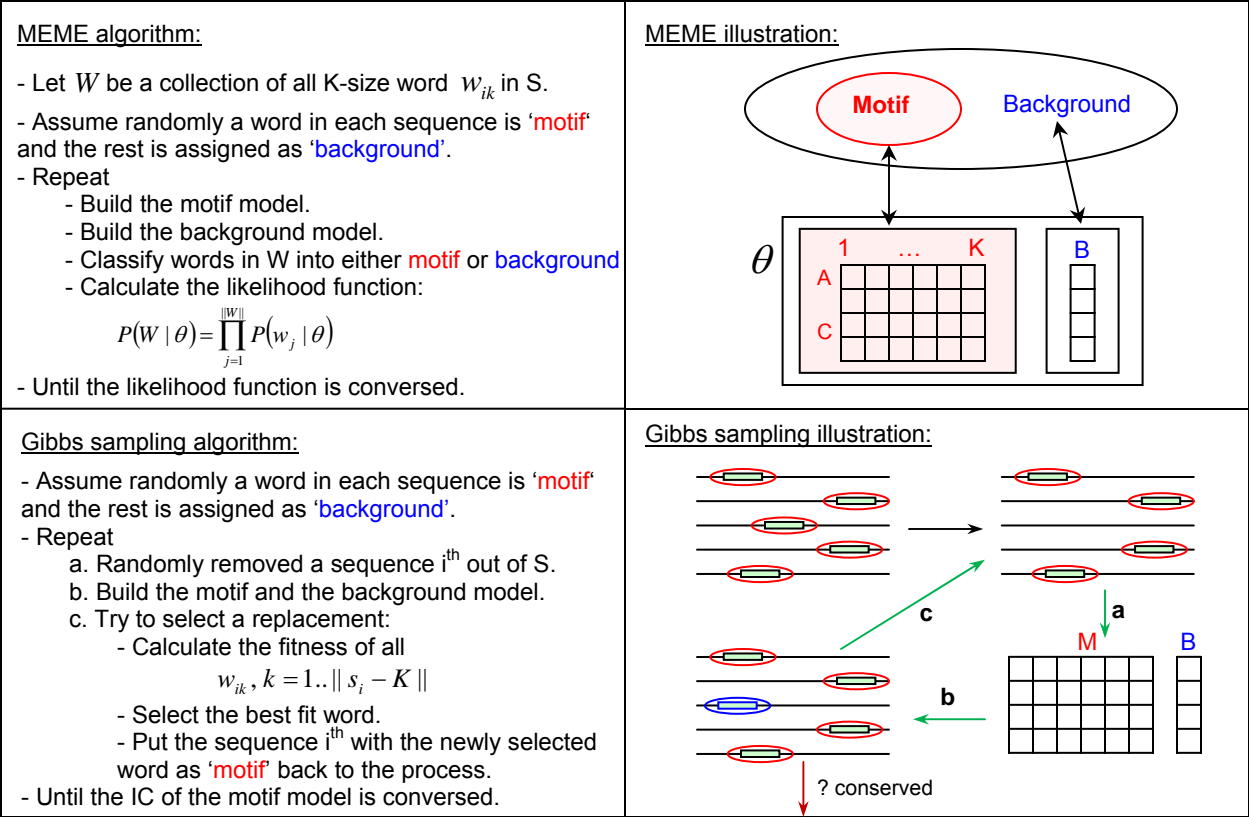


Figure 8: Basic ideas of MEME and Gibbs sampling methods. At the top is a basic MEME algorithm and the graphic representation; θ contains both the motif and the background model and is considered as a parameter to optimize. At the bottom is the basic Gibbs sampling algorithm and a brief demonstration; Step (c) can be replaced by randomly selecting a new word instead of examining all words in sequence s_i for the best fit word. Gibbs sampling is less likely to get stuck in a local optimum because it randomly selects a new oligo for each loop whereas MEME updates all oligos of the motif for every time, causing it easily to fall into local optima, in theory.

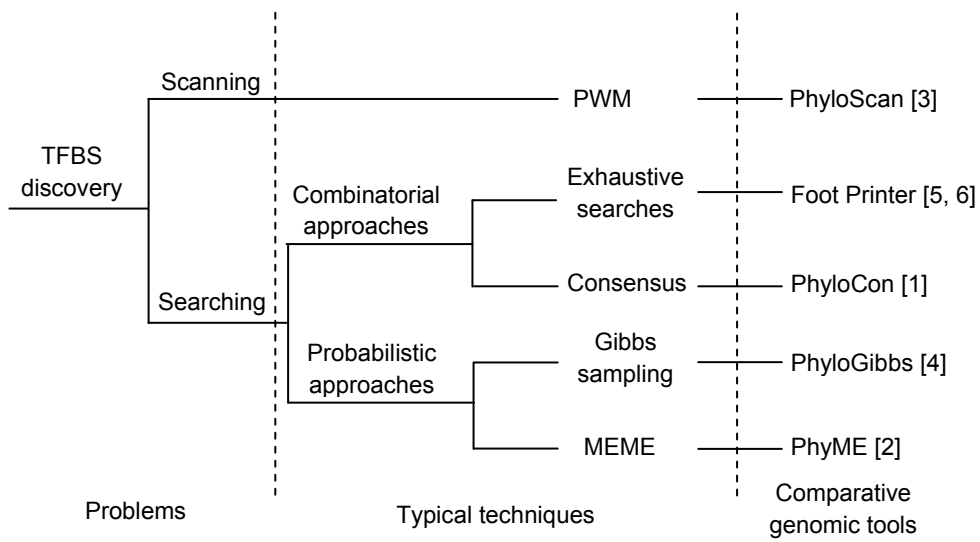


Figure 9: Basic TFBS discovery techniques. The core techniques in this aspect are classified into two categories; each is represented by two typical methods. Those methods are become standard methods in developing tools for discovering overrepresented motifs in a given set of promoter sequences (Table 5). They are also applied when combining with gene expression data whereas specific tools are developed for comparative genomics when combining with orthologous information. But these are two different directions; one consider the activity of a TFBS to know how much it can become a real TFBS via the expression level of its TF and its gene set; the other only finds conserved regions in the set of sequences and considered how much those becomes real regulatory regions. There are some other popular tools e.g. ConSite [117], rVISTA [118], or some designed as collections of useful functions for regulatory studies such as Expander [119], TOUCHCAN [120].

Table 1: Promoter-relevant databases

No.	Database	URL	Reference
1.	DBTSS – Database of Transcription Start Sites	http://dbtss.hgc.jp/	[23, 41]
2.	EPD – Eukaryotic Promoter Database	http://www.epd.isb-sib.ch/	[21, 22]
3.	UTRdb – UTR databvase.	bighost.area.ba.cnr.it/BioWWW/Bio-WWW.htm	[43]
4.	TRANSFAC – Transcription Factor and Gene Regulatory Database	http://www.gene-regulation.com/	[70, 71]
5.	Genomatix	http://www.genomatix.de/	[72]
6.	JASPAR	http://jaspar.genereg.net/	[121]
7.	TRED – Transcriptional Regulatory Element Database	http://rulai.cshl.edu/TRED	[122]
8.	PlantPromDB – Database of Plant Promoter Sequences	http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom	[123]
9.	PLACE – A Database of Plant Cis-acting Regulatory DNA Elements	http://www.dna.affrc.go.jp/PLACE/	[124]
10.	SCPD – Yeast Promoter Database	http://rulai.cshl.edu/SCPD	[125]
11.	TRRD - Transcription Regulatory Regions Database	http://www.mgs.bionet.nsc.ru/mgs/gnw/trrd/	[126]
12.	Ensemble	http://www.ensemble.org	[127]

Table 2: Promoter prediction programs (PPPs) in more details.

Prediction programs	Learning techniques	Technical details	URL	Reference
ARTS	SVM with string kernel	Use suffix trees in the string kernel to exploit significant k-size words in the promoter regions; the technique can be referred to taxonomy classification.	http://www.fml.tuebingen.mpg.de/raetsch/projects/arts	[61]
Promoter Explorer	AdaBoost algorithm	Analyze local distribution of 5-size words, CpG island, and digitized DNA sequence; then combine them for a cascade learning process of classifiers to lower false positive prediction.	http://www.hy8.com/~tec/papers/pexp01.zip	[63]
N-SCAN	Multi-genome alignments, tree-structured HMM	Comparative gene prediction by integrating multi-genome and 5'UTR modeling.	http://mblab.wustl.edu/nscan/submit/	[128, 129]
PromPredictor	Feature selection, Artificial Neural Network (ANN)	Select significant 5-size words in the promoter regions for 4 basic classifiers: promoter-, exon-, intron-, and 3'UTR- classifier.	http://www.whtelecom.com/Prompredictor.htm	[51]
DragonPF (ver1.5)	Composite-model structure, ANN, statistics techniques	Divide into G+C-rich and G+C-poor models; each has a number of sub-models (ANNs) corresponding to which type of sequences (introns, exons, or promoter) is sensitive.	http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm	[49]
*DragonGSF	ANN	Combine information about CpG islands, predicted TSSs, and signals downstream of the predicted TSSs; predicted TSSs are from DragonPF.	http://research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm	[47, 48]
McPromoter	Interpolated Markov model	Exploit the statistical properties of the core promoter (significant motifs); apply stochastic segment models in which each state is a functional part of the promoter.	http://genes.mit.edu/McPromoter.html	[59, 60]
PromH	Linear discriminant function	Based on conservative features of promoter regions in pairs of orthologous genes.	http://www.softberry.com/berry.phtml?topic=promhg&group=programs&subgroup=promoter	[55]
Eponine	Relevance Vector Machine	Constrain position by a collection of motif weight matrixes; focus on TATA-box motif and G+C rich domain.	http://www.sanger.ac.uk/Users/td2/eponine/	[56]
CpGProD	Statistical rules	Detect only CpG islands-related promoters; parameters are higher sensitive with species; suggest strand.	http://pbil.univ-lyon1.fr/software/cpgprod.html	[130]
CONPRO	Consensus promoter prediction	Restrict the searched genomic regions using gene transcript alignment as anchors; then, use GENESCAN to build the model and combine with the results of some previous programs to infer the promoter regions.	http://stl.bioinformatics.med.umich.edu/conpro/	[131]
FirstEF	Quadratic discriminant function	Recognize CpG islands, promoter regions, and first splice-donor sites by a posterior probability $P(\text{promoter} \text{window})$.	http://rulai.cshl.org/tools/FirstEF	[54]
NNPP2.2	Time-delay neural network	Train independently two time-delay ANNs for recognizing TATA-box and Inr within a corresponding window; then combine to enhance the ability.	http://www.fruitfly.org/seq_tools/promoter.html	[52]
*Promoter Inspector	ANNs	Focus on promoter context rather than exact locations; extract IUPAC words for 3 classifiers to differentiate between promoters and introns, exons, and 3'UTR.	http://www.genomatix.de/promoterinspector.html	[53]
Promoter2.0	ANN	Use GA to optimize NN's weights; train the NN to recognize a set of sub-patterns (6-size words) and specialize on 4 TFBSs (TATA, CAAT, Inr, and GC) between promoter- and non-promoter seq.	http://www.cbs.dtu.dk/services/promoter/	[50]

* : implies recommended PPPs.

Table 3: IUPAC-IUB single-letter codes recommended for ambiguous positions in DNA sequences [132]

IUPAC	Nucleotides	Mnemonics
A		Adenine
C		Cytosine
G		Guanine
T		Thymine
R	A or G	puRines
Y	C or T	pYrimidines
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C, T	not G
B	C, G, T	not A
V	A, C, G	not T
D	A, G, T	not C
N	A, C, G, T	aNy

Table 4: Typically selected a couple of tools for overrepresented motif discovery [133]

Programs	Operating principle	Technical data	URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are overrepresented in the input set	Judges alignments sampled during the course of the algorithm using a maximum a priori log likelihood score, which gauges the degree of over-representation. Provides an adjunct measure (group specificity score) that takes into account the sequence of the entire genome and highlights those motifs found preferentially in association with the genes under consideration.	http://atlas.med.harvard.edu/	[134]
ANN-Spec	Models the DNA-binding specificity of a transcription factor using a weight matrix	Objective function based on log likelihood that transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples.	http://www.cbs.dtu.dk/~workman/ann-spec/	[135]
Consensus	Models motifs using weight matrices, searching for the matrix with maximum information content	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif resulting in greatest information content, and so on.	http://bifrost.wustl.edu/consensus/	[90]
MEME	Optimizes the E-value of a statistic related to the information content of the motif	Rather than sum of information content of each motif column, statistic used is the product of the P values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	http://meme.sdsc.edu/	[91]
MotifSampler	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence.	http://www.esat.kuleuven.ac.be/~dna/Biol/Software.html	[136]
Oligo/dyad	Detects overrepresented oligo-nucleotides with oligo-analysis and spaced motifs with dyad-analysis	These algorithms detect statistically significant motifs by counting the number of occurrences of each word or dyad and comparing these with expectation. Most crucial parameter is choice of appropriate probabilistic model for the estimation of occurrence significance. In this study, a negative binomial distribution on word distributions was obtained from 1,000 random promoter selections of the same size as the test sets	http://rsat.scmdb.ulb.ac.be/rsat/	[137, 138]
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences	Each motif evaluated according to number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the oligo frequency analysis of all the available upstream sequences of the same organism. Different combinations of 'canonical' motif parameters derived from the analysis of known instances of yeast transcription factor binding sites (length ranging from 6 to 12, number of substitutions from 1 to 4) are automatically tried by the algorithm in different runs. It also analyzes and compares the top-scoring motifs of each run with a simple clustering method to detect which ones could be more likely to correspond to transcription factor binding sites. Best instances of each motif are selected from sequences using a weight matrix built with sites found by consensus-based algorithm.	http://159.149.109.16/Tool/ind.php	[73]
YMF	Uses an exhaustive search algorithm to find motifs with the greatest z-scores	A P value for the z-score is used to assess significance of motif. Motifs themselves are short sequences over the IUPAC alphabet, with spacers ('N's) constrained to occur in the middle of the sequence.	http://bio.cs.washington.edu/software.html#ymf	[139]

: this table is drawn out directly from Table 1 of [133]