

Review

## Recent Advances in the Computational Discovery of Transcription Factor Binding Sites

Tung T. Nguyen <sup>1,\*</sup> and Ioannis P. Androulakis <sup>2,\*</sup>

<sup>1</sup> BioMaPS Institute for Quantitative Biology, Rutgers University, New Jersey 08854, USA

<sup>2</sup> Biomedical Engineering Department, Rutgers University, New Jersey 08854, USA

E-mails: nhttung@biomaps.rutgers.edu, yannis@rci.rutgers.edu

\* Author to whom correspondence should be addressed.

Received: 6 January 2009 / Accepted: 17 March 2009 / Published: 24 March 2009

---

**Abstract:** The discovery of gene regulatory elements requires the synergism between computational and experimental techniques in order to reveal the underlying regulatory mechanisms that drive gene expression in response to external cues and signals. Utilizing the large amount of high-throughput experimental data, constantly growing in recent years, researchers have attempted to decipher the patterns which are hidden in the genomic sequences. These patterns, called motifs, are potential binding sites to transcription factors which are hypothesized to be the main regulators of the transcription process. Consequently, precise detection of these elements is required and thus a large number of computational approaches have been developed to support the *de novo* identification of TFBSs. Even though novel approaches are continuously proposed and almost all have reported some success in yeast and other lower organisms, in higher organisms the problem still remains a challenge. In this paper, we therefore review the recent developments in computational methods for transcription factor binding site prediction. We start with a brief review of the basic approaches for binding site representation and promoter identification, then discuss the techniques to locate physical TFBSs, identify functional binding sites using orthologous information, and infer functional TFBSs within some context defined by additional prior knowledge. Finally, we briefly explore the opportunities for expanding these approaches towards the computational identification of transcriptional regulatory networks.

**Keywords:** transcription factor binding sites, binding site representation, promoter analysis, phylogenetic footprinting, context-specific, transcriptional regulatory networks.

---

## 1. Introduction

The gene is the fundamental unit on the genomic DNA which contains the required information to carry out the biological functions of cells. The expression of genes i.e. mRNA synthesis can be measured efficiently in a high-throughput fashion and such expression patterns are characteristic of cellular responses to external stimuli [1]. It is widely accepted that these responses are mainly driven by the interactions between transcription factors (TFs) and their corresponding transcription factor binding sites (TFBSs) on the proximal promoters of the target genes [2, 3]. However, with a large number of genes in eukaryotic genomes, deciphering how these interactions evolve to control the expression of tens of thousands of genes (~ 35,000 genes in human) remains an open question. Recent studies [4] have shown that the underlying regulatory mechanisms are complex, dynamic (especially in higher organisms) and can be arranged in multiple hierarchical levels such as the sequence, the chromatin, and the nuclear level.

The sequence level, also the best-studied level of gene regulation, is characterized by the linear organization of transcription units and *cis*-regulatory elements considered as the regulatory code which governs gene expression. These *cis*-regulatory elements i.e. binding sites which are more important when found on the proximal promoters form a highly flexible and context-dependent structure [5] for each gene [6-8]. Furthermore, in eukaryotic cells genomic DNA is 'packed' into an efficient structure, called chromatin, composed of nucleosomes that consist of approximately 147bp of DNA wrapped around a protein octamer [9, 10]. This structure not only packs DNA but also creates an added layer of gene regulation which ensures correct gene expression and accessibility to DNA-dependent processes e.g. gene transcription, DNA repair, and DNA replication. The overall process of the transcription process encompassing the nuclear architecture and/or the complex spatial arrangement of genes, gene clusters, chromatin, and regulatory DNA elements [11, 12] is far beyond the scope of any single review and hence we only focus on the sequence level aiming at discovering *cis*-regulatory elements on the proximal promoters.

Two of the most important functional elements in gene regulation are transcription factors and their binding sites on the promoters of their target genes. A TF is a protein which binds to specific DNA binding motifs that can be present multiple times on the same promoter of a gene or on different promoters of different genes. The transcription factor binding sites where a TF binds are usually short (5 – 15bp) and degenerate but highly selective through evolution [13]. A gene can have multiple alternative promoters [14, 15] and each promoter frequently contains a large number of binding sites (10 – 50 binding sites) for 5 – 15 different TFs [16]. Therefore, a more comprehensive understanding of these elements and their interactions will provide a deeper understanding of the regulatory pathways within cells and potential functions of individual genes and/or gene clusters [17].

Although various approaches have been developed, we are still limited by both experimental and computational techniques in order to detect these binding sites and understand their interplay with corresponding TFs as well as their role in the transcription process. However, recent high-throughput technologies which identify high-affinity binding sequences e.g. ChIP-chip [18, 19], SELEX [20, 21] revealed the genomic regions to which a particular TF is bound, providing a powerful resource for discovering binding sites of transcription factors. Even though the information collected through such

methods is substantial, how to best annotate and interpret such a large volume of data and how the data can be explored to predict novel binding sites are still open issues.

Traditionally, one has attempted to extract a set of promoters of a set of genes similar in function or expression pattern using a fixed-size length of the promoter sequence and then search for statistically overrepresented subsequences, also called motifs and considered as TFBSs if they match with some known TF profiles i.e. similar to known binding sites of a TF. A large number of tools have been developed using a variety of algorithmic approaches and underlying models. The relative advantages and disadvantages of each approach is making selection among them as well as developing novel tools become a non-trivial task. Consequently, one tried to build up testing datasets to measure their performance as well as identify the strengths and limitations [22-24]. Although there remain a number of difficulties in constructing the testing datasets and the accuracy measurements, a general view from these benchmarks is that *in silico* predictions is still lack of corresponding to *in vivo* experiments. The main reason is that binding motifs are short, degenerate and contain lack of information i.e. encoded by only four types of characters (A, T, C, and G), leading to the fact that most binding sites are found as random hits throughout the genomic DNA. Additionally, regulatory elements are not randomly distributed; they tend to form clusters with a particular structure, *cis*-regulatory modules [5, 25]. Therefore, recent studies have tried to combine with additional information such as gene expression, gene annotation, phylogenetic footprinting and/or search for composite motifs instead of single motifs to increase the sensitivity of the methods [25-28].

A number of excellent reviews have addressed a variety of critical issues. Typically, Brazma et al. [29] classified motif discovery methods following the motif model (deterministic or statistical, pattern driven or sequence driven), the scoring function, and the search strategy. Pavese et al. [30] provided a very comprehensive discussion about different algorithmic methods and approaches to the problem (similarly in Wasserman [31] and Bulyk [17]). And then since individual binding sites are lack of information for algorithmic methods and recent advances have moved to model the regulatory regions or combine with other biological lines of evidence, Sandve et al. [16] proposed an integrated framework to divide the trend following the description of motif discovering models such as single motifs, composite motifs (*cis*-regulatory modules), gene level – how several modules interact together to regulate a gene, and genome level – how several classes of modules interact together to regulate a set of genes. Alternatively, with the idea of exploring the interdependence between computational and experimental techniques in this aspect, Elnitski et al. [32] made a summary on the synergism between *in silico*, *in vitro* and *in vivo* identification of transcription factor binding sites. And later, Das et al. [26] surveyed again different approaches combining with other biological evidence in motif discovery. Therefore, in this review we would like to concentrate on the developing strategies that detect physical TFBSs on the proximal promoters of target genes and some promising preliminary results on identifying functional-relevant binding sites. The remainder of this manuscript will discuss issues related to the representation of binding motifs, promoter identification and then review the basic approaches for the identification of TFBSs. Finally, we briefly explore the possibility to infer transcriptional regulatory networks under the aspect of promoter analysis.

## 2. Binding site representation

Assuming a list of DNA binding sites for some TF is available, one of the very first questions is how to best represent and characterize the information contained in these sites for further analysis. The goal is to find a representation that matches as closely as possible all the binding sites in the collection and is clearly distinguished from the background. From the point of view of string processing, a simple and widely-used concept is the consensus sequence in which the most frequent character at each position is chosen to represent in the binding motif at that position. However, some positions might consist of characters of equivalent frequencies and thus a more complex pattern, the IUPAC sequence [33, 34] was used to characterize the diversity of those binding sites (**Figure 1a**). Although this representation works well for highly conserved and short binding motifs, it is defined somewhat arbitrarily and removes much of the information in the original set of binding sites. In a case for yeast TF ABF1, for instance, two IUPAC sequences (RTCRYNNNNACG or RTCRYNNNNNACG) have been published and used as a relatively precise description of ABF1 binding sites [35]. However, these representations failed to recognize the binding site SCPK01 on PYK1 promoter from position -610 to -598 which was showed to be bound by TF ABF1 experimentally [36]. Consequently, a more precise representation was proposed to utilize almost all binding site information, known as the nucleotide distribution matrix or position weight matrix (PWM) [35, 37, 38], which has been proven very successful in various problems in DNA and protein sequence analysis [35, 39]. The PWM is a matrix of scores (e.g. occurrences, frequencies) with four rows corresponding to four DNA bases and  $m$  columns, each of which is a position in the binding motif. The basic assumption of the PWM is that the base-pairs at different positions are statistically independent and thus the fitness score of a matched oligonucleotide 'p' with this profile is the sum of the fitness at each position. This representation reflects the extent to which a position is conserved within the binding motifs and thus the higher the similarity, the higher the fitness is.

The main weakness of the PWM approach stems from the assumption is that the positions contribute independently and additively to the total activity of the binding site. However, position dependence may exist on the binding sites and has been experimentally and/or statistically verified in some cases [40]. For example, using a new quantitative multiple fluorescence relative affinity assay Man et al. [41] showed that position 16 and 17 on the operator DNA were not independent in the interactions with its TF, *Salmonella* bacteriophage repressor Mnt; or in another case, when Ellrott et al. [42] applied  $\chi^2$  test on the 71 binding sites of TF *hepatocyte nuclear factor 4 $\alpha$*  HNF4 $\alpha$ , a significant dependence was found between several pairs of positions e.g. position 4 and 8, 4 and 11. Therefore, more comprehensive representations were introduced to capture the potential dependence between positions in binding sites, such as maximal dependence decomposition [43], hidden Markov model [44, 45], Markov chain optimization [42], as well as a more flexible approach based on variable-order Bayesian network which combines PWM, Markov models and Bayesian network model to fit with each particular subset of binding sites of a TF [46].

However, despite the limitations of the basic PWM approach, it is still the leading model in the search for discovering potential TFBSs. In fact, besides its intuitive representation and fast computation, it has been shown to be comparable at least, and in some case outperforms, other more complicated models e.g. fixed-order Markov models that are usually over-fitted due to a limited



training data [46]. Therefore, emphasis has been given to strategies that optimize the PWM instead of building more complicated models. For example, the scores in the cells of the matrix can be transformed to improve the specificity of the binding motif model (e.g. convert frequencies to probabilities, adding pseudo-count, taking logarithms, etc. [30, 37]) and the binding sites can be aligned before creating the PWM [35]. In some cases, the information content (IC) of the PWM, or some similar form, is made use to select a suitable number of binding sites for creating the binding motif model [30, 47, 48];  $IC = \sum_i \sum_{b \in \{A,C,G,T\}} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$  where  $f_{b,i}$  is the observed frequency of base  $b$  at position  $i$  and  $p_b$  is the background frequency of base  $b$  (usually 25% as neutral distribution across the genome is assumed).

Additionally, other significant efforts have been devoted towards enhancing the power of the PWM in order to better discriminate between real binding sites and the background e.g. random data or non-regulatory regions (**Figure 1b**). In this direction, Gershenzon et al. [49] proposed 16–row matrices to replace the 4–row PWMs; Sandelin et al. [50] tried to classify TFBSs into TF families based on the constrained binding sequence diversity for groups of structurally related TFs to create familial binding profiles; Hannenhalli et al. [51] computationally divided the binding site collection of a TF into two subsets corresponding to two-child PWMs to increase the binding specificity of TF profiles. As earlier noted, however, the short length of the binding sites makes them appear fairly redundant and predictive methods are often replete with false positives. Therefore, given that the main question concerns the actual identification of TFBSs and effective the location of the promoter, searching becomes a more critical issue than simply optimizing the representation.

### 3. Promoter identification

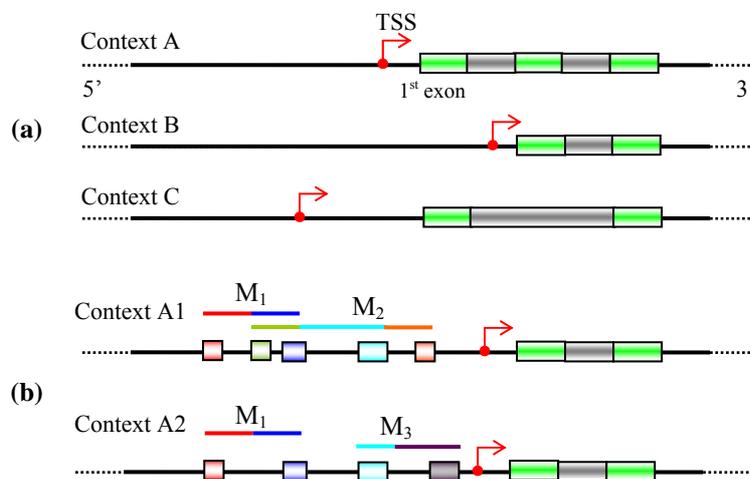
The first step towards discovering TFBSs is identifying the set of promoters. In principle, they are defined as the upstream regions proximal to the transcription start sites (TSSs) of genes; however, their length is still not clearly defined among different studies although it is one of the most important factors affecting to the computational predictions. Numerous activities have been proposed such as the recent experiment known as genome-wide open chromatin map that integrates high-throughput sequencing and genome-wide tiled array technologies has been performed to identify DNase I hypersensitive sites within human primary CD4<sup>+</sup> T cells [52]. Such activities aim at better defining proximal promoter lengths which are subsequently incorporated in commercial tools, such as [53].

Besides experimentally identified promoters, a number of computational methods have been proposed to predict promoter regions. Available tools include PromoterInspector [54], DragonGSF [55], EnSemPro [56], and have all been thoroughly reviewed [57, 58]. Prediction tools can be classified into two main categories, signal-based approaches which rely on conserved signals relevant to promoters, e.g. TATA box, CAAT box, CpG islands, and content-based approaches that utilize conserved motifs to distinguish between promoters and non-promoter regions [59]. Several models have been shown to be promising but due to the complexities of the genome structure, large-scale predictions are still difficult [60].

The structure of promoters, especially in mammals, is a complex which can be considered as a mini-structure of a gene where regulatory elements are interspersed within a large number of regions

non-conserved and unknown function [60]. Traditionally, it has been assumed that the combinatorial interaction of multiple transcription factors with the gene promoter is sufficient to explain the process of transcription. However, recent studies provided results to show that a large proportion of mammalian genes possess multiple transcription start sites (TSSs) and thus multiple promoters driving gene expression in a context-specific manner [61-63]. Specifically, in a recent study Singer et al. [15] developed and employed a custom microarray platform to show that nearly 35,000 alternative putative promoters are present on around 7,000 human genes. Furthermore, each set of unique combination of TFBSs in the promoter will determine its temporal and spatial expression in a specific context [60] (**Figure 2**). These observations significantly increase the complexity of understanding gene regulation and the transcription process in general, and create a huge challenge for both computational and experimental methodologies of TFBS identification.

**Figure 2.** Data complexities in TFBS prediction. **(a)** Alternative promoters usually occur for genes in higher eukaryotes e.g. nearly 35,000 alternative putative promoters are present on around 7,000 human genes [15]. For a specific gene, different promoters are activated to drive the gene expression in different corresponding contexts. **(b)** Alternative sets of combinatorial TFs regulate the transcription process even though only one promoter is activated in these contexts.  $M_1$ ,  $M_2$ ,  $M_3$  are three example transcriptional modules (a set of TFs or corresponding TFBSs) activated to regulate the transcription process; module  $M_1$  is present on two cases whereas only a part of  $M_2$  is functional in the other case e.g. human RANTES/CCL5 gene consists of different set of functional TFBSs in different cell types [60].



#### 4. Discovery of physical TFBSs

One of the first questions related to TFBS identification would be how to detect a conserved motif in a given set of sequences. The problem can be simply stated as follows: given a set of  $N$  sequences  $S = \{s_i\}_{i=1}^N$ ,  $s_i = \{s_{il}\}_{l=1}^{|s_i|}$ ,  $s_{il} \in A$ ,  $A = \{A, C, G, T\}$ , identify conserved motifs  $p = \{p_k\}_{k=1}^K$ ,  $p_k \in A$  that are overrepresented, i.e. motifs present in  $S$  at a statistical significant rate. The fundamental assumption is that if the sequences are promoters of genes, then conserved motifs can be assumed to be potential binding sites for TFs.

There have been a wide range of possible applications for such *in silico* motif discovery methods. First, they greatly assist experimental studies aiming towards detection of the collection of binding sites for a given TF [32]. ChIP-chip assays, for example, identify genomic regions to which a TF of interest binds. However, locating exact sites where the TF binds might be very difficult due to the limitations of the assays. As a result, once the DNA sequences to which the TF binds have been collected motif discovery algorithms, e.g. consensus [64], Gibbs sampling [65], MEME [66], are then applied to locate the exact binding sites. Secondly, if one identifies a set of genes that can be considered as regulated by some common TF(s), then one can begin to search computationally for conserved motifs in the corresponding promoter to infer regulating TFs. The underlying assumption of such a computation is that the common patterns are the likely functional ones. Furthermore, motif discovery algorithms can also assist in cross-species extrapolation to improve the specificity of finding TFBSs on a gene promoter. Once a set of corresponding promoters of a gene across multiple species have been extracted, motif discovery algorithms are used to detect conserved sub-sequences in this promoter across species in an attempt to identify all potential *cis*-regulatory elements (discussed more details in the next section).

**Table 1.** Selected resources and relevant tools for *in silico* TFBS identification.

<b>Genome Browsers</b>			
UCSC	<a href="http://genome.ucsc.edu">genome.ucsc.edu</a>	VISTA	<a href="http://genome.lbl.gov/vista">http://genome.lbl.gov/vista</a>
<b>Promoter resources</b>			
<i>Databases</i>		<i>Prediction Tools</i>	
Genomatix	<a href="http://genomatix.de/products/Gene2Promoter">genomatix.de/products/Gene2Promoter</a>	PromoterInspector	<a href="http://genomatix.de/promoterinspector.html">genomatix.de/promoterinspector.html</a>
CSHL	<a href="http://rulai.cshl.edu/CSHLmpd2">rulai.cshl.edu/CSHLmpd2</a>	DragonGSF	<a href="http://research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm">research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm</a>
DBTSS	<a href="http://dbtss.hgc.jp">dbtss.hgc.jp</a>	Eponine	<a href="http://www.sanger.ac.uk/Users/td2/eponine">www.sanger.ac.uk/Users/td2/eponine</a>
EPD	<a href="http://www.epd.isb-sib.ch">www.epd.isb-sib.ch</a>	FirstEF	<a href="http://rulai.cshl.org/tools/FirstEF">rulai.cshl.org/tools/FirstEF</a>
<b>Transcription factor resources</b>			
<i>PWM databases</i>		<i>Phylogenetic footprinting tools</i>	
Genomatix	<a href="http://genomatix.de/products/MatBase">genomatix.de/products/MatBase</a>	FootPrinter	<a href="http://bio.cs.washington.edu/software.html#footprinter">bio.cs.washington.edu/software.html#footprinter</a>
TRANSFAC	<a href="http://www.gene-regulation.com/pub/databases.html">www.gene-regulation.com/pub/databases.html</a>	PhyloME	<a href="http://bio.cs.washington.edu/software.html#phyme">bio.cs.washington.edu/software.html#phyme</a>
JASPAR	<a href="http://jaspar.cgb.ki.se">jaspar.cgb.ki.se</a>	PhyloGibbs	<a href="http://www.phylogibbs.unibas.ch/cgi-bin/phylogibbs.pl">www.phylogibbs.unibas.ch/cgi-bin/phylogibbs.pl</a>
		PhyloGibbs-MP	<a href="http://www.imsc.res.in/~rsidd/phylogibbs-mp">www.imsc.res.in/~rsidd/phylogibbs-mp</a>
		MONKEY	<a href="http://rana.lbl.gov/monkey">rana.lbl.gov/monkey</a>
<i>Single-motif discovery tools</i>		<i>Cis-regulatory module discovery tools</i>	
MatInspector	<a href="http://genomatix.de/products/MatInspector">genomatix.de/products/MatInspector</a>	FrameWorker	<a href="http://genomatix.de/frameworker.html">genomatix.de/frameworker.html</a>
P-Match	<a href="http://www.gene-regulation.com/pub/programs.html">www.gene-regulation.com/pub/programs.html</a>	CMA	<a href="http://www.gene-regulation.com/pub/programs.html">www.gene-regulation.com/pub/programs.html</a>
AlignACE	<a href="http://atlas.med.harvard.edu">atlas.med.harvard.edu</a>	CisModule	<a href="http://www.stat.ucla.edu/~zhou/CisModule">www.stat.ucla.edu/~zhou/CisModule</a>
Consensus	<a href="http://bifrost.wustl.edu/consensus">bifrost.wustl.edu/consensus</a>	CisPlusFinder	<a href="http://jakob.genetik.uni-koeln.de/bioinformatik/people/nora/nora.html">jakob.genetik.uni-koeln.de/bioinformatik/people/nora/nora.html</a>
MEME	<a href="http://meme.sdsc.edu">meme.sdsc.edu</a>	DiRE	<a href="http://dire.dcode.org">dire.dcode.org</a>

Because of the importance of this problem, a variety of algorithms as well as computational tools have been developed for those problems above for the past twenty years (**Table 1**). However, generally speaking the core algorithms can be classified into two categories: combinatorial and probabilistic [26,

30, 67]. Exhaustive search with pattern-based scoring (combinatorial category) is the starting point of discovering conserved motifs in a set of promoter sequences [67]. Due to magnitude of the search space, methods were further improved by exploring sequence-based exhaustive search [68] and also consensus search [69]. The probabilistic-based methods employ two main algorithms e.g. Gibbs sampling [65] and MEME [66] and have also been used extensively for motif discovery tools. The basic idea is to continuously reduce the search space and the false positive matches by more accurately representing the motif models.

However, it is important to realize that although a large number of TFs has already been identified, and more are being identified, through numerous high-throughput activities emanating from the decoding of the human, *in silico* analysis is further hindered by the fact that only a fraction of those can currently be mapped to known and well characterized profiles [53, 70, 71] (around 600 human TFs in [www.genomatix.de](http://www.genomatix.de) vs. approximately 1,850 TFs found in human [72]). When conserved motifs are predicted computationally that are not present in available collections, these are then considered as novel binding sites and/or regulatory regions but they are set aside for further investigation. Therefore, besides such motif discovery methods, another approach to detect potential TFBSs is directly scanning known TF profiles and scoring to determine whether or not the matches are potential binding sites.

Given that the scoring metric would assign relative importance to alternative binding sites in motif discovery methods [29, 73, 74], it is of equal importance to score directly the subsequences of interest in terms of their potential of being binding sites compared to known TF profiles. Despite the large number of alternative representation models and their associated scoring function, the most widely-used approach is still the one based on the PWM model and the sum fitness function, as discussed above. Given, therefore, that the sum fitness is used, which based on the relative abundance of bases in a specific position based on scanning the TF profiles, the strategy to predict whether or not a site is a binding site is among the most critical factors. Therefore, major emphasis is placed on developing strategies that score a candidate oligo and identify the thresholds for the prediction. A typical approach is based on core similarity matches (**Figure 1a**) to reduce the number of false positive matches [47]. Furthermore, the threshold for each PWM is optimized so that a maximum of three matches are allowed in 10,000bp of non-regulatory test sequences (coding sequences excluding first exons and genomic repeats). This is the approach used in tool MatInspector in Genomatix [47]. As an alternative strategy, [48] implemented P-Match in TRANSFAC to select the optimized thresholds so that the false positive rate is minimum and/or the false negative rate reaches some user-defined threshold. The threshold for minimum false positive rate is the one at which no match is found on the background set of exon sequences; and the threshold for false negative rate  $\alpha$  is the rate at which  $\alpha\%$  of binding sites in the collection used to build the TF profile are not detected by that threshold using leave-one-out cross validation. Besides determining is the magnitude of a score threshold, both approaches also make use of the concept of TF family profiles [50, 51] with some variations to reduce the redundant matches in scanning TF profiles on a promoter sequence. Generally speaking, the key idea here is using prior knowledge such as known TF profiles to predict the most probable TFBSs on promoter sequences with a minimum false positive matches; for example, those PWMs that represent similar DNA patterns will be assigned into the same TF family [47].

## 5. Discovery of functional relevant TFBSs

All the above approaches focus on identifying either experimentally or computationally putative binding sites. Regardless of the approach used, however, physical binding does not necessarily imply functional activity. As such, a major question concerns the functional characterization of the putative binding sites. Although, one can never be certain of the true activity of a TF, unless an appropriate experiment is conducted, computational approaches aim at reducing the number of alternatives on which further experimentation is conducted. Therefore, the next critical question to explore is whether we can identify those TFBS that are more likely to be functional among a set of candidates. Probably as expected, this has turned to be a very challenging question, particularly in higher eukaryotes. However, it has also served as an endless source inspiration for developing numerous computational strategies.

Before delving into the specific details of the computational approaches, it is instructive to classify the putative functional binding sites into two major categories: (a) general vs. context specific functional binding sites. The former aims at identifying *cis*-regulatory elements of a single gene that are conserved across multiple species based on the evolutionary hypothesis (so-called phylogenetic footprinting) whereas the latter searches for overrepresented TFBSs across multiple genes of a single species that share common characteristics in a specific context e.g. co-expression and/or co-function.

### 5.1 Phylogenetic footprinting

With the advent of novel high-throughput technologies, a large number of genomic sequences of different species have been sequenced, catalogued and annotated, making it possible to explore the conserved information among orthologous genes in an effort to enhance TFBS prediction. The basic underlying assumption of comparative genomics, or phylogenetic footprinting, is that functional regions evolve under constraints and thus at a lower rate than non-functional regions. Therefore, it is hypothesized that well conserved regions in a set of orthologous sequences survived due to their special functional implications, making them become promising candidates for functional *cis*-regulatory elements [75]. Preliminary evidence seems to support the hypothesis that conservation does imply so kind of, yet to be determined, significance. For instance, Cliften et al. [76] sequenced six *Saccharomyces* species and verified that many TFBSs are conserved across species and also located in conserved blocks although the blocks are often times much longer than the binding sites. Similarly, Gibbs et al. [77] demonstrated that regions with high-scoring PWM matches that are conserved across human-mouse-rat genomic alignment provided a 44-fold increase in the specificity of the predictions compared to those that are not conserved. Therefore, utilizing the information from orthologous genes across multiple species is becoming a useful paradigm in predicting putatively functional binding sites as well as reducing the false positive matches in motif discovering methods.

Now given a gene of interest, one begins constructing a global [78, 79] or local multiple sequence alignment [80-82] of orthologous promoter sequences and then identifies conserved regions which are considered as regulatory regions of the gene. However, all these tools assume that all nucleotides are alike or use a well-established substitution matrix to penalize the insertion, deletion, or substitution in the alignment, and thus they may not align properly non-coding DNA sequences of orthologous genes

[82]. Another critical limitation is that for closely related species, the alignment is obvious but impossible to distinguish functional elements from the surrounding non-functional regions like the case of four *Saccharomyces* species from the sensu stricto group, two species from the sensu lato group and one petite-negative species for example [83]. On the other hand, when the species are too far apart evolutionarily as the case of *rbcS* gene in 10 plants shown span ~ 760 million years of evolution, only 3 conserved sites that are known regulatory elements each of 9 bp long are present in around 500 bp in the 5' upstream regions [84]. Therefore, given that *cis*-regulatory elements are short, degenerate fragments, and present on such a large number of non-functional, diverged regions, regular sequence alignment methods are usually failed to detect properly these short conserved regions [85].

To overcome the problems associated with alignment, motif discovery algorithms e.g. Consensus, Gibbs sampling, MEME have been utilized along with the set of orthologous promoter sequences as the input data, [83, 86]. As a result, it is more likely to have a functional binding site as well as reducing false positive matches if some TF profile is located on a conserved region in the set of promoter sequences; especially, if the region is conserved among sequences from distantly related species. As such the match would be subject to selection pressure and more likely to be functional. A number of available tools such as PhyloCon [87] and CONREAL [88], explored those ideas.

However, in the aforementioned approaches phylogenetic relationships of the given sequences are not explored. Therefore, results are still highly biased to favor relationships between sequences located on closely related species [85]. A series of later models have attempted to incorporate the phylogenetic information into the search strategy, including the phylogenetic relationships between sequences and/or the binding site evolution model. Specifically, FootPrinter [85, 89] used a standard phylogenetic tree to estimate the significance of each conserved motif; EMnEM [90] applied the Jukes-Cantor (JC) model [91] with a fixed substitution rate for the evolutionary model of regulatory elements; PhyME [92, 93] and PhyloGibbs [94, 95] used a model suggested by Sinha et al. [96] which is similar to Felsenstein's molecular evolution model [97] to model the binding site evolution; MONKEY [98] employed Felsenstein's molecular evolution model [97] and allows users to select between the JC [91] and HKY [99] models for the background. Finally, Gertz et al. [100] proposed a model that employed a more detailed evolution model for binding sites based on the work of [99]. Besides such tools that find conserved regions in a set of orthologous promoters independently with known TF profiles, some attempt has been made to incorporate both into a single search method e.g. PhyloScan [101].

## 5.2 Context-specific search

While it is recognized that not all binding sites found on a promoter will be functional elements, it is also recognized that functional sites are not activated simultaneously or independently of condition, or environment, since the cooperation of TFs is highly dependent on context [102-106]. Human RANTES/CCL5, a member of the CC- or  $\beta$ -subfamily chemotactic cytokines for instance, appears to have six functionally characterized short regulatory elements on its promoter that mediate its transcription initiation. However, not all six elements are activated simultaneously in any specific tissue in five cell types analyzed and the elements are also highly selective under different stimulating signals regulating gene expression [107]. Consequently, a critical question is to establish a relationship between binding sites and the context in which these sites become functional. The term 'context' here

is used in a way that implicitly refers to a set of potentially co-regulated genes e.g. genes that appear either to exhibit correlation in their expression patterns or to be involved in similar functions in a specific condition and/or tissue [103, 108, 109]. Two elements become critical in this direction: (i) knowledge of the set of potentially co-regulated genes, and (ii) the context-specific nature of functionality.

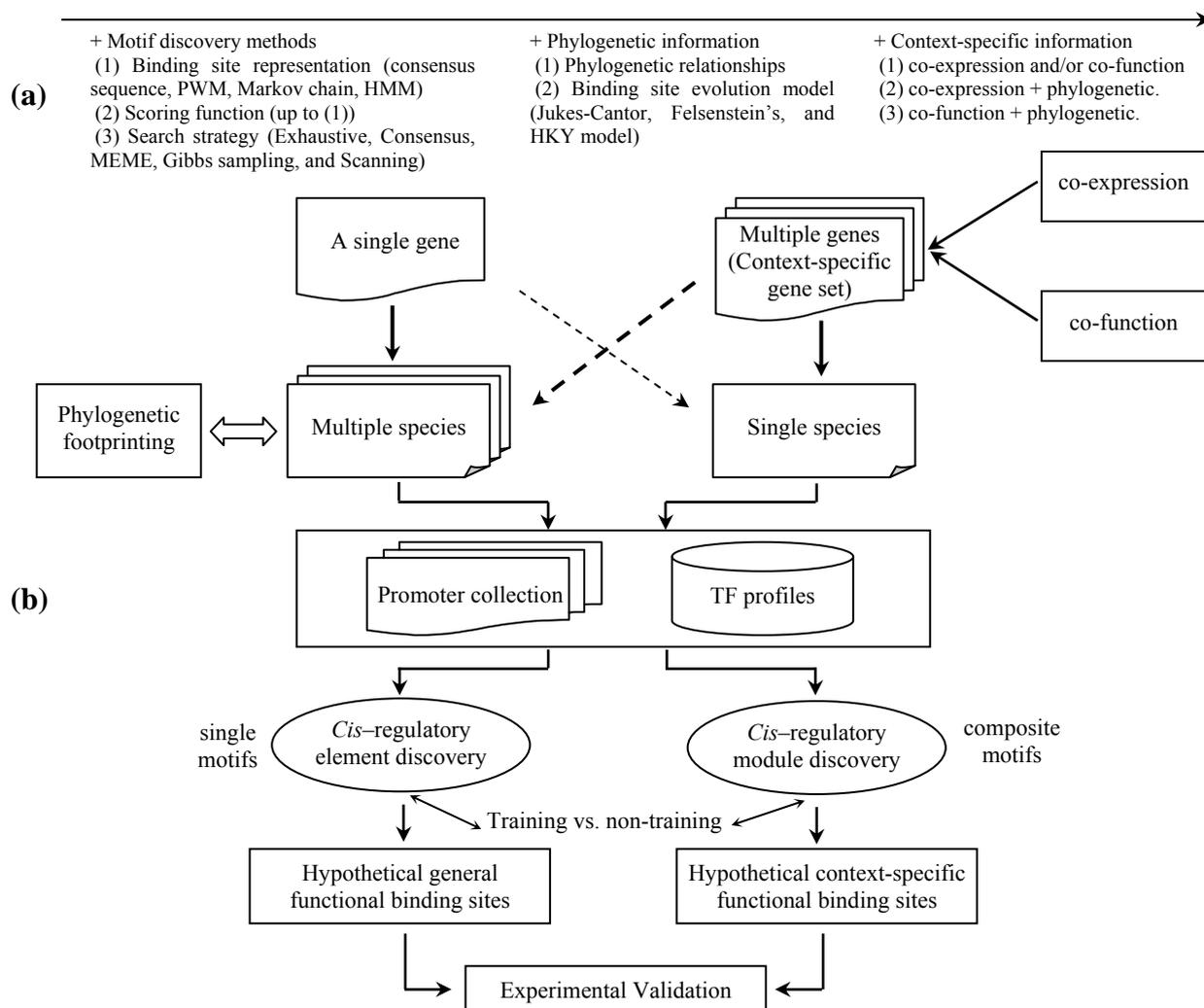
We must, however, realize that TFs in higher organisms are more likely to cooperate with nearby bound factors in a combinatorial manner to regulate the transcription process rather than function isolation. As an example, gene *even skipped* (*eve*) is known to be regulated by at least five TFs (Twi, Tin, dTcf, Mad, and Pnt) binding to a 312bp MHE enhancer located ~6kbp downstream of the *eve* coding region. The corresponding binding sites of these TFs form a *cis*-regulatory modules that has been shown to occur significantly less than randomly expected and also more likely responsible for the regulation of other genes if present on their promoters. Therefore, the field has been shifting from the single motif detection to the discovery of composite motifs i.e. *cis*-regulatory modules. From a computational point of view, the concept of *cis*-regulatory modules also helps to reduce the number of false positive matches and make the search strategy more efficient. A *cis*-regulatory module is in general defined as a *cis*-regulatory element which is the smallest functional unit to have a biological role in transcriptional regulation [53]. It consists of a set of individual binding sites of TFs on the proximal promoter region of a gene. A module is mainly characterized by two factors: composition and structural constraints. Composition is a set of non-overlapping binding sites, whereas structural constraints are the strand orientation to which the corresponding TF binds, the order and the distance between binding sites [110]. A variety of tools have been developed to search for modules in a set of promoter sequences without taking into account the structural constraints (also called composite motifs) e.g. Cluster-Buster [111], CisModule [112], MSCAN [113], CisPlusFinder [114], ModuleMiner [115], DiRE [116]. Alternatively, when *cis*-regulatory modules are associated with their structural constraints (now so-called TF-modules), it refers to methods that detect TFBSs using known TF profiles (FrameWorker [110], CMA [117]).

The main idea in this direction is to use prior knowledge to identify the set of potentially co-regulated genes and then search the corresponding promoter set for common and/or significant *cis*-regulatory modules (**Figure 3**). Earlier studies assumed that a cluster of coexpressed genes could be under the same regulatory mechanism, e.g. co-regulation [118, 119] or co-function [120]. However, more recent evidence suggests that co-expression alone is not enough to infer the existence of common regulatory mechanisms and instead additional information is required [108, 121], especially in higher organisms. Specifically, recent studies have shown that genes sharing similar expression patterns can participate in a number of different biological functions and/or genes in the same pathway can exhibit different patterns of expression [122, 123]. Moreover, the underlying gene regulation is shown to be tissue and/or condition specific and the TFs that drive the gene expression are very flexible in function and activity under different conditions [103-106]. Therefore, defining the context in which a set of genes are more likely to be co-regulated poses a formidable challenge to researchers.

A number of assumptions have been suggested and preliminary results appear promising. For example, Segal et al. [124] attempted to generate testable hypotheses like ‘regulator X regulates co-expression module Y under conditions W’ with the assumption that co-expressed genes across a set of conditions will be co-regulated. The work was done on yeast with some cases experimentally verified

successfully. Similarly, Elkon et al. [125] analyzed a set of coexpressed gene that are cell-cycle-dependent and found eight TFs whose binding sites are significantly overrepresented in their promoters, or Long et al. [109] identified statistically overrepresented cooperating TFBSs in the 1 kbp upstream sequence set of each biological-process gene group in GO.

**Figure 3. (a)** A brief overview of how computational models are developed. Motif discovery methods consist of three main components: the binding site representation, the scoring function, and the search strategy. There are alternative approaches for searching novel motifs, including scanning for TFs with known profiles. Phylo- tools incorporate phylogenetic relationships among species or their corresponding promoter sequences, and the binding site evolution model to improve the search strategies. Besides, context-specific information can also be explored to predict functional binding sites and then infer a set of context-relevant TFs for further analysis. **(b)** Different strategies to predict *cis*-regulatory elements using additional biological knowledge. Two main strategies exist: single gene, multiple species and single motif discovery methods vs. multiple genes, single species and *cis*-regulatory module discovery methods. The concept of *cis*-regulatory modules was introduced to capture the biological aspect as well as enhance the specificity of the search, especially in higher organisms; besides, other combinatorial strategies have also been developed in the literature (dash arrows).



Besides the two main strategies for discovery functional binding sites, i.e. single gene, multiple species, single motifs for detecting general functional binding sites; and multiple genes, single species, composite motifs plus context for predicting context-specific functional binding sites as well as relevant condition-specific activated TFs, other combinatorial strategies have also been developed in the literature. For example, combining co-expressed genes with phylogenetic footprinting for single motif discovery [87, 126-128] or for composite motif discovery [129]. On the other hand, some studies require a collection of promoter sequences with known module sites to serve as training data for building predictive models [130-133]. These approaches seem to be more accurate both in detecting single TFBSs [42, 46] and in predicting TF-modules [132, 133]. However, these models have to be built carefully, cannot be easily inferred automatically, and a number of parameters need to be determined using training data.

## 6. Inference of transcriptional regulatory networks

Automatic inference of regulatory networks is an essential step in bridging the gap between the raw expression data and the mechanistic understanding at the molecular level. Better predictions of such networks will find widespread application towards efforts to delineate the impact of external stimuli on cellular responses. Although gene expression can reveal some part of the picture, such results are often difficult to interpret without an understanding of relevant pathways and networks [134]. Promoter analysis provides suggestions to which TFs are relevant to the response as well as sketch out a preliminary picture of the interplays between TFs and gene promoters that orchestrate the gene expression of thousands of genes due to changes in the environment. With the assumption that if a corresponding binding site of TF A is present on the promoter region of gene B, or statistically over-represented on a gene set B, B will be considered as regulated by A, while the regulation can be either activation or repression. Several computational tools such as PAINT [135, 136], CARRIE [137, 138] have been developed to automatically produce the transcriptional regulatory network given a set of genes. Although much work needs to be done, these tools can provide preliminary testable hypothesis.

## 7. Concluding Remarks

In this review, we have summarized the current state in characterizing promoter sequences for the search for putative transcription factor binding sites. We addressed the elements associated with the representation and mining of the genomic information and characterized the basic methods, algorithms and computational tools. Future success of such endeavors is expected to have major impact on biological and clinical applications. However, a major point of concern and a most critical open question refers to the possibility of establishing *de novo* link between putative binding sites and actually functional binding sites. Function prediction from sequence information is an open question in a number of areas of computational biology and transcription regulation is no exception. Regardless of our present inability to establish that link, the availability of methods, like the ones described in this review, are of paramount important as they allow for the systematic generation of critical testable hypotheses.

## Acknowledgements

The authors acknowledge financial support from the National Institutes of Health (R01GM082974), the National Science Foundation (NSF-BES 0519563) and the EPA (GAD R 832721-010).

## References and Notes

1. Kafatos, F.C. A revolutionary landscape: the restructuring of biology and its convergence with medicine. *J Mol Biol* **2002**, *319*(4), 861-867.
2. Lemon, B.; Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* **2000**, *14*(20), 2551-2569.
3. Levine, M.; Tjian, R. Transcription regulation and animal diversity. *Nature* **2003**, *424*(6945), 147-151.
4. van Driel, R.; Fransz, P.F.; Verschure, P.J. The eukaryotic genome: a system regulated at different hierarchical levels. *J Cell Sci* **2003**, *116*(Pt 20), 4067-4075.
5. Werner, T.; Fessele, S.; Maier, H.; Nelson, P.J. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *Faseb J* **2003**, *17*(10), 1228-1237.
6. Cooper, S.J.; Trinklein, N.D.; Anton, E.D.; Nguyen, L.; Myers, R.M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* **2006**, *16*(1), 1-10.
7. Maston, G.A.; Evans, S.K.; Green, M.R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **2006**, *7*, 29-59.
8. Heintzman, N.D.; Ren, B. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol Life Sci* **2007**, *64*(4), 386-400.
9. Barrera, L.O.; Ren, B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* **2006**, *18*(3), 291-298.
10. Dillon, N. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* **2006**, *14*(1), 117-126.
11. Mateos-Langerak, J.; Goetze, S.; Leonhardt, H.; Cremer, T.; van Driel, R.; Lanctot, C. Nuclear architecture: Is it important for genome function and can we prove it? *J Cell Biochem* **2007**, *102*(5), 1067-1075.
12. Schneider, R.; Grosschedl, R. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev* **2007**, *21*(23), 3027-3043.
13. Wray, G.A.; Hahn, M.W.; Abouheif, E.; Balhoff, J.P.; Pizer, M.; Rockman, M.V.; Romano, L.A. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **2003**, *20*(9), 1377-1419.
14. Landry, J.R.; Mager, D.L.; Wilhelm, B.T. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* **2003**, *19*(11), 640-648.
15. Singer, G.A.; Wu, J.; Yan, P.; Plass, C.; Huang, T.H.; Davuluri, R.V. Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics* **2008**, *9*, 349.

16. Sandve, G.K.; Drablos, F. A survey of motif discovery methods in an integrated framework. *Biol Direct* **2006**, *1*, 11.
17. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol* **2003**, *5*(1), 201.
18. Qi, Y.; Rolfe, A.; MacIsaac, K.D.; Gerber, G.K.; Pokholok, D.; Zeitlinger, J.; Danford, T.; Dowell, R.D.; Fraenkel, E.; Jaakkola, T.S.; Young, R.A.; Gifford, D.K. High-resolution computational models of genome binding events. *Nat Biotechnol* **2006**, *24*(8), 963-970.
19. Ren, B.; Robert, F.; Wyrick, J.J.; Aparicio, O.; Jennings, E.G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T.L.; Wilson, C.J.; Bell, S.P.; Young, R.A. Genome-wide location and function of DNA binding proteins. *Science* **2000**, *290*(5500), 2306-2309.
20. Roulet, E.; Busso, S.; Camargo, A.A.; Simpson, A.J.; Mermod, N.; Bucher, P. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **2002**, *20*(8), 831-835.
21. Stoltenburg, R.; Reinemann, C.; Strehlitz, B. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng* **2007**, *24*(4), 381-403.
22. Hu, J.; Li, B.; Kihara, D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* **2005**, *33*(15), 4899-4913.
23. Sandve, G.K.; Abul, O.; Walseng, V.; Drablos, F. Improved benchmarks for computational motif discovery. *BMC Bioinformatics* **2007**, *8*, 193.
24. Tompa, M.; Li, N.; Bailey, T.L.; Church, G.M.; De Moor, B.; Eskin, E.; Favorov, A.V.; Frith, M.C.; Fu, Y.; Kent, W.J.; Makeev, V.J.; Mironov, A.A.; Noble, W.S.; Pavese, 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.
25. Klepper, K.; Sandve, G.K.; Abul, O.; Johansen, J.; Drablos, F. Assessment of composite motif discovery methods. *BMC Bioinformatics* **2008**, *9*, 123.
26. Das, M.K.; Dai, H.K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* **2007**, *8* Suppl 7, S21.
27. Kato, M.; Hata, N.; Banerjee, N.; Futcher, B.; Zhang, M.Q. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol* **2004**, *5*(8), R56.
28. Wang, J. A new framework for identifying combinatorial regulation of transcription factors: a case study of the yeast cell cycle. *J Biomed Inform* **2007**, *40*(6), 707-725.
29. 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.
30. Pavese, G.; Mauri, G.; Pesole, G. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform* **2004**, *5*(3), 217-236.
31. Wasserman, W.W.; Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **2004**, *5*(4), 276-287.
32. Elnitski, L.; Jin, V.X.; Farnham, P.J.; Jones, S.J. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res* **2006**, *16*(12), 1455-1464.

33. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **1985**, *13*(9), 3021-3030.
34. Stormo, G.D. Consensus patterns in DNA. *Methods Enzymol* **1990**, *183*, 211-221.
35. Quandt, K.; Frech, K.; Karas, H.; Wingender, E.; Werner, T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **1995**, *23*(23), 4878-4884.
36. Chambers, A.; Stanway, C.; Tsang, J.S.; Henry, Y.; Kingsman, A.J.; Kingsman, S.M. ARS binding factor 1 binds adjacent to RAP1 at the UASs of the yeast glycolytic genes PGK and PYK1. *Nucleic Acids Res* **1990**, *18*(18), 5393-5399.
37. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **2000**, *16*(1), 16-23.
38. Kel, A.E.; Gossling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **2003**, *31*(13), 3576-3579.
39. Salzberg, S.L. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput Appl Biosci* **1997**, *13*(4), 365-376.
40. Bulyk, M.L.; Johnson, P.L.; Church, G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* **2002**, *30*(5), 1255-1261.
41. Man, T.K.; Stormo, G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* **2001**, *29*(12), 2471-2478.
42. Ellrott, K.; Yang, C.; Sladek, F.M.; Jiang, T. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics* **2002**, *18 Suppl 2*, S100-109.
43. Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **1997**, *268*(1), 78-94.
44. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
45. 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.
46. Ben-Gal, I.; Shani, A.; Gohr, A.; Grau, J.; Arviv, S.; Shmilovici, A.; Posch, S.; Grosse, I. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **2005**, *21*(11), 2657-2666.
47. 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.
48. Chekmenev, D.S.; Haid, C.; Kel, A.E. P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res* **2005**, *33*(Web Server issue), W432-437.

49. Gershenzon, N.I.; Stormo, G.D.; Ioshikhes, I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res* **2005**, *33*(7), 2290-2301.
50. Sandelin, A.; Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* **2004**, *338*(2), 207-215.
51. Hannenhalli, S.; Wang, L.S. Enhanced position weight matrices using mixture models. *Bioinformatics* **2005**, *21 Suppl 1*, i204-212.
52. Boyle, A.P.; Davis, S.; Shulha, H.P.; Meltzer, P.; Margulies, E.H.; Weng, Z.; Furey, T.S.; Crawford, G.E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **2008**, *132*(2), 311-322.
53. Genomatix <http://www.genomatix.de/>.
54. 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.
55. Bajic, V.B.; Seah, S.H. 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.
56. Won, H.H.; Kim, M.J.; Kim, S.; Kim, J.W. EnsemPro: an ensemble approach to predicting transcription start sites in human genomic DNA sequences. *Genomics* **2008**, *91*(3), 259-266.
57. Bajic, V.B.; Brent, M.R.; Brown, R.H.; Frankish, A.; Harrow, J.; Ohler, U.; Solovyev, V.V.; Tan, S.L. Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol* **2006**, *7 Suppl 1*, S3 1-13.
58. Pedersen, A.G.; Baldi, P.; Chauvin, Y.; Brunak, S. The biology of eukaryotic promoter prediction—a review. *Comput Chem* **1999**, *23*(3-4), 191-207.
59. Qiu, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun* **2003**, *309*(3), 495-501.
60. Werner, T. The state of the art of mammalian promoter recognition. *Brief Bioinform* **2003**, *4*(1), 22-30.
61. Davuluri, R.V.; Suzuki, Y.; Sugano, S.; Plass, C.; Huang, T.H. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **2008**, *24*(4), 167-177.
62. Kapranov, P.; Willingham, A.T.; Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **2007**, *8*(6), 413-423.
63. Sandelin, A.; Carninci, P.; Lenhard, B.; Ponjavic, J.; Hayashizaki, Y.; Hume, D.A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **2007**, *8*(6), 424-436.
64. Hertz, G.Z.; Hartzell, G.W., 3rd; Stormo, G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* **1990**, *6*(2), 81-92.
65. Lawrence, C.E.; Altschul, S.F.; Boguski, M.S.; Liu, J.S.; Neuwald, A.F.; Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**, *262*(5131), 208-214.
66. Bailey, T.L.; 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.

67. Tung, N.T.; Yang, E.; Androulakis, I.P. Machine learning approaches in promoter sequence analysis; In *Machine Learning Research Progress*, Peters, H., Vogel, Mia, Eds.; Nova Science Publishers, Inc, 2008.
68. Marsan, L.; Sagot, M.F. 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.
69. Hertz, G.Z.; Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **1999**, *15*(7-8), 563-577.
70. Vlieghe, D.; Sandelin, A.; De Bleser, P.J.; Vleminckx, K.; Wasserman, W.W.; van Roy, F.; Lenhard, B. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* **2006**, *34*(Database issue), D95-97.
71. Wingender, E.; Dietze, P.; Karas, H.; Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **1996**, *24*(1), 238-241.
72. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; Gocayne, J.D.; Amanatides, P.; Ballew, R.M.; Huson, D.H.; Wortman, J.R.; Zhang, Q.; Kodira, C.D.; Zheng, X.H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P.D.; Zhang, J.; Gabor Miklos, G.L.; Nelson, C.; Broder, S.; Clark, A.G.; Nadeau, J.; McKusick, V.A.; Zinder, N.; Levine, A.J.; Roberts, R.J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T.J.; Higgins, M.E.; Ji, R.R.; Ke, Z.; Ketchum, K.A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G.V.; Milshina, N.; Moore, H.M.; Naik, A.K.; Narayan, V.A.; Neelam, B.; Nusskern, D.; Rusch, D.B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M.L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferreira, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y.H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N.N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J.F.; Guigo, R.; Campbell, M.J.; Sjolander, K.V.; Karlak, B.; Kejariwal, A.; Mi H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y.H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly,

- M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. The sequence of the human genome. *Science* **2001**, *291*(5507), 1304-1351.
73. Friberg, M.; von Rohr, P.; Gonnet, G. Scoring functions for transcription factor binding site prediction. *BMC Bioinformatics* **2005**, *6*, 84.
74. Li, N.; Tompa, M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol* **2006**, *1*, 8.
75. Doniger, S.W.; Huh, J.; Fay, J.C. Identification of functional transcription factor binding sites using closely related *Saccharomyces* species. *Genome Res* **2005**, *15*(5), 701-709.
76. Cliften, P.; Sudarsanam, P.; Desikan, A.; Fulton, L.; Fulton, B.; Majors, J.; Waterston, R.; Cohen, B.A.; Johnston, M. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **2003**, *301*(5629), 71-76.
77. Gibbs, R.A.; Weinstock, G.M.; Metzker, M.L.; Muzny, D.M.; Sodergren, E.J.; Scherer, S.; Scott, G.; Steffen, D.; Worley, K.C.; Burch, P.E.; Okwuonu, G.; Hines, S.; Lewis, L.; DeRamo, C.; Delgado, O.; Dugan-Rocha, S.; Miner, G.; Morgan, M.; Hawes, A.; Gill, R.; Celera; Holt, R.A.; Adams, M.D.; Amanatides, P.G.; Baden-Tillson, H.; Barnstead, M.; Chin, S.; Evans, C.A.; Ferriera, S.; Fosler, C.; Glodek, A.; Gu, Z.; Jennings, D.; Kraft, C.L.; Nguyen, T.; Pfannkoch, C.M.; Sitter, C.; Sutton, G.G.; Venter, J.C.; Woodage, T.; Smith, D.; Lee, H.M.; Gustafson, E.; Cahill, P.; Kana, A.; Doucette-Stamm, L.; Weinstock, K.; Fechtel, K.; Weiss, R.B.; Dunn, D.M.; Green, E.D.; Blakesley, R.W.; Bouffard, G.G.; De Jong, P.J.; Osoegawa, K.; Zhu, B.; Marra, M.; Schein, J.; Bosdet, I.; Fjell, C.; Jones, S.; Krzywinski, M.; Mathewson, C.; Siddiqui, A.; Wye, N.; McPherson, J.; Zhao, S.; Fraser, C.M.; Shetty, J.; Shatsman, S.; Geer, K.; Chen, Y.; Abramzon, S.; Nierman, W.C.; Havlak, P.H.; Chen, R.; Durbin, K.J.; Egan, A.; Ren, Y.; Song, X.Z.; Li B.; Liu, Y.; Qin, X.; Cawley, S.; Worley, K.C.; Cooney, A.J.; D'Souza, L.M.; Martin, K.; Wu, J.Q.; Gonzalez-Garay, M.L.; Jackson, A.R.; Kalafus, K.J.; McLeod, M.P.; Milosavljevic, A.; Virk, D.; Volkov, A.; Wheeler, D.A.; Zhang, Z.; Bailey, J.A.; Eichler, E.E.; Tuzun, E.; Birney, E.; Mongin, E.; Ureta-Vidal, A.; Woodwark, C.; Zdobnov, E.; Bork, P.; Suyama, M.; Torrents, D.; Alexandersson, M.; Trask, B.J.; Young, J.M.; Huang, H.; Wang, H.; Xing, H.; Daniels, S.; Gietzen, D.; Schmidt, J.; Stevens, K.; Vitt, U.; Wingrove, J.; Camara, F.; Mar Alba, M.; Abril, J.F.; Guigo, R.; Smit, A.; Dubchak, I.; Rubin, E.M.; Couronne, O.; Poliakov, A.; Hubner, N.; Ganten, D.; Goesele, C.; Hummel, O.; Kreitler, T.; Lee, Y.A.; Monti, J.; Schulz, H.; Zimdahl, H.; Himmelbauer, H.; Lehrach, H.; Jacob, H.J.; Bromberg, S.; Gullings-Handley, J.; Jensen-Seaman, M.I.; Kwitek, A.E.; Lazar, J.; Pasko, D.; Tonellato, P.J.; Twigger, S.; Ponting, C.P.; Duarte, J.M.; Rice, S.; Goodstadt, L.; Beatson, S.A.; Emes, R.D.; Winter, E.E.; Webber, C.; Brandt, P.; Nyakatura, G.; Adetobi, M.; Chiaromonte, F.; Elnitski, L.; Eswara, P.; Hardison, R.C.; Hou, M.; Kolbe, D.; Makova, K.; Miller, W.; Nekrutenko, A.; Riemer, C.; Schwartz, S.; Taylor, J.; Yang, S.; Zhang, Y.; Lindpaintner, K.; Andrews, T.D.; Caccamo, M.; Clamp, M.; Clarke, L.; Curwen,

- V.; Durbin, R.; Eyraas, E.; Searle, S.M.; Cooper, G.M.; Batzoglou, S.; Brudno, M.; Sidow, A.; Stone, E.A.; Venter, J.C.; Payseur, B.A.; Bourque, G.; Lopez-Otin, C.; Puente, X.S.; Chakrabarti, K.; Chatterji, S.; Dewey, C.; Pachter, L.; Bray, N.; Yap, V.B.; Caspi, A.; Tesler, G.; Pevzner, P.A.; Haussler, D.; Roskin, K.M.; Baertsch, R.; Clawson, H.; Furey, T.S.; Hinrichs, A.S.; Karolchik, D.; Kent, W.J.; Rosenbloom, K.R.; Trumbower, H.; Weirauch, M.; Cooper, D.N.; Stenson, P.D.; Ma, B.; Brent, M.; Arumugam, M.; Shteynberg, D.; Copley, R.R.; Taylor, M.S.; Riethman, H.; Mudunuri, U.; Peterson, J.; Guyer, M.; Felsenfeld, A.; Old, S.; Mockrin, S.; Collins, F. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **2004**, *428*(6982), 493-521.
78. Brudno, M.; Do, C.B.; Cooper, G.M.; Kim, M.F.; Davydov, E.; Green, E.D.; Sidow, A.; Batzoglou, S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **2003**, *13*(4), 721-731.
79. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. 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.
80. Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **1999**, *15*(3), 211-218.
81. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **2000**, *302*(1), 205-217.
82. Siddharthan, R. Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. *BMC Bioinformatics* **2006**, *7*, 143.
83. Cliften, P.F.; Hillier, L.W.; Fulton, L.; Graves, T.; Miner, T.; Gish, W.R.; Waterston, R.H.; Johnston, M. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* **2001**, *11*(7), 1175-1186.
84. Tompa, M. Identifying functional elements by comparative DNA sequence analysis. *Genome Res* **2001**, *11*(7), 1143-1144.
85. Blanchette, M.; Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* **2002**, *12*(5), 739-748.
86. McCue, L.; Thompson, W.; Carmack, C.; Ryan, M.P.; Liu, J.S.; Derbyshire, V.; Lawrence, C.E. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* **2001**, *29*(3), 774-782.
87. Wang, T.; Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **2003**, *19*(18), 2369-2380.
88. Berezikov, E.; Guryev, V.; Plasterk, R.H.; Cuppen, E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res* **2004**, *14*(1), 170-178.
89. Blanchette, M.; Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* **2003**, *31*(13), 3840-3842.
90. Moses, A.M.; Chiang, D.Y.; Eisen, M.B. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput* **2004**, 324-335.
91. Jukes, T.H.C.R.C. Evolution of protein molecules. In *Mammalian protein metabolism*; Munro, H.N. (Ed.); Academic Press: New York, 1969; pp. 21-123.

92. Sinha, S. PhyME: a software tool for finding motifs in sets of orthologous sequences. *Methods Mol Biol* **2007**, *395*, 309-318.
93. Sinha, S.; Blanchette, M.; Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **2004**, *5*, 170.
94. Siddharthan, R. PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput Biol* **2008**, *4*(8), e1000156.
95. Siddharthan, R.; Siggia, E.D.; van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **2005**, *1*(7), e67.
96. Sinha, S.; van Nimwegen, E.; Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics* **2003**, *19 Suppl 1*, i292-301.
97. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **1981**, *17*(6), 368-376.
98. Moses, A.M.; Chiang, D.Y.; Pollard, D.A.; Iyer, V.N.; Eisen, M.B. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* **2004**, *5*(12), R98.
99. Hasegawa, M.; Kishino, H.; Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **1985**, *22*(2), 160-174.
100. Gertz, J.; Fay, J.C.; Cohen, B.A. Phylogeny based discovery of regulatory elements. *BMC Bioinformatics* **2006**, *7*, 266.
101. Carmack, C.S.; McCue, L.A.; Newberg, L.A.; Lawrence, C.E. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol* **2007**, *2*, 1.
102. Harbison, C.T.; Gordon, D.B.; Lee, T.I.; Rinaldi, N.J.; Macisaac, K.D.; Danford, T.W.; Hannett, N.M.; Tagne, J.B.; Reynolds, D.B.; Yoo, J.; Jennings, E.G.; Zeitlinger, J.; Pokholok, D.K.; Kellis, M.; Rolfe, P.A.; Takusagawa, K.T.; Lander, E.S.; Gifford, D.K.; Fraenkel, E.; Young, R.A. Transcriptional regulatory code of a eukaryotic genome. *Nature* **2004**, *431*(7004), 99-104.
103. Lee, H.G.; Lee, H.S.; Jeon, S.H.; Chung, T.H.; Lim, Y.S.; Huh, W.K. High-resolution analysis of condition-specific regulatory modules in *Saccharomyces cerevisiae*. *Genome Biol* **2008**, *9*, R2.
104. McCord, R.P.; Berger, M.F.; Philippakis, A.A.; Bulyk, M.L. Inferring condition-specific transcription factor function from DNA binding and gene expression data. *Mol Syst Biol* **2007**, *3*, 100.
105. Smith, A.D.; Sumazin, P.; Zhang, M.Q. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol* **2007**, *3*, 73.
106. Yu, X.; Lin, J.; Zack, D.J.; Qian, J. Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors. *BMC Bioinformatics* **2007**, *8*, 437.
107. Fessele, S.; Maier, H.; Zischek, C.; Nelson, P.J.; Werner, T. Regulatory context is a crucial part of gene function. *Trends Genet* **2002**, *18*(2), 60-63.
108. Allocco, D.J.; Kohane, I.S.; Butte, A.J. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **2004**, *5*, 18.
109. Long, F.; Liu, H.; Hahn, C.; Sumazin, P.; Zhang, M.Q.; Zilberstein, A. Genome-wide prediction and analysis of function-specific transcription factor binding sites. *In Silico Biol* **2004**, *4*(4), 395-410.

110. Frech, K.; Danescu-Mayer, J.; Werner, T. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J Mol Biol* **1997**, *270*(5), 674-687.
111. Frith, M.C.; Li, M.C.; Weng, Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* **2003**, *31*(13), 3666-3668.
112. Zhou, Q.; Wong, W.H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A* **2004**, *101*(33), 12114-12119.
113. Alkema, W.B.; Johansson, O.; Lagergren, J.; Wasserman, W.W. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res* **2004**, *32*(Web Server issue), W195-198.
114. Pierstorff, N.; Bergman, C.M.; Wiehe, T. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics* **2006**, *22*(23), 2858-2864.
115. Van Loo, P.; Aerts, S.; Thienpont, B.; De Moor, B.; Moreau, Y.; Marynen, P. ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol* **2008**, *9*(4), R66.
116. Gotea, V.; Ovcharenko, I. DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* **2008**, *36*(Web Server issue), W133-139.
117. Waleev, T.; Shtokalo, D.; Konovalova, T.; Voss, N.; Cheremushkin, E.; Stegmaier, P.; Kel-Margoulis, O.; Wingender, E.; Kel, A. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res* **2006**, *34*(Web Server issue), W541-545.
118. Roth, F.P.; Hughes, J.D.; Estep, P.W.; Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **1998**, *16*(10), 939-945.
119. Tavazoie, S.; Hughes, J.D.; Campbell, M.J.; Cho, R.J.; Church, G.M. Systematic determination of genetic network architecture. *Nat Genet* **1999**, *22*(3), 281-285.
120. Lockhart, D.J.; Winzler, E.A. Genomics, gene expression and DNA arrays. *Nature* **2000**, *405*(6788), 827-836.
121. Flintoft, L. Gene regulation: The many paths to coexpression. *Nature Reviews Genetics* **2007**, *8*, 827.
122. Choi, D.; Fang, Y.; Mathers, W.D. Condition-specific coregulation with cis-regulatory motifs and modules in the mouse genome. *Genomics* **2006**, *87*(4), 500-508.
123. Huang, R.; Wallqvist, A.; Covell, D.G. Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics* **2006**, *87*(3), 315-328.
124. Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Botstein, D.; Koller, D.; Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **2003**, *34*(2), 166-176.
125. Elkon, R.; Linhart, C.; Sharan, R.; Shamir, R.; Shiloh, Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* **2003**, *13*(5), 773-780.

126. Cora, D.; Herrmann, C.; Dieterich, C.; Di Cunto, F.; Provero, P.; Caselle, M. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics* **2005**, *6*, 110.
127. Defrance, M.; Touzet, H. Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics* **2006**, *7*, 396.
128. Monsieurs, P.; Thijs, G.; Fadda, A.A.; De Keersmaecker, S.C.; Vanderleyden, J.; De Moor, B.; Marchal, K. More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics* **2006**, *7*, 160.
129. Vandepoele, K.; Casneuf, T.; Van de Peer, Y. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol* **2006**, *7*(11), R103.
130. King, D.C.; Taylor, J.; Elnitski, L.; Chiaromonte, F.; Miller, W.; Hardison, R.C. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **2005**, *15*(8), 1051-1060.
131. Kolbe, D.; Taylor, J.; Elnitski, L.; Eswara, P.; Li, J.; Miller, W.; Hardison, R.; Chiaromonte, F. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* **2004**, *14*(4), 700-707.
132. Taylor, J.; Tyekucheva, S.; King, D.C.; Hardison, R.C.; Miller, W.; Chiaromonte, F. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* **2006**, *16*(12), 1596-1604.
133. Wang, H.; Zhang, Y.; Cheng, Y.; Zhou, Y.; King, D.C.; Taylor, J.; Chiaromonte, F.; Kasturi, J.; Petrykowska, H.; Gibb, B.; Dorman, C.; Miller, W.; Dore, L.C.; Welch, J.; Weiss, M.J.; Hardison, R.C. Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res* **2006**, *16*(12), 1480-1492.
134. Seifert, M.; Scherf, M.; Epple, A.; Werner, T. Multievidence microarray mining. *Trends Genet* **2005**, *21*(10), 553-558.
135. Gonye, G.E.; Chakravarthula, P.; Schwaber, J.S.; Vadigepalli, R. From promoter analysis to transcriptional regulatory network prediction using PAINT. *Methods Mol Biol* **2007**, *408*, 49-68.
136. Vadigepalli, R.; Chakravarthula, P.; Zak, D.E.; Schwaber, J.S.; Gonye, G.E. PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. *Omics* **2003**, *7*(3), 235-252.
137. Haverty, P.M.; Frith, M.C.; Weng, Z. CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res* **2004**, *32*(Web Server issue), W213-216.
138. Haverty, P.M.; Hansen, U.; Weng, Z. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res* **2004**, *32*(1), 179-188.