Identifying regulatory elements in eukaryotic genomes

Leelavati Narlikar and Ivan Ovcharenko

Advance Access publication date 4 June 2009

Abstract

Proper development and functioning of an organism depends on precise spatial and temporal expression of all its genes. These coordinated expression-patterns are maintained primarily through the process of transcriptional regulation. Transcriptional regulation is mediated by proteins binding to regulatory elements on the DNA in a combinatorial manner, where particular combinations of transcription factor binding sites establish specific regulatory codes. In this review, we survey experimental and computational approaches geared towards the identification of proximal and distal gene regulatory elements in the genomes of complex eukaryotes. Available approaches that decipher the genetic structure and function of regulatory elements by exploiting various sources of information like gene expression data, chromatin structure, DNA-binding specificities of transcription factors, cooperativity of transcription factors, etc. are highlighted. We also discuss the relevance of regulatory elements in the context of human health through examples of mutations in some of these regions having serious implications in misregulation of genes and being strongly associated with human disorders.

Keywords: *transcriptional regulation; enhancers; silencers; tissue-specific regulatory elements; population variation; non-coding diseases; computational analysis of regulatory element sequence composition*

INTRODUCTION

A human body has numerous different types of cells [1], which are further organized into tissues with distinct structures and functions. The proper development and functioning of all these tissues require precise spatial and temporal expression of the thousands of genes encoded in the human genome. A cell achieves this primarily by regulating the rate of transcription of its genes, a mechanism commonly referred to as transcriptional regulation.

Transcriptional regulation is mostly mediated by sequence-specific binding of proteins called transcription factors (TFs) to regions on the DNA called TF binding sites. Rarely does a single TF–DNA binding event control the transcription of the target gene in eukaryotes. Instead, different

combinations of ubiquitous and cell type-specific TFs act together, binding to regulatory elements, which harbour the respective TF binding sites. As a result of this combinatorial control, a human cell is able to regulate the transcription of its large number of protein coding genes (between 20 000 and 25 000 [2]) by a relatively small number of TFs (8% of all proteins [3]). Furthermore, a regulatory element can operate tens of thousands of base pairs (bp) away from the target gene [4], adding another layer of complexity to transcriptional regulation.

While most genes have been successfully annotated in the human genome, our knowledge of regulatory elements controlling these genes in different cell types, at various time-points and under different environmental stimuli is still limited. Recent studies

Corresponding author. Ivan Ovcharenko, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD, USA. Tel: (301) 435-8944; Fax: (301) 480-2290; Zip Code: 20894; E-mail: ovcharei@ncbi.nlm.nih.gov

Leelavati Narlikar is a postdoctoral fellow at the Computational Biology Branch of the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH). Her research interests include modeling the architecture of tissue-specific enhancers and developing computational techniques to identify novel regulatory elements.

Ivan Ovcharenko is a Principal Investigator at the Computational Biology Branch of the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH). His research is focused on the computational analysis of gene regulation in the human and other vertebrate genomes. Ovcharenko laboratory is particularly interested in determining the genomic code of tissue-specific regulatory elements, evolutionary divergence of enhancers and silencers, population variation in non-coding DNA and non-coding polymorphisms associated with genetic disorders.

have shown that mutations in many of the known regulatory elements are associated with diseases [5], indicating the important role regulatory elements could play in disease diagnostics and drug discovery.

Here we review the problem of identifying regulatory elements and their functions in various cellular processes. We briefly look at the different bio-molecules that participate in transcriptional regulation and examine the distinct roles of regulatory elements. We then discuss the close relationship between mutations within these elements and diseases emphasizing the importance of identifying regulatory elements. Finally, we survey the popular methods used to identify regulatory elements, with a special focus on computational approaches.

MAIN PLAYERS IN EUKARYOTIC TRANSCRIPTIONAL REGULATION

The transcript of a protein-coding gene, which originates from the transcription start site (TSS), is produced by the enzyme RNA polymerase II. However, the enzyme by itself does not directly recognize the TSS, and requires the presence of other factors called general transcription factors (GTFs). These GTFs assemble on the DNA at a region known as the core promoter, which includes the TSS as well as other binding sites recognized by different subunits of the GTFs. See Thomas and Chiang [6] for a review. After the GTFs form a complex with the core promoter, the polymerase binds to it, forming a transcription initiation complex (TIC). The main players regulating the formation and activity of the TIC can be classified into two groups based on their mode of activity: *trans*-acting factors that are not part of the DNA and *cis*-acting elements that are regions along the DNA.

Role of *trans*-regulatory factors

Gene activation

The assembly of the TIC on DNA has been shown to produce RNA transcripts from DNA templates *in vitro* [7]. However, *in vivo*, the recruitment of TIC requires additional factors, which can be classified into two groups:

- (i) DNA-binding proteins (activators). Activator proteins bind DNA, usually in a sequence-specific manner, making contact with 5–15 bp of DNA [8]. These TF-DNA interactions activate transcription by attracting the TIC to the

core promoter. Activators may operate by binding close to the core promoter, at proximal promoters and 5' untranslated regions (UTRs) or by binding distal regions on the DNA, at enhancers [4].

- (ii) Non-DNA-binding proteins (co-activators). There are several kinds of co-activator proteins. Some are recruited by protein–protein interactions and act as a bridge between the GTFs and activators, thereby stimulating TIC formation. Others play an important role in chromatin remodelling, around the gene as well as the regulatory regions, to allow the TIC and activators to access their binding sites on the DNA [9].

Gene repression

The final rate of transcription of a gene depends on the combined effect of gene activating and gene repressing mechanisms. As in the case of gene activation, cells employ many different mechanisms to repress genes [10]. The proteins involved in repression can be similarly classified into two groups:

- (i) DNA-binding proteins (repressors). Repressors bind DNA and inhibit transcription of the gene in several ways: (a) they may have the same specificity as an activator and compete with it for DNA binding sites; (b) they may bind close to the activator and interfere with its activity; (c) they may bind to silencers and inhibit TIC formation/activity, via protein–protein interactions; or (d) they may bind DNA elements between enhancers and promoters at regions called insulators, thereby blocking communication between them. In addition, nucleosomes themselves make excellent repressors, and help ensure that transcription does not start at incorrect sites [11].
- (ii) Non-DNA-binding proteins (co-repressors). Co-repressors do not bind DNA directly, but inhibit transcription via protein–protein interactions: (a) they may reorganize the chromatin structure to hinder the binding of activators or TIC to DNA; (b) they may bind to the activator to form a complex incapable of binding DNA and/or co-activators; or (c) directly bind to the TIC and inhibit its activity.

Role of *cis*-regulatory elements

In this review, we primarily focus on the DNA elements that contribute to transcriptional regulation.

There are several different kinds of such regulatory elements that are utilized by activators or repressors, or are responsible in changing the chromatin landscape to either activate or repress transcription.

Promoters

Promoters can be classified as core promoters (regions within 100 bp around the TSS [12]) and proximal promoters (regions further away from the TSS, but generally limited to a few hundred base pairs [5]). As mentioned previously, core promoters contain binding sites for ubiquitous GTFs, which are instrumental in recruiting the polymerase to the TSS. The proximal promoters contain binding sites for activator proteins that interact with the GTFs and can drive tissue-specific expression [13–15].

Enhancers

An enhancer is a regulatory region found at a greater distance from the TSS (compared to promoters), and can be either upstream or downstream of the gene or within an intron [16]. Most enhancers act as modules independent of orientation and distance from the TSS of the target gene [4]; however, some cases have been reported where this is not true [17, 18]. An enhancer usually harbours binding sites of multiple activators spatially constrained to allow for a stable DNA–protein complex. Two mechanisms have been proposed for enhancer activity. The popular looping theory states that once activators bind the enhancer, the DNA between the enhancer and the core promoter loops out, bringing the activators close to the promoter [12]. Specific protein–protein interactions between activators binding the enhancer and the promoter ensure that the correct target gene is activated. DNA scanning is the alternative proposed mechanism, where after binding enhancers, activators move continuously along the DNA until they encounter their target promoter [16].

Silencers

Silencers are regulatory regions that have a repressing effect on the target gene. Many silencers act in a distance- and orientation-independent manner, although some have been reported to act only within promoters and UTRs. Silencers can be present within enhancers or can act as independent modules with binding sites for repressors [19].

Insulators

Insulators are regulatory regions that stop the activating or repressing transcriptional activity in a locus

from spreading to an adjacent locus. They can do so in one of two ways, either by inhibiting the interaction between enhancer/silencer and promoter, or by preventing the spread of heterochromatin through the formation of a barrier [20]. An insulator usually contains multiple binding sites for TFs and the strength of the insulator is directly proportional to the number of binding sites [21].

Other elements

Eukaryotic genomes contain two additional types of regulatory regions: locus control regions (LCRs) and matrix attachment regions (MARs). LCRs are regulatory elements that enhance the expression of a cluster of genes in a specific cell type. LCRs can contain several different enhancers, silencers or insulators, each of which can be bound by different TFs [22]. MARs are elements on the DNA that make contact with the nuclear matrix. These regions are AT-rich and are believed to facilitate dynamic changes in chromatin structure to allow accessibility to TFs at their binding sites [23].

Tissue-specific control of transcription

As described in the previous subsections, transcriptional regulation is a collaborative effort between different TFs, chromatin remodeling complexes and other non-DNA-binding co-factors. These proteins can be either ubiquitous or cell type specific, but together activate or repress genes by targeting specific regulatory elements. As a result, regulatory elements harbouring multiple TF binding sites are often referred to as *cis*-regulatory modules (CRMs), with each CRM contributing to a specific spatial and temporal expression pattern of the gene [24]. CRMs typically range from 50 bp to a few 100 bp, and rarely >1 kb in length [25].

Activation or repression is seldom a binary switch of ON and OFF. Rather, the rate of transcription is modulated between the two extremes by the relative concentrations of activators and repressors in each cell type. For example, the *Yellow* gene in *Drosophila* has two upstream tissue-specific enhancers. One of the enhancers drives expression of *Yellow* at low levels in large parts of the wings, giving them a light grey colour. The other enhancer is stronger, driving expression at high levels in the abdomen, giving it a darker hue [26].

IMPORTANCE OF REGULATORY ELEMENTS IN HUMAN HEALTH: DISEASES DUE TO MISREGULATION

Genetic disorders are commonly associated with mutations in protein coding genes, including non-synonymous nucleotide substitutions, deletions, insertions and introduction of premature stop codons. Genetic abnormalities associated with gene mutations have been reported for Parkinson's disease [27], breast cancer [28], cystic fibrosis [29] and hundreds of other diseases [30].

Mutations in regulatory elements are generally assumed less likely to have a pronounced phenotypic impact, as these affect the expression pattern of a gene, not the structure or function of a protein. Furthermore, it is common for a gene to have multiple regulatory elements with each one having a small contributory function [31, 32]. This redundancy in the function of regulatory elements in a locus [31, 33, 34] has been argued to provide an explanation to why deletions of certain ultraconserved non-coding elements with enhancer activity lead to no observable phenotype [35]. However, contrary to these findings, the number of recorded cases of non-coding mutations linked to human diseases has been growing rapidly. HTRA1 promoter mutation has been linked to macular degeneration [36], PKLR promoter mutation to pyruvate kinase deficiency [37], erythropoietin promoter mutation to diabetic eye and kidney complications [38]. Multiple other promoter mutations have been also associated with different diseases [5]. An intronic mutation in the RET gene has been linked to Hirschsprung disease risk with a 20-fold greater contribution to risk than rare alleles [39]. DAX-1 intronic mutation has been shown to cause X-linked adrenal hypoplasia [40]. There are many cases of distant intergenic mutations as well. One of the classical examples is the SHH mutation that causes pre-axial polydactyly [41] and resides 1Mb away from the misregulated gene in a well-conserved region. Another polymorphism in the IRF6 enhancer located 10 kb upstream of the TSS is associated with cleft lip [42].

Genome-wide association studies (GWAS) provide a high-throughput approach to rapidly identify disease-causing polymorphisms by scanning markers across genomes of many people. Results of many GWAS are available in the database of genotypes and phenotypes (dbGaP) [43]. A GWAS study involving myocardial infarction (MI) warrants a special

mention in the context of disease-associated mutations in non-coding elements. MI is a common presentation of ischaemic heart disease; a disease accounting for >12% of deaths worldwide [44]. A recent study of early-onset MI performed by the Myocardial Infarction Genetics Consortium limited strong genetic associations of the disease to nine single nucleotide polymorphisms (SNPs) [45]. All of them reside in non-coding regions of the human genome.

Identifying mutations within coding regions that cause diseases and/or disrupt normal biological processes is generally an easier task than identifying those within non-coding regulatory regions. This is primarily due to the inherent difficulty in identifying non-coding regulatory regions. Whereas coding regions have characteristic exonic features making them relatively easy to spot, non-coding regulatory regions have few distinguishing sequence signatures. Moreover, establishing the identity of the target gene can be a further challenge since these elements do not necessarily control the gene closest to them. As a result, although several regulatory elements have so far been identified, the list is far from comprehensive. Indeed, the fraction of genes regulated by each type of regulatory element (enhancers, silencers, insulators, LCRs and MARs) has not yet been established [12]. In the next two sections, we discuss current approaches geared towards identifying and characterizing regulatory elements.

EXPERIMENTAL APPROACHES TOWARDS IDENTIFYING REGULATORY ELEMENTS

Since the discovery of the first long-range regulatory element acting upon a mammalian gene [46] almost three decades ago, technologies for detecting regulatory elements have advanced tremendously. In this section, we examine experimental strategies, both small-scale and high-throughput, for identifying regulatory elements.

Assay-based experiments

One of the most effective ways of examining the regulatory activity of a DNA region is with a reporter gene assay. In such assays, plasmids containing the region of interest and a reporter gene whose expression level can be measured accurately (e.g. green fluorescent protein) are introduced into cells of the organism of interest. The structure of the plasmid

depends largely on the kind of role the element is expected to play in regulation. If the element is being tested for promoter activity, it is placed immediately upstream of the reporter gene. If the element is suspected of being an enhancer, a weak promoter that needs an enhancer to drive expression is placed immediately upstream of the reporter gene and the element to be tested is placed either upstream or downstream of the promoter-gene construct. If the element is a silencer, the weak promoter is replaced by a strong promoter that is sufficient to drive ubiquitous expression. If the element to be tested is an insulator, it is placed between a well-characterized enhancer-promoter pair, upstream of the reporter gene. In this case, it is important to check that the placement of the element beyond the enhancer (upstream of both enhancer and promoter) does not repress transcription. For a review on designing reporter gene assays, see Carey and Smale [12].

Transfection assay is a commonly used reporter gene assay where the plasmid is introduced into cultured cells using a transfection procedure. This can be done in a transient or stable manner. In the former, the plasmid usually remains episomal and does not get integrated into the host genome. As a result, these regions may not be in an appropriate chromatin configuration and could lead to aberrant observations. This limitation can be overcome in stable transfection assays, where special measures are taken to ensure that the plasmid gets integrated into the host genome. A major advantage of transfection assays is that they can be performed in a high-throughput manner [47, 48].

Transfection assays are performed in immortalized cell lines that may not resemble environments naturally occurring in the organism. Transgenic assays overcome this limitation by employing animal models. In these assays, the plasmid is integrated into a fertilized egg at several random locations within the host genome. The *in vivo* expression pattern of the reporter gene in the embryo indicates the tissue-specific activity of the inserted element. These assays have been successful in various animal models like fly, fish, frogs, chickens and mice [49–54]. A large-scale study involving human regions tested in mouse embryos identified 75 enhancers active at a particular time-point [33]. Since the publication of the original study, the data set of these tissue-specific enhancers has grown to 497 enhancers. [55]

High-throughput experiments

Assay-based methods are usually time-consuming and expensive. In addition, they are limited by the number of elements that can be tested at a time. In this section, we review some high-throughput techniques which can locate regulatory elements on a genome-wide scale.

A chromatin immunoprecipitation (ChIP) experiment is used to determine the genomic sequences bound by a particular protein *in vivo*. The protein of interest is cross-linked to the chromatin in the cells, which are then lysed and the DNA is sheared into pieces of desired size. Using an antibody specific to the protein of interest, protein-DNA complexes are precipitated from the mixture. The identity of DNA regions that are part of the complex can be determined either by using microarrays (ChIP-chip) or by high-throughput sequencing (ChIP-seq). A major advantage of this technology is that the whole genome is tested for *in vivo* binding of the protein of interest. Also, this method can detect different kinds of regulatory elements depending on the function of the profiled protein. For instance, ChIP experiments profiling the insulator protein CTCF have identified locations of putative CTCF-binding insulators in multiple organisms and cell types [56–58]. Similar experiments with different proteins have been used to identify promoters, enhancers and silencers [59–62].

One drawback of this technology is that a specific antibody needs to be created for every TF of interest. Furthermore, finding all regulatory elements active in a particular cell type in principle requires the identity of all TFs likely to act in that cell type. Visel *et al.* [63] approached this problem in a different manner: they profiled co-activator p300 associated with enhancers [64] instead of a specific DNA-binding TF. ChIP-seq profiling of this protein in three different tissues of the developing mouse embryo identified distinct putative enhancers. Visel *et al.* further demonstrated that a large fraction of these regions were indeed tissue-specific enhancers.

Chromatin structure is another indirect indicator of regulatory elements. DNase hypersensitive sites (HSs), i.e. nucleosome-depleted regions that are easily digested by DNase I enzyme, have long been associated with regulatory elements that are bound by TFs. The novel DNase-chip [65] (or DNase-seq) technology has provided a genome-wide view of DNase HS in T cells, which are indeed enriched for binding sites of TFs active in the same cell

type [66]. Similarly, regulatory elements are known to be enriched for certain histone modifications [181, 182]. Recent genome-wide profiles of various histone modifications using ChIP experiments have revealed the location of several putative regulatory regions in different cell types [183, 184].

COMPUTATIONAL APPROACHES TOWARDS IDENTIFYING REGULATORY ELEMENTS

Small-scale experiments, while specific and generally reproducible, are labour-intensive and impractical when many elements need to be tested. Current high-throughput experiments test several regions simultaneously, but are usually noise-prone and still limited to a few cell types and environmental conditions. With 98% of the 3 Gb human genome being non-coding and therefore likely to harbour regulatory signals, computational approaches towards detecting them are proving invaluable. In this section, we discuss two related parts of the problem. As mentioned previously, TFs bind DNA in a sequence-specific manner, and hence, detecting binding specificities of individual TFs constitutes the first part of the problem. These binding specificities can then be used to determine potential binding sites of TFs in the genome, which leads to the second part of the problem: identifying functional clusters of binding sites of TFs constituting regulatory elements.

Identification of TF binding specificities

TF binding specificities are often represented as position weight matrices (PWMs) [67] with each position in the binding site modelled as a multinomial distribution over the four nucleotides. Small-scale experiments like electrophoretic mobility shift arrays [68] and DNA footprinting [69] can test the binding affinity of a TF with a few DNA templates at a time; doing so for a large number of DNA templates is highly impractical. Indeed, only a small fraction of human TFs have been well characterized using such methods and are listed in the TRANSFAC [70] and JASPAR [71] databases. Recently, Berger and Bulky [72] developed a novel large-scale technology where a large number of DNA substrates can be tested simultaneously for binding by a purified protein using protein binding microarrays. The database UniPROBE [73] contains over 200 eukaryotic TFs characterized by this methodology.

Large-scale *in vivo* experiments like ChIP-chip or ChIP-seq can locate all genomic regions bound by the profiled TF. A common overrepresented signature or 'motif' can then be identified from these regions using *de novo* motif discovery programs, yielding a PWM for the TF. Similar programs are also applied to detect motifs in promoters of co-expressed genes, the assumption being that such a set of genes is likely to be regulated (and therefore bound) by a common TF. A plethora of *de novo* motif discovery programs have been developed so far, from early ones that identified signals close to TSSs [74] in prokaryotes to the more recent ones focused on eukaryotes [75].

Motif discovery methods usually fall in one of two main categories: (i) enumerative, which examine the frequency of all DNA strings and compute over-represented strings to form a PWM [76–79] and (ii) probabilistic, which tackle the problem by creating a multiple local alignment of all sequences while simultaneously learning the PWM parameters using methods like expectation–maximization [80–83], Gibbs sampling [84–89] or greedy approaches [90]. Each category has certain advantages over the other. Enumerative approaches exhaustively search the whole space and therefore (unlike probabilistic methods) do not run the risk of getting stuck in a local optimum. In contrast, probabilistic methods can handle arbitrary variations in the motif model and are not affected by the length of the motif. A combination of the two approaches has also been proposed [91, 92].

An assessment of 13 publicly available methods [75] showed that no method consistently surpassed others in all data sets, indicating that the problem of motif discovery is far from solved. In addition, most tools performed better on yeast data sets than similarly created data sets from more complex organisms like human and mouse.

Recently developed methods have approached the problem of improving the detection of motifs in two ways [93]. The first is by improving the model for representing binding sites. Since a PWM cannot model the dependence of nucleotide preferences between positions, more flexible models like pair-correlation models [94], trees [95], mixtures of PWMs and trees [96], non-parametric models [97] and feature-based models [98] have been developed and shown to be more effective for some TFs. The other direction has been towards using additional biological information like sequence

conservation [99–107], TF concentration [108] computed based on gene-expression data, locational preferences of binding sites within co-bound sequences [109, 110], chromatin state of the genome [111], TF structural information [112–117] and DNA structural information [118].

Identification of cis-regulatory modules

The aforementioned methodologies identify PWMs recognized by TFs and the short 5–15 bp regions most likely to be bound by TFs within the set of input sequences. These methods usually treat each site independently and are employed when the search is carried out in a set of co-bound sequences not longer than a few 100 bp. However, searching for regulatory elements, even if the TF PWM is known, is trickier: a simple scan of the genome for sequences similar to learned PWMs can often lead to spurious matches which occur frequently in the genome by chance and are not necessarily utilized by the TFs *in vivo*. One way to solve this problem is by finding clusters of TF binding sites. As mentioned previously, transcriptional regulation is a collaborative effort between different TFs binding next to each other forming CRMs. Simply put, solitary binding sites are less likely to act as regulatory

elements than binding sites occurring in clusters, which is the primary basis in the rapidly growing field of CRM detection.

CRM detection was initially developed to identify core promoters. Two early programs used different approaches to solve the problem: PromoterScan [119] used known GTF motifs, the TATA box and motifs of other TFs known to be enriched near the core, while PromFind [120] used the variations in hexamer frequencies across promoter, coding and non-coding regions. Since then, several programs have employed more complex computational techniques [121–127] to solve this problem (see Bajic *et al.* [128] and Bajic *et al.* [129] for an assessment of various promoter prediction methods on human genomic data and Table 1 for details of some of these methods). Not surprisingly, the accuracy of these programs increases with an increase in high-quality hand-curated training data. Incorporation of large-scale data from recent cap analyses of gene expression (CAGE) [130] experiments, which identify the 5' ends of cDNAs, has enabled computational approaches to detect core promoters and TSSs with high resolution and remarkable accuracy [131, 132].

Many CRM-detection algorithms have also been developed to detect distal regulatory elements, from

Table 1: Computational methods applied to detecting core-promoters/TSSs

Algorithm	Methodology	Data
PromoterScan [119]	Relative densities of matches to known TF motifs in promoters and non-promoters are used to compute a 'promoter recognition profile'	Primate promoters
PromFind [120]	Relative densities of each hexamer in promoter and non-promoter regions are used to compute a linear scoring scheme	Vertebrate promoters
PromoterInspector [121]	A context-based system using IUPAC words are used to make predictions	Primate promoters
McPromoter [122]	Markov models and neural networks are used to train a model from sequence and structural features	<i>Drosophila</i> promoters
FirstEF [123]	A decision tree based on quadratic discriminant functions is constructed to exploit CpG islands and sequence signatures of promoters and first exons	Human promoters and first exons
Eponine [124]	A hybrid relevance vector machine is trained on a large number of arbitrary PWMs to learn a sparse model of PWMs relevant for discrimination	Mammalian promoter sequences
DragonGSF [125]	A neural network is trained on a large window around TSSs based on CpG islands, GC content and differential PWMs learned from sequences downstream of TSS	Human TSSs
ARTS [126]	A support vector machine is trained on many different kernels derived from sequence and DNA local structure to discriminate between sequences that contain a TSS and those that do not	Human TSSs
ProSOM [127]	Unsupervised clustering based on self-organizing maps is used to classify between structural profiles of promoter and non-promoter sequences	Human TSSs
Frith <i>et al.</i> [131]	A position-specific Markov model is built on different overrepresented k-mers	Human TSSs
Megraw <i>et al.</i> [132]	Positional biases of PWMs of various TFs and with other sequence features like GC content are exploited to train a linear model on a large subset of TSSs determined by CAGE	Mammalian TSSs

early ones which modelled the co-occurrence of two TF binding sites [133, 134] to more complex ones which use sequence conservation, gene-expression data, inter-dependence between various TF binding sites, etc. Predictions based solely on sequence conservation have been shown to achieve remarkable success [31, 33, 135], although they are likely to miss many species-specific elements or functional elements that do not produce 'high scoring' alignments with currently available tools [136]. Conversely, conservation across species does not necessarily imply regulatory functionality of the region [137, 138]. However, when interpreted appropriately, sequence conservation holds tremendous potential in reducing the vast search space of the non-coding genome and is used extensively by most CRM detection algorithms.

TF motifs have been used to produce a genome-wide map of TF binding sites [139], and predicting CRMs based on their higher densities has been shown to be beneficial [140–143]. If the identity of TFs active in the cell type of interest and their motifs is known, the predictive power of the methods increases for that cell type [144–150]. In a complementary approach, the loci of genes with a similar function can be searched for common TF binding sites [151–154]. In such approaches, TFs specific to that function can also be learned. This has also been attempted without prior knowledge of motifs, by learning overrepresented words [155, 156] in loci of co-regulated genes.

Methods have been targeted to find a special class of CRMs, those containing binding site clusters of the same TF, also known as homotypic clusters. Homotypic clusters have been widely studied in *Drosophila* [157–159], but are yet relatively unexplored in mammalian genomes. A large fraction of methods use a set of elements believed to be functional in a particular process or cell type and train a model based on the frequencies and relative distributions of motifs within them. One of the earliest CRM discoveries in mammalian co-regulated sequences was performed by Wasserman and colleagues in muscle cells [160] and later in liver cells [161]. The approach there was to compile PWMs of known muscle (liver) TFs and use them to learn a logistic regression model to classify between muscle (liver) and non-muscle (non-liver) regulatory regions. Since then many methods have been developed that train a model based on TF motifs occurring in a set of CRMs to make novel predictions;

in many cases, a set of motifs is needed to be provided by the user [149, 150, 162–164], in others overrepresented motifs or words are learned *de novo* from the data [165–169].

Table 2 shows a description of several CRM-detection methods grouped according to the type of data they require. All these methods make use of a subset of the following biological data: libraries of binding specificities of known TFs, PWMs of TFs known to act in a cooperative manner, cross-species sequence conservation, known CRMs and gene-expression data. More recently, methods have been devised to exploit other kinds of biological information. Quantitative high-resolution imaging has made available concentrations of regulatory proteins targeting segmentation genes in the nuclei of a *Drosophila* embryo at different time-points during development [170]. These concentrations of TFs and their PWMs have been used to model the likelihood of a DNA sequence driving expression of a segmentation gene and hence being a regulatory element [171, 172]. Some methods have shown significant improvements in the accuracy of detecting CRMs using chromatin information, either in the form of histone modification data [173, 174] or DNase HS data [175].

CONCLUSIONS

During the last decade that encompassed the sequencing of the human and many other vertebrate genomes, our understanding of mechanisms of gene regulation has grown remarkably. The convergence of computer algorithms and bio-technology has played a major role in deciphering the architecture of the regulatory landscape of complex genomes.

While this review provides additional details, we summarize main aspects of the developments that have had the most notable impact on the advancement of the field. ChIP-chip (and later ChIP-seq) experiments have been instrumental in describing the genome map of active TF binding sites, histone modifications and chromatin structure. With the rapid sampling of additional TFs and broader sets of cell lines, we are moving towards a comprehensive landscape of the regulatory genome. Assay-based testing in model organisms like *Drosophila*, mouse and zebrafish, has produced large data sets of tissue-specific developmental regulatory sequences. On the computational front, modelling the composition of TF binding sites and inter-TF interactions in CRMs

Table 2: Computational methods applied to detecting CRMs (not restricted to core-promoters)

Algorithm	Methodology	Data
Predictions using densities of binding sites matching known PWMs		
Crowley <i>et al.</i> [140]	A two-state Bayesian hidden Markov model (HMM) is learned from positions of predicted binding sites matching PWMs from a database	Viruses and the β -globin human locus
Wagner [141]	Clusters of TF binding sites are identified based on deviation from the null hypothesis modelled by a Poisson distribution	Yeast data
SCORE [157]	CRMs based on homotypic clusters within a specified range of window sizes are identified with significance assessed using Monte Carlo simulations	<i>Drosophila</i>
Lifanov <i>et al.</i> [159]	CRMs based on homotypic clusters within a specified range of window sizes are identified with significance assessed as deviation from the null hypothesis of a Poisson distribution	<i>Drosophila</i>
MSCAN [142]	<i>P</i> -values of multiple hits in a window, of all PWMs from a database are computed to predict CRMs	Human muscle and liver sets
Blanchette <i>et al.</i> [143]	Non-overlapping binding site predictions based on all TFs in a database in human and their presence in orthologous rat and mouse sequences are used to devise a linear scoring scheme	Human regions that can be aligned with mouse and rat
Predictions using combinatorial effects of TFs known to function in similar cell types		
Wasserman and Fickett [160]	Logistic regression is used to train a model based on match scores of each PWM in a set of similarly acting CRMs	Human muscle data and later on liver data [140]
Cister [144]	An HMM is learned where each DNA position can either be in one of 'motif', 'intra-CRM background', or 'inter-CRM background' states	Human genome and muscle data; eukaryotic promoters
Ahab [145]	A window-based model is learned using number of PWM matches, strength of each match, and the weights of PWMs; can also be used when no PWMs are given, which are in that case learned <i>de novo</i>	<i>Drosophila</i> developmental data
CIS-ANALYST [146]	A window-based model is learned based on number of matches stronger than a threshold	<i>Drosophila</i> developmental data
COMET [147]	An HMM similar to Cister is learned, but Viterbi decoding is used (instead of posterior decoding), <i>E</i> -values of predicted clusters are also computed	Human promoter and muscle data
MCAST [148]	An HMM similar to comet is learned, with differences in the modelling of the background	<i>Drosophila</i> developmental data; human promoter and muscle data
Cluster-Buster [149]	An HMM similar to Cister is learned, but the algorithm is much faster thanks to a linear-time heuristic; the program can also learn weights of motifs if training CRMs are provided	Human genome
Stubb [150]	An HMM similar to Cister is trained, but takes into account correlations between binding sites and exploits phylogenetic data; like Cluster-Buster, the program can also learn weights of motifs if training CRMs are provided	Yeast genome and <i>Drosophila</i> developmental data
ModuleFinder [176]	A window-based model that incorporates homotypic and heterotypic clusters and sequence conservation is learned	<i>Drosophila</i> developmental data; human muscle data
EEL [177]	Affinities of TFs are used to detect locally aligned clusters of binding sites across orthologous regions	Mammalian genomes
BayCis [164]	A Bayesian hierarchical HMM is used, which models complex inter-motif length distributions, correlations between binding sites, and different motif, intra-CRM and inter-CRM background distributions	<i>Drosophila</i> developmental data
Exploring loci or proximal upstream regions of co-expressed genes		
CREME [162]	From a library of PWMs, a set of co-occurring matches to PWMs in conserved upstream regions of co-expressed genes are identified using a window-based approach and used to make novel predictions	Human promoters

(continued)

Table 2: Continued

Algorithm	Methodology	Data
ModuleSearcher [151]	Similar CRMs from co-expressed genes are learned in conserved regions within 10 kb upstream sequences using a library of PWMs	Human cell-cycle data
Gibbs module sampler [166]	A CRM model is trained to simultaneously infer PWMs, distributions of TF binding sites per CRM and frequencies of neighboring pairs of TF binding sites (all learned de novo) from upstream regions of co-expressed genes; can also be trained on known CRMs	Human muscle data
CisModule [167]	A two-level hierarchical mixture-model is trained to simultaneously infer PWMs, first layer being a mixture of CRMs and background and the second a mixture of motifs in the CRM and intra-CRM background; can also be trained on known CRMs	Homotypic clusters in <i>Drosophila</i> ; human muscle data
PRF-Sampler [155]	Overrepresented motifs are learned <i>de novo</i> from conserved regions of loci of co-expressed genes, while simultaneously learning regions most likely to be CRMs	<i>Drosophila</i> blastoderm expression data
EMCMODULE [163]	Starting from a library of PWMs, a small number of PWMs that best model CRMs in upstream regions of co-expressed genes are selected using a Monte Carlo method; can also be applied on known CRMs	<i>Drosophila</i> developmental genes; human muscle data
HexDiff [168]	The frequency of all hexamers is computed for known CRMs and a set of control sequences; hexamers which have a higher frequency in CRMs are chosen and used to build a linear model which can be used to scan and score new DNA windows	<i>Drosophila</i> developmental data
CMA [153]	Upstream regions of co-regulated genes are modelled as a combination of one or more composite modules, each of which contains one TF binding site or a pair (from a library of PWMs) constrained by a spacer length distribution and orientation	Human T-cell data; yeast cell-cycle data
EI [152], DiRE [178]	A linear model based on motifs from a library of TFs is learned, which selects combinations of motifs in conserved regions of loci of co-expressed genes, while simultaneously learning regions most likely to be CRMs	Human, mouse and rat expression and conservation data
CSam [156]	Similar CRMs within loci of co-regulated genes are detected by learning a model based on overrepresented words the set CRMs using simulated annealing	<i>Drosophila</i> data
D2Z-set [156]	CRMs within loci of co-regulated genes are detected using a similar strategy as CSam, but with a different statistic to measure similarity between CRMs	<i>Drosophila</i> data
ModuleMiner [154]	Similar CRMs from co-expressed genes are learned in conserved regions within 10 kb upstream sequences using a library of PWMs; the scoring scheme focuses on increasing specificity by performing a whole genome optimization	Human microarray data

has greatly improved the precision of CRM predictors. Most importantly, it is the clever use of high-throughput data arising from various experiments (which directly or indirectly indicate functionality of DNA regions) that has enabled machine learning approaches to make accurate novel predictions. We must emphasize the leading role of the ENCODE project [179] in facilitating and supporting many of these studies; first by targeting only 1% of the human genome, and then by expanding to the entire human genome and genomes of model organisms (modENCODE) [180].

With the increasing amount of publicly available GWAS data for multiple diseases, we have begun to observe the large role that the gene regulation plays in human disorders. As our understanding of the functions of non-coding mutations in diseases grows, our ability to effectively screen patients for fitness and survival will increase. The former is closely tied with advances in computational and high-throughput technologies that accurately identify regulatory elements and predict their function. Having such tools will reduce the search space from 2.9 Gb of non-coding DNA in the 3 Gb human

genome to a manageable subset of functionally relevant regulatory elements, thereby ensuring stronger associations. Additionally, knowing the structure of regulatory elements and the TFs that utilize these elements will benefit drug therapeutics through characterization of novel candidate drug targets.

Key points

- Regulatory elements play a major role in controlling temporal and spatial expression of genes in the cellular environment. The genomic code of gene regulatory elements is encrypted by combinatorial patterns of TF binding sites.
- Identification of regulatory elements has received considerable interest in the experimental and computational community. Several small-scale assay-based technologies as well as high-throughput technologies like ChIP-chip and ChIP-seq have helped elucidate key mechanisms involving regulation by these elements.
- Increasing numbers of regulatory elements are being validated in vertebrate embryos *in vivo* and stored in specialized databases, which provide valuable training data for the development of reliable computational tools aimed at predicting tissue-specific regulatory elements.
- Many human disorders, including cancer and MI, have been linked to non-coding polymorphisms. Further success in the detection of disease-causing non-coding mutations strongly depends on the development of computational tools capable of predicting the mechanistic effect of mutations disrupting TF binding sites.

Acknowledgements

We are grateful to Leila Taher and Valer Gotea for critical comments and assistance with manuscript preparation.

FUNDING

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Alberts B, Johnson A, Lewis J, *et al.* In: Marjorie Anderson, Sherry Granum (eds) *Molecular Biology of the Cell*. USA: Garland Science Publishing, 2008:1417–76.
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.
3. Babu M, Luscombe N, Aravind L, *et al.* Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 2004;**14**:283–91.
4. Khoury G, Gruss P. Enhancer elements. *Cell* 1983;**33**:313–14.
5. Maston G, Evans S, Green M. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006;**7**:29–59.
6. Thomas M, Chiang C. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* 2006;**41**:105–78.
7. Weil P, Luse D, Segall J, *et al.* Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. *Cell* 1979;**18**:469–84.
8. Bulyk M. Computational prediction of transcription-factor binding site locations. *Genome Biol* 2003;**5**:201–11.
9. Narlikar G, Fan H, Kingston R. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 2002;**108**:475–87.
10. Latchman D. Inhibitory transcription factors. *Int J Biochem Cell Biol* 1996;**28**:965–74.
11. Kaplan C, Laprade L, Winston F. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* 2003;**301**:1096–99.
12. Carey M, Smale S. In: Judy Cuddihy, Patricia Barker, Danny deBruin, Susan Schaefer (eds) *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*. Cold Spring Harbor Laboratory Press, USA, 2001:3–20, 194–212.
13. Chen S, Ruteshouser E, Maity S, *et al.* Cell-specific *in vivo* DNA-protein interactions at the proximal promoters of the pro alpha 1(I) and the pro alpha2(I) collagen genes. *Nucleic Acids Res* 1997;**25**:3261–68.
14. Lewis B, Sims R, Lane W, *et al.* Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. *Mol Cell* 2005;**18**:471–81.
15. Smith A, Sumazin P, Xuan Z, *et al.* DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci USA* 2006;**103**:6275–80.
16. Blackwood E, Kadonaga J. Going the distance: a current view of enhancer action. *Science* 1998;**281**:60–63.
17. Das G, Henning D, Wright D, *et al.* Upstream regulatory elements are necessary and sufficient for transcription of a U6 RNA gene by RNA polymerase III. *EMBOJ* 1988;**7**:503–12.
18. Arthaningtyas E, Kok C, Mordvinov V, *et al.* The conserved lymphokine element 0 is a powerful activator and target for corticosteroid inhibition in human interleukin-5 transcription. *Growth Factors* 2005;**23**:211–21.
19. Ogbourne S, Antalis T. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *BiochemJ* 1998;**331**:1–14.
20. Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 2006;**7**:703–13.
21. Georgiev P, Murav'eva E, Golovnin A, *et al.* Insulators and interaction between long-distance regulatory elements in higher eukaryotes. *Genetika* 2000;**36**:1588–97.
22. Li Q, Peterson KR, Fang X, *et al.* Locus control regions. *Blood* 2002;**100**:3077–86.
23. Hart CM, Laemmli UK. Facilitation of chromatin dynamics by SARs. *Curr Opin Genet Dev* 1998;**8**:519–25.
24. Davidson E. *Genomic regulatory systems. Development and Evolution*. USA: Academic press, 2001.
25. Arnosti D. Analysis and function of transcriptional regulatory elements: Insights from *Drosophila*. *Annu Rev Entomol* 2003;**48**:579–602.

26. Geyer P, Corces V. Separate regulatory elements are responsible for the complex pattern of tissue-specific and developmental transcription of the yellow locus in *Drosophila melanogaster*. *Genes Dev* 1987;**1**:996–1004.
27. Valente E, Abou-Sleiman P, Caputo V, *et al*. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* 2004;**304**:1158–60.
28. Narod S, Lynch H, Conway T, *et al*. Increasing incidence of breast cancer in family with BRCA1 mutation. *Lancet* 1993;**341**:1101–02.
29. Bobadilla J, Macek M, Fine J, *et al*. Cystic fibrosis: a world-wide analysis of CFTR mutations—correlation with incidence data and application to screening. *Hum Mutat* 2002;**19**:575–606.
30. Wynbrandt J, Ludman M. *Encyclopedia of Genetic Disorders and Birth Defects*. Facts on File, 2000.
31. Nobrega M, Ovcharenko I, Afzal V, *et al*. Scanning human gene deserts for long-range enhancers. *Science* 2003;**302**:413.
32. Engström P, Ho Sui S, Drivenes O, *et al*. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res* 2007;**17**:1898–1908.
33. Pennacchio L, Ahituv N, Moses A, *et al*. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 2006;**444**:499–502.
34. Navratilova P, Fredman D, Hawkins T, *et al*. Systematic human/zebrafish comparative identification of *cis*-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol* 2009;**327**:526–40.
35. Ahituv N, Zhu Y, Visel A, *et al*. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* 2007;**5**:e234.
36. Dewan A, Liu M, Hartman S, *et al*. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 2006;**314**:989–92.
37. Manco L, Ribeiro M, Mximo V, *et al*. A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. *Br J Haematol* 2000;**110**:993–97.
38. Tong Z, Yang Z, Patel S, *et al*. Promoter polymorphism of the erythropoietin gene in severe diabetic eye and kidney complications. *Proc Natl Acad Sci USA* 2008;**105**:6998–7003.
39. Emison E, McCallion A, Kashuk C, *et al*. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* 2005;**434**:857–63.
40. Goto M, Katsumata N. X-linked adrenal hypoplasia congenita caused by a novel intronic mutation of the DAX-1 gene. *Horm Res* 2009;**71**:120–24.
41. Lettice L, Horikoshi T, Heaney S, *et al*. Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci USA* 2002;**99**:7548–53.
42. Rahimov F, Marazita M, Visel A, *et al*. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 2008;**40**:1341–47.
43. Mailman M, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;**39**:1181–86.
44. World Health Organization. *The world health report 2004 - changing history*. <http://www.who.int/whr/2004/en/> Last accessed March 30, 2009
45. Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* 2009;**41**:334–41.
46. Banerji J, Rusconi S, Schaffner W. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* 1981;**27**:299–308.
47. Noonan D, Henry K, Twaroski M. A high-throughput mammalian cell-based transient transfection assay. *Methods Mol Biol* 2004;**284**:51–65.
48. Cooper S, Trinklein N, Anton E, *et al*. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* 2006;**16**:1–10.
49. Müller F, Williams D, Kobolák J, *et al*. Activator effect of coinjecting enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol Reprod Dev* 1997;**47**:404–12.
50. Ellingsen S, Laplante M, König M, *et al*. Large-scale enhancer detection in the zebrafish genome. *Development* 2005;**132**:3799–811.
51. Khokha M, Loots G. Strategies for characterising *cis*-regulatory elements in *Xenopus*. *Brief Funct Genomic Proteomic* 2005;**4**:58–68.
52. Woolfe A, Goodson M, Goode D, *et al*. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 2005;**3**:e7.
53. Pereira P, Pinho S, Johnson K, *et al*. A 3' *cis*-regulatory region controls wingless expression in the *Drosophila* eye and leg primordia. *Dev Dyn* 2006;**235**:225–34.
54. Allende M, Manzanares M, Tena J, *et al*. Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods* 2006;**39**:212–19.
55. VISTA enhancer browser. <http://enhancer.lbl.gov/> Last accessed March 30, 2009
56. Mukhopadhyay R, Yu W, Whitehead J, *et al*. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res* 2004;**14**:1594–1602.
57. Cuddapah S, Jothi R, Schones D, *et al*. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 2009;**19**:24–32.
58. Smith S, Wickramasinghe P, Olson A, *et al*. Genome wide ChIP-chip analyses reveal important roles for CTCF in *Drosophila* genome organization. *Dev Biol* 2009;**328**:518–28.
59. Kim T, Barrera L, Zheng M, *et al*. A high-resolution map of active promoters in the human genome. *Nature* 2005;**436**:876–80.
60. Johnson D, Mortazavi A, Myers R, *et al*. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007;**316**:1497–1502.
61. Zeitlinger J, Zinzen R, Stark A, *et al*. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev* 2007;**21**:385–90.
62. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
63. Visel A, Blow M, Li Z, *et al*. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**:854–58.

64. Merika M, Williams A, Chen G, *et al.* Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol Cell* 1998;**1**: 277–87.
65. Crawford G, Davis S, Scacheri P, *et al.* DNase-chip: a high-resolution method to identify DNase I hyper-sensitive sites using tiled microarrays. *Nat Methods* 2006;**3**:503–09.
66. Boyle A, Davis S, Shulha H, *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008;**132**:311–22.
67. Staden R. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* 1984;**12**:505–19.
68. Fried M, Crothers D. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* 1981;**9**:6505–25.
69. Brenowitz M, Senear D, Shea M, *et al.* Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol* 1986;**130**:132–81.
70. Wingender E, Chen X, Fricke E, *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 2001;**29**:281–83.
71. Sandelin A, Alkema W, Engström P, *et al.* JASPAR: An open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–D94.
72. Berger M, Bulyk M. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 2008;**4**: 393–411.
73. Newburger D, Bulyk M. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 2008;**37**:D77–D82.
74. Korn L, Queen C, Wegman M. Computer analysis of nucleic acid regulatory sequences. *Proc Natl Acad Sci USA* 1977;**74**:4401–4405.
75. Tompa M, Li N, Bailey T, *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–144.
76. van Helden J, André 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**:827–842.
77. Jensen L, Knudsen S. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 2000;**16**: 326–333.
78. Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. *Proceedings of the eighth annual conference on Int Sys for Mol Biol*, AAAI Press, 2000:344–54.
79. van Helden J, André B, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 2000;**28**:1808–1818.
80. Lawrence C, Reilly A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct Funct Genet* 1990;**7**:41–51.
81. Cardon L, Stormo G. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* 1992;**223**: 159–70.
82. Bailey T, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the second annual conference on Int Sys for Mol Biol*. AAAI Press, 1994:28–36.
83. Liu D, Brutlag X, Liu J. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat Biotechnol* 2002;**20**:835–839.
84. Lawrence C, Altschul S, Boguski M, *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;**262**:208–214.
85. Roth F, Hughes J, Estep P, *et al.* Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 1998;**16**:939–945.
86. Liu X, Brutlag D, Liu J. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, Vol. 6, 2001:127–38.
87. Thijs G, Lescot M, Marchal K, *et al.* A higher-order background model improves the detection of potential promoter regulatory elements by Gibbs sampling. *Bioinformatics* 2001;**17**:1113–22.
88. Frith M, Hansen U, Spouge J, *et al.* Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res* 2004;**32**:189–200.
89. Jensen S, Liu J. BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics* 2004;**20**: 1557–64.
90. Hertz G, Hartzell G, Stormo G. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics* 1990;**6**:81–92.
91. Ohler U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res* 2006;**34**:5943–50.
92. Barash Y, Bejerano G, Friedman N. A simple hypergeometric approach for discovering putative transcription factor binding sites. *Lect Notes Comput Sci* 2001;**2149**:278–93.
93. Hännenhalli S. Eukaryotic transcription factor binding sites modeling and integrative search methods. *Bioinformatics* 2008;**24**:1325–31.
94. Zhou Q, Liu J. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 2004;**20**:909–16.
95. Agarwal P, Bafna V. Detecting non-adjacent correlations within signals in DNA. *Proceedings of the Second Annual International Conference on Research in Computational Molecular Biology*, March 22–25. New York, USA: ACM, 1998:2–8.
96. Barash Y, Elidan G, Friedman N, *et al.* Modeling dependencies in protein-DNA binding sites. *RECOMB*, 2003.
97. King O, Roth F. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res* 2004;**31**:e116.
98. Sharon E, Lubliner S, Segal E. A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* 2008;**4**:e1000175.
99. Kellis M, Patterson N, Endrizzi M, *et al.* Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003;**432**:241–54.
100. Wang T, Stormo G. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 2003;**19**:2369–80.

101. Harbison C, Gordon D, Lee T, *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;**431**: 99–104.
102. Sinha S, Blanchette M, Tompa M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 2004;**5**:170.
103. Prakash A, Blanchette M, Sinha S, *et al.* Motif discovery in heterogeneous sequence data. *Pacific Symposium on Biocomputing*, Vol. 9, 2004:348–59.
104. Moses A, Chiang D, Eisen M. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pacific Symposium on Biocomputing*, Vol. 9, 2004:324–35.
105. Liu Y, Liu X, Wei L, *et al.* Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res* 2004;**14**:451–58.
106. Siddharthan R, Siggia E, van Nimwegen E. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 2005;**1**:e67.
107. Gordán R, Narlikar L, Hartemink A. A fast, alignment-free, conservation-based method for transcription factor binding site discovery. *Proceedings of the Twelfth Annual International Conference on Research in Computational Molecular Biology, March 30-April 2, 2008*. Singapore, Lecture Notes in Computer Science 4955, Springer, 2008:98–111.
108. Segal E, Barash Y, Simon I, *et al.* From promoter sequence to expression: A probabilistic framework. *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology, April 18-21, 2002*. Washington, DC, USA: ACM, 2002:263–272.
109. Ao W, Gaudet J, Kent J, *et al.* Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 2004;**305**:1743–46.
110. Tharakaraman K, Mariño Ramírez L, Sheetlin S, *et al.* Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* 2005;**21**:i440–48.
111. Narlikar L, Gordán R, Hartemink A. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol* 2007;**3**:e215.
112. E. Xing, R. Karp. MotifPrototyper: A Bayesian profile model for motif families. *Proc Natl Acad Sci USA* 2004;**101**: 10523–28.
113. Sandelin A, Wasserman W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* 2004;**338**: 207–15.
114. Kaplan T, Friedman N, Margalit H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol* 2005;**1**:e1.
115. Mahony S, Golden A, Smith T, *et al.* Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics* 2005;**21**:i283–91.
116. Narlikar L, Gordán R, Ohler U, *et al.* Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics* 2006;**22**:e384–92.
117. Morozov A, Siggia E. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci USA* 2007;**104**:7068–73.
118. Gordán R, Hartemink A. Using DNA duplex stability information to discover transcription factor binding sites. *Pacific Symposium on Biocomputing*, Vol. 13, 2008:453–64.
119. Prestige D. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* 1995;**249**: 923–32.
120. Hutchinson G. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci* 1996;**12**:391–98.
121. 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**:599–606.
122. Ohler U, Niemann H, Liao G, *et al.* Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* 2001;**17**:S199–S206.
123. Davuluri R, Grosse I, Zhang M. Computational identification of promoters and first exons in the human genome. *Nat Genet* 2001;**29**:412–17.
124. Down T, Hubbard T. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 2002;**12**:458–61.
125. Bajic V, Seah S. Dragon Gene Start Finder identifies approximate locations of the 5' ends of gene. *Nucleic Acids Res* 2003;**31**:3560–63.
126. Sonnenburg S, Zien A, Ratsch G. ARTS: accurate recognition of transcription starts in human. *Bioinformatics* 2006;**22**: e472–80.
127. Abeel T, Saeys Y, Rouzé P, *et al.* ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* 2008;**24**:24–31.
128. Bajic V, Tan S, Suzuki Y, *et al.* Promoter prediction analysis on the whole human genome. *Nat Biotechnol* 2004;**22**: 1467–73.
129. Bajic V, Brent M, Brown R, *et al.* Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol* 2006;**7**(Suppl 1):S3.
130. Carninci P, Sandelin A, Lenhard B, *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;**38**:626–35.
131. Frith M, Valen E, Krogh A, *et al.* A code for transcription initiation in mammalian genomes. *Genome Res* 2008;**18**: 1–12.
132. Megraw M, Pereira F, Jensen S, *et al.* A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res* 2009;**19**:644–56.
133. GuhaThakurta D, Stormo G. Identifying target sites for cooperatively binding factors. *Bioinformatics* 2001;**17**:608–21.
134. Sudarsanam P, Pilpel Y, Church G. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res* 2002;**12**:1723–31.
135. Prabhakar S, Poulin F, Shoukry M, *et al.* Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 2006;**16**:855–63.
136. Fisher S, Grice E, Vinton R, *et al.* Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 2006;**312**:276–79.
137. Nobrega M, Zhu Y, Pla jzer-Frick I, *et al.* Megabase deletions of gene deserts result in viable mice. *Nature* 2004;**431**: 988–93.
138. Cooper G, Brown C. Qualifying the relationship between sequence conservation and molecular function. *Genome Res* 2008;**18**:201–05.

139. Loots G, Ovcharenko I. ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics* 2007;**23**: 122–24.
140. Crowley E, Roeder K, Bina M. A statistical model for locating regulatory regions in genomic DNA. *J Mol Biol* 1997;**268**:8–14.
141. Wagner A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 1999;**15**:776–84.
142. Johansson O, Alkema W, Wasserman W, *et al.* Identification of functional clusters of transcription factor binding motifs in genome sequences: The MSCAN algorithm. *Bioinformatics* 2003;**19**:i169–76.
143. Blanchette M, Bataille A, Chen X, *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006;**16**:656–68.
144. Frith M, Hansen U, Weng Z. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* 2001;**17**: 878–89.
145. Ra Jewsky N, Vergassola M, Gaul U, *et al.* Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* 2002;**3**:30.
146. Berman B, Nibu Y, Pfeiffer B, *et al.* Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci USA* 2002;**99**:757–62.
147. Frith M, Spouge J, Hansen U, *et al.* Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 2002;**30**:3214–24.
148. Bailey T, Noble W. Searching for statistically significant regulatory modules. *Bioinformatics* 2003;**19**(Suppl 2):ii16–25.
149. Frith M, Li M, Weng Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* 2003;**31**:3666–68.
150. Sinha S, Schroeder M, Unnerstall U, *et al.* Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics* 2004;**5**:129.
151. Aerts S, van Loo P, Thijs G, *et al.* Computational detection of *cis*-regulatory modules. *Bioinformatics* 2003;**19**(Suppl 2): 5–14.
152. Pennacchio L, Loots G, Nobrega M, *et al.* Predicting tissue-specific enhancers in the human genome. *Genome Res* 2007;**17**:201–11.
153. Kel A, Kononova T, Waleev T, *et al.* Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics* 2006;**22**: 1190–97.
154. Van Loo P, Aerts S, Thienpont B, *et al.* 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**:R66.
155. Grad Y, Roth F, Halfon M, *et al.* Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics* 2004;**20**: 2738–50.
156. Ivan A, Halfon M, Sinha S. Computational discovery of *cis*-regulatory modules in *Drosophila* without prior knowledge of motifs. *Genome Biol* 2008;**9**:R22.
157. Rebeiz M, Reeves N, Posakony J. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. *Proc Natl Acad Sci USA* 2002;**99**:9888–93.
158. Markstein M, Markstein P, Markstein V, *et al.* Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc Natl Acad Sci USA* 2002;**99**:763–68.
159. Lifanov A, Makeev V, Nazina A, *et al.* Homotypic regulatory clusters in *Drosophila*. *Genome Res* 2003;**13**: 579–88.
160. Wasserman W, Fickett J. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 1998;**278**:167–81.
161. Krivan W, Wasserman W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* 2001;**11**:1559–66.
162. Sharan R, Ovcharenko I, Ben-Hur A, *et al.* CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Bioinformatics* 2003;**19**(Suppl 1):i283–91.
163. Gupta M, Liu J. De novo *cis*-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci USA* 2005;**102**: 7079–84.
164. Lin T, Ray P, Sandve G, *et al.* BayCis: A Bayesian hierarchical HMM for *cis*-regulatory module decoding in metazoan genomes. *Proceedings of the Twelfth Annual International Conference on Research in Computational Molecular Biology, March 30–April 2, 2008*. Singapore, Lecture Notes in Computer Science 4955, Springer, 2008:66–81.
165. Nazina A, Papatsenko D. Statistical extraction of *Drosophila cis*-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics* 2003;**4**:65.
166. Thompson W, Palumbo M, Wasserman W, *et al.* Decoding human regulatory circuits. *Genome Res* 2004;**14**: 1967–74.
167. Zhou Q, Wang W. CisModule: A Bayesian module sampler by hierarchical mixture modeling. *Proc Natl Acad Sci USA* 2004;**101**:12114–119.
168. Chan B, Kibler D. Using hexamers to predict *cis*-regulatory motifs in *Drosophila*. *BMC Bioinformatics* 2005;**6**:262.
169. Pierstorff N, Bergman C, Wiehe T. Identifying *cis*-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics* 2006;**22**: 2858–64.
170. Surkova S, Kosman D, Kozlov K, *et al.* Characterization of the *Drosophila* segment determination morphome. *Dev Biol* 2008;**313**:844–62.
171. Segal E, Raveh-Sadka T, Schroeder M, *et al.* Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 2008;**451**:535–40.
172. Bauer D, Bailey T. Studying the functional conservation of *cis*-regulatory modules and their transcriptional output. *BMC Bioinformatics* 2008;**9**:220.
173. Whittington T, Perkins A, Bailey T. High-throughput chromatin information enables accurate tissue-specific prediction

- of transcription factor binding sites. *Nucleic Acids Res* 2008; **37**:14–25.
174. Wang X, Xuan Z, Zhao X, *et al.* High-resolution human core-promoter prediction with CoreBoost HM. *Genome Res* 2009; **19**:266–75.
175. Noble W, Kuehn S, Thurman R, *et al.* Predicting the *in vivo* signature of human gene regulatory sequences. *Bioinformatics* 2005; **21**(Supp 1):i338–43.
176. Philippakis A, He F, Bulyk M. ModuleFinder: a tool for computational discovery of *cis* regulatory modules. *Pacific Symposium on Biocomputing*, Vol. 10, 2005:519–30.
177. Hallikas O, Palin K, Sinjushina N, *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 2006; **124**:47–59.
178. Gotea V, Ovcharenko I. DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* 2008; **36**:W133–39.
179. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**:799–816.
180. The modENCODE project: *Model organism encyclopedia of DNA elements*. URL <http://www.modencode.org/> Last accessed March 30, 2009
181. Hatzis P, Talianidis I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell*, 2002; **10**:1467–77.
182. Kouzarides T. Chromatin modifications and their function. *Cell*, 2007; **128**:693–705.
183. Barski A, Cuddapah S, Cui K, *et al.* High-resolution profiling of histone methylations in the human genome. *Cell*, 2007; **129**:823–37.
184. Heintzman N, Stuart R, Hon G, *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 2007; **39**:311–8