Advances in the discovery of cis-regulatory elements

Pan, Youlian

Publisher's version / Version de l'éditeur:

https://doi.org/10.2174/157489306777828026

Current Bioinformatics, 1, 3, pp. 321-336, 2006

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Advances in the Discovery of cis-*Regulatory Elements* *

Pan, Y.
August 2006

* published in Current Bioinformatics. Volume 1, number 3. August 2006. pp. 321-336. Bentham Science Publishers Ltd. NRC 48464.

Canada

Advances in the Discovery of *cis*-Regulatory Elements

Youlian Pan*

Integrated Reasoning Group, Institute for Information Technology, National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada

Abstract: Discovery of transcription regulatory elements has been an enormous challenge, both to biologists and computational scientists. Over the last three decades, significant progress has been achieved by various laboratories around the world. Earlier, laborious experimental methods were used to detect one or handful of elements at a time. With recent advances in DNA sequencing technology, many completed genomes became available. High throughput biological techniques and computational methods emerged. Comparative genomic approaches and their integration with microarray gene expression data provided promising results. In this review, we discuss the development of technology to decipher the complex transcription regulation system with a focus on the discovery of *cis*-regulatory elements in eukaryotes.

Keywords: *cis*-regulatory elements, motifs, transcription factor binding sites, transcriptional regulation, gene expression, transcription factor.

1. INTRODUCTION

Completion of genome sequences provides enormous information and opportunity for understanding molecular machinery in cells. The first challenge in the scientific community is how to extract knowledge from such massive sequences of nucleotides. With the contribution of various laboratories, a significant number of genes have been predicted and annotated. However, the understanding of how and when these genes are expressed and how they are related to each other is very minimal. One of the key steps toward this understanding is to decipher the genes' transcription regulatory machinery that promote, suppress, or enhance gene expression.

Transcription is initiated by a transcription initiation complex composed of the DNA sequences and the proteins binding on them, usually described as *cis*-acting elements (also called *cis*-regulatory elements, *cis*-elements, or DNA motifs) and *trans*-acting factors (or transcription factors) respectively. Binding between the *cis*-acting DNA elements and the *trans*-acting protein factors is a prerequisite to promotion, enhancement, or suppression of a transcription process. To understand whether a gene is regulated by a transcription factor (TF), we first have to know whether the *cis*-element, to which the TF binds, exists in the regulatory region surrounding the gene. Identification of the *cis*-regulatory elements has been a great challenge to both the biology and the computation communities. This challenge is mainly due to the fact that (1) the *cis*-elements themselves are variable in their nucleotide composition and their location with regard to the transcription start site, (2) the activity of TFs depends heavily on interactions with co-factors and other TFs, (3) TFs may need to be modified (e.g. phosphorylated) in order to be active, and (4) expression of TFs is subject to regulation by other factors.

*Address correspondence to this author at the Integrated Reasoning Group, Institute for Information Technology, National Research Council Canada, Bldg M-50, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada; Tel: 1-613-993-8556; Fax: 1-613-952-0215; E-mail: youlian.pan@nrc-cnrc.gc.ca

There is a need for efficient and reliable identification of *cis*-elements. Significant progress has been achieved in the development of both computational and biological methods to detect these regulatory elements. This paper provides an overview of the current state of technologies in the field with a focus on the principles that underlie each method. First, it addresses the biological problem of transcription and its regulation with emphasis on eukaryotes. Then it provides a survey and an assimilation of existing (*i*) biological, (*ii*) computational, and (*iii*) integrations of both biological and computational methods. The strengths and weaknesses of each method are identified. Finally it is concluded with a discussion and perspective on future directions.

2. TRANSCRIPTION AND ITS INITIATION

Transcription is a biological process through which the genetic information encoded in a DNA sequence is enzymatically copied by an RNA polymerase to produce a complementary RNA. This is the first step of gene expression. In DNA sequences, the base where transcription initiates is called the *transcription initiation site*, or more commonly, the *transcription start site* (TSS). The TSS of a transcription unit is conventionally numbered +1. Bases extending in the direction of transcription (*downstream*) are assigned positive numbers and those extending in the opposite direction (*upstream*) are assigned negative numbers. A DNA sequence region (a few tens of bases in bacteria and several hundreds up to one thousand bases in eukaryotes) upstream of the TSS are usually called the promoter.

Bacteria have only a single RNA polymerase, while eukaryotes have three: RNA polymerases I, II and III. RNA polymerase is a multi-protein enzyme and is the target, directly or indirectly, of most regulation of transcription [1]. To transcribe a gene, RNA polymerase proceeds through a series of well-defined steps in three phases: *initiation*, *elongation* and *termination*. In this review, we focus only on initiation. The initiation phase itself can be divided into a series of defined steps which differ between prokaryotes and eukaryotes.

1574-8936/06 $50.00+.00

In prokaryotes, such as *E. coli*, an initiation factor called σ first binds to the promoter on two distinct DNA promoter sequence elements, at -35 and -10, and then mediates binding of the core RNA polymerase to the promoter. Once bound to the promoter, the RNA polymerase spontaneously undergoes a configuration change, becomes more intimately engaged with the promoter and opens the DNA double helix to reveal the template strand. Binding of the σ factor and RNA polymerase to DNA sequences is subject to regulation by activators and repressors bound on other *cis*-elements.

In eukaryotic cells, a given region of DNA wraps around a core of eight histone proteins and forms the basic building block of chromosomes called the *nucleosome*. Nucleosomes are assembled into a higher order structure called chromatin with different properties depending on the regulatory context. Chromatin maintains genes in an inactive state by restricting access to RNA polymerase and its accessory factors [2]. To activate a gene, the chromatin encompassing that gene and its control regions must be altered to permit transcription. High order chromatin structures must be decondensed, the specific nucleosomes over gene-specific regulatory regions must be made accessible to regulatory proteins, and nucleosomes within the gene itself must be remodeled to permit passage of RNA polymerase for transcription. Once the regulatory region of a gene becomes accessible and bound by a combination of TFs, the RNA polymerase is recruited to transcribe the gene.

Transcription of a eukaryotic gene requires assembly of a preinitiation complex (PIC), which consists of template DNA, RNA polymerase II complex and five general TFs (TFIID, TFIIB, TFIIF, TFIIE, TFIIH). Transcription initiation involves two steps. First, binding of various TFs to promoters and enhancers constitutes a multi-protein complex. This multi-protein complex then directly or indirectly recruits the PIC to the core promoter in the vicinity of the TSS. Subsequently, transcription is started by the polymerase II complex, which itself is subject to regulatory influence of TFs [3].

3. *CIS*-ELEMENTS AND THEIR ROLES IN REGULATION OF GENE EXPRESSION

In eukaryotes, thousands of genes are differentially expressed in accordance with cell types, developmental stages, physiological conditions, and in response to a wide variety of intra- and extra-cellular signals. Multiple events are involved in the initiation of transcription of a gene. Details of such events have been extensively reviewed earlier [4-7]. In this section, we summarize the *cis*-regulatory elements and their general roles in transcription initiation.

3.1. Core Promoter

A core (or basal) promoter is located between approximately -40 and +35 relative to the TSS of the metazoan genes [4]. Four major core promoter elements have been identified: a TATA box, an initiator element (Inr), a downstream promoter element (DPE), and a TFIIB recognition element (BRE) [4, 8] (Table **1**). The BRE is bound by TFIIB while the other three elements are bound by components in the TFIID complex (Table **1**). The TATA box (also called a Goldberg-Hogness box) was the first core promoter element identified in eukaryotic protein-coding genes [10]. In yeast, this element is present at 40-120 base pairs (bp) upstream of the TSS. In higher eukaryotes and viral protein-coding genes, the TATA box is present 25 to 30 bp upstream of the TSS. The TATA box is bound by the TATA-binding protein (TBP), a key element in the TFIID complex. Binding of TBP to the TATA box is a key early step in transcription initiation. In promoters lacking TATA boxes, proteins that bind to other promoter motifs facilitate TBP association with DNA in a sequence-independent manner [7]. Once TBP binds, several TBP-associated factors (TAFs) guide the RNA polymerase II complex onto the core promoter region.

Comparison of promoter sequences from transcribed protein-coding genes revealed that most of them contain an adenosine (A) at TSS (+1) and a few pyrimidines surround this nucleotide [11]. This 7-8 bp motif (Table 1) was defined as a discrete core promoter element and named as the initiator. The initiator functions similarly, with regard to transcription initiation, to the TATA box and often independently of a TATA box. The Inr is recognized by the TFIID, and also by RNA polymerase II, TFII-I and YY-1 [4]. Studies of TATA-Inr spacing showed that the two elements act synergistically when separated by 25-30 bp, but independently when separated by more than 30 bp [12]. When separated by 15-20 bp, synergy is retained, but the location of the TSS is dictated by the TATA box rather than the Inr (i.e. transcription initiation occurs 25 bp downstream of the TATA box) [12].

The DPE motif is located at 27 to 31 bp downstream of the TSS. In a subset of TATA-less promoters, the DPE motif is required for binding of TFIID [8]. A typical DPE-

Table 1. Core Promoter Elements

	BRE	TATA	Inr	DPE
Consensus motif	SSRCGCC	TATAWAAR	YY**A**NWYY	RGWYV
Location	-37 ~ -31	-30 ~ -25 -120 ~ -40 (yeast)	-2 ~ +5	+27 ~ +31
Binding Protein	TFIIB	TBP	TFIID, RNA polymerase II, TFII-I and YY-1	TFIID
Interaction with other core elements	N/A	Inr	TATA box DPE	Inr
Organism found in	All except for plants and yeast	All	All	All

BRE: TFIIB recognition element, TATA: TATA box, Inr: initiator, DPE: downstream promoter element. Note: each of the motifs is found in only a subset of core promoters. A particular core promoter may contain some, all or none of these motifs. In the initiator motifs, the underline at "A" indicates the TSS. The degenerate nucleotides are described by IUPAC-IUB recommended code [9].

dependent promoter also contains an Inr. In these promoters, mutation of either DPE or Inr results in a loss of TFIID binding and basal transcription activity [13]. A single nucleotide increase or decrease in the spacing between DPE and Inr results in a several fold decrease in TFIID binding and transcription activity [14]. The DPE and Inr function together as a single core promoter unit. In this regard, the DPE differs from the TATA box, which is able to function independently. If a TATA-dependent promoter is inactivated by mutation of the TATA-box, core promoter activity can be restored by addition of a DPE downstream of the Inr [14].

Some promoters contain a BRE motif immediately upstream of a TATA box. This is the only well-characterized element in the core promoters of protein-coding genes. The BRE is recognized by a single factor (TFIIB) rather than by a complex. The interaction between TFIIB and BRE was found to clearly enhance the assembly of a preinitiation complex and transcription initiation in archaea [15], but it was observed to repress basal transcription in humans [16]. This repression by the TFIIB-BRE interaction in humans was relieved when transcriptional activators were bound to a distal site, resulting in increased amplitude of transcriptional activation. In humans, TFIIB has a helix-turn-helix (HTH) motif that binds to the BRE. Interestingly, neither a comparable HTH motif nor evidence of sequence-specific DNA binding of TFIIB has been reported for yeast [17]. No BREs have been reported in plants [4] either.

Most core promoters do not have all four elements. In humans, it is estimated that only 32% of promoters contain a TATA box, while 85% contain an Inr [18]. In *Drosophila*, only 33-43% of core promoters contain a TATA box [4]. In a database of 205 core promoters in *Drosophila*, it is estimated that 29% contain a TATA box but no DPE, 26% contain a DPE but no TATA, 14% have both, and 31% have neither [14]. Promoters of some genes contain neither a TATA-box nor an Inr and are called null core promoters [2, 19]. Some genes may have multiple core promoters with different TSSs; both TATA and TATA-less core promoters can be associated with alternate TSSs in the same gene [20].

There are several other DNA sequence elements that contribute to core promoter activity. For example, the downstream core element in the human ß-globin gene located between +10 and +45 contributes to transcription activity and binding of TFIID [21]. A downstream promoter element (from +11 to +50) in the human glial fibrillary acidic protein gene is required for TFIID binding and transcription activity [22, 23].

3.2. Transcription Factor Binding Sites

Although necessary for transcription, the core promoter is not a common point of gene regulation. It cannot by itself generate functionally significant levels of mRNA. Most proteins that bind to the core promoter are ubiquitously expressed and provide little regulatory specificity. Various other transcription factor binding sites (TFBSs) confer specificity of transcription. The production of functionally significant levels of mRNA is controlled through sequence-specific binding of TFs to DNA sequences outside of the core promoter (see [2, 7] and references there in). Also, in many cases, the association of a TF with a TFBS and the function of a TF require the presence of co-factors [24].

Identifying true TFBSs is not straight forward (see [2, 7] and sections below). It is difficult to be certain that all functional TFBSs within a promoter have been identified and therefore it is prudent to assume that some TFBSs remain to be characterized even within a well-studied promoter. Because of this uncertainty, the range and average number of TFBSs in a typical promoter is unknown. However, an examination of well-characterized eukaryotic promoters suggests that it is not unusual to have 10-50 TFBSs on a promoter for 5-15 different factors (see [7, 25] and refs therein).

TFBSs are short and imprecise; most of them span 5-15 bp, conferring binding specificity, while a flanking region of 10-20 bp may also contribute to affinity [7]. Most binding sites can tolerate at least one, and often more, specific nucleotide substitutions without completely losing functionality. A full range of sequence variants for a particular factor with significant binding specificity is often described by a position weight matrix (Section 5.1)

Given that binding sites are short and imprecise, one can expect many potential binding sites that have the same nucleotide composition as the real sites on the basis of random distribution of genomic DNA. Many of these matches either do not bind to any proteins or do not exhibit any gene regulatory functions. Identification of a TFBS that both binds to a protein and regulates activity of a gene requires biochemical and experimental validation. Most TFs can bind to degenerate sequence motifs with alternative nucleotides in one or more positions likely with different kinetics and different protein concentrations [2]. The binding affinity of a site to certain factor also contributes to the selection of a TFBS for the factor [7].

The positions of TFBSs relative to TSSs differ enormously among genes. Often, they are located within a few kilo-bases (kb) upstream of the TSS [7], but they can be found at >30 kb upstream of the TSS [26-28], within the 5' UTR [29], within introns [30, 31], >30 kb downstream of the transcription unit [32], and, in rare instances, even in coding exons [33, 34]. Some TFBSs lie on the far side of an adjacent locus (see [7] and refs therein). The diversity of TFBS positions is possibly because of DNA looping and bending (see Section 2) that allow the interaction between proteins binding on DNA at distant sites with regard to the primary structure [7]. However, binding sites for some factors are functionally constrained. For instance, CCAAT binding sites for CBP (CREB binding protein) are generally located 50-100 upstream of the TSS, and those for Sp1 are often located near the core promoter of mammalian genes [7].

Since TFs can act upon distant basal promoters, they are potentially able to influence transcription of multiple loci. One example is a "divergent promoter" that lies between opposite stranded paralogous loci with their 5' end centrally located (see Fig. **2M** in [7]). Extensive review in this regard can be found in [7].

3.3. TFBS Modules

Distribution of TFBSs is sparse and uneven. Within promoter regions, only small proportions (10% - 20%) of the nucleotides are TFBSs. Although TFBSs often occupy a single, discrete region near the TSS, in many cases, they are

dispersed into several distinct clusters. They are usually interspersed by regions with no known functions with regard to transcription. Spacing between TFBSs varies from partial overlap to hundreds of kilo-bases [7]. One extreme example is a regulatory module of the *Shh* locus in both humans and mice that lie ~800 kb from the TSS [35].

Clusters of nearby TFBSs sometimes operate as functional coherent modules. A module is operationally defined as a cluster of TFBSs that produces a discrete aspect of the total transcription profile [7]. A single module typically contains 2-15 TFBSs for 1-8 different factors [25]. Transcription factors within a module cooperate both homotypically (involve binding sites for the same TF) and heterotypically (involve different TFs) [36].

The *cis*-elements in a promoter module can also exhibit cooperative protein binding, in which a strong binding TF/*cis*-element pair can stabilize a weak binding of a TF to an adjacent *cis*-element. A promoter module often includes a degenerate binding site for a specific TF [37]. A weak binding site embedded in the correct context can be functionally as important as a strong binding site [38]. Functional interactions between TFs not only require their co-occurrence on the same promoter (enhancer), but often with positional [39] and orientational [40] constraints as well.

Promoter modules work in collaboration. Two aspects of promoter function are suggested as analog logic circuits [7]. First, an individual module can function as a boolean (on/off) or scalar (quantitative) element whose interactions with others have predictable, additive effects on transcription. Multiple modules are sometimes required to produce a single phase expression profile. Conversely, a single module may be involved in several different phases of an expression profile. Second, promoters integrate multiple, diverse input signals and produce a single scalar output - the rate of transcription initiation. In many promoters, signal integration is done at the core promoter. In some promoters, however, a distinct module may integrate signals from other modules.

A single module may carry out one or a combination of the following functions [7]: (1) initiate transcription, (2) boost the transcription rate, (3) mediate intra- and extra-cellular signals, (4) repress transcription, (5) restrict the effect of another module to a core promoter, (6) selectively link other modules by bringing them to proximity with the core promoter (see Section 3.2), or (7) integrate the functions of other modules by influencing transcription differently depending on what proteins are bound elsewhere [41]. The most common term for a promoter module in the literature is an *enhancer*. However, the other terms, such as *booster, activator, insulator, locus control region, upstream activating sequence* and *upstream repressing sequence*, also refer to various kind of modules.

Promoter modules can be pathway or cell type specific [42]. They can mediate the transcription response to specific signal transduction pathways [37, 43], cell type specific gene expression, and events in developmental regulation [44]. A given promoter module which has a strong response in one cell type may not be functional in another [38].

3.4. CpG Islands and DNA Methylation

In vertebrates, promoter regions of genes are usually GC rich as compared to the genome average. CpG dinucleotides are more often seen in promoters than in other regions of the genome. To measure regional richness of CpG dinucleotides, Gardiner-Garden and Frommer [45] proposed a ratio of observed/expected CpG (= $\frac{Number\_Of\_CpG}{Number\_Of\_C \times Number\_Of\_G} \times N$, where N = total number of nucleotides within the window) and defined a CpG island to be a 200-bp stretch of DNA with a C+G content greater than 50% and a ratio of observed/expected CpG greater than 0.6. Even though this definition is still used until today, Duret and Galtier [46] showed the observed/expected CpG frequency underestimates the real CpG deficiency in the G+C-rich sequence and cautioned readers in the interpretation of the published dinucleotides frequency. Antequera and Bird [47] reported an estimate of 45,000 and 37,000 CpG islands in human and mouse haploid genomes, respectively. Recent computational predictions have lowered these numbers to 27,000 and 15,500 [48-50]. Only about 60% of all human genes are associated with CpG islands. All housekeeping genes (those expressed in all cell types) and about half of all tissue specific genes associate with CpG islands [47]. Many tissue-specific genes have CpG islands in their 5' promoters. Some tissue-specific genes have CpG islands at their 3' end. More recently, an analysis of chromosomes 21 and 22 indicated that regions of DNA of 500 bp with a G+C equal to or greater than 55% and a ratio of observed/expected CpG 0.65 or above are more likely to be associated with the 5' region than other regions of the genes [51].

Promoters with CpG islands usually lack TATA boxes, DPE elements, or Inr elements [52]. Additionally, they are often characterized by the presence of multiple TSSs that span a region of 100bp or more [4]. As a consequence, it has generally been difficult to identify core promoter elements within these CpG islands. Binding of basal TFs in these CpG islands is strongly dependent on recruitment by activator proteins bound to distal promoter elements [4].

The CpG dinucleotide is unique in that it usually contains 5-methylcytosine; about 80% of CpGs are methylated (a methyl group is added) at position 5 of the cytosine ring in both humans and mice. Deamination of these methylated cytosines (C) yields thymines (T) resulting in a decrease in CpG frequency and GC content. However, CpGs in CpG islands usually remain unmethylated, especially in the early developmental stage. In humans and mice, G+C contents in CpG islands are approximately 67 and 64%, while the genome averages are 41 and 42%, respectively [53].

DNA methylation appears to silence active promoters, especially during embryonic development [54], but in many cases affects genes that are already silent, thus contributing to the stability of gene silence [55]. A detailed mechanism of methylation is studied in [56-58]. A growing number of human diseases have been found to be associated with aberrant DNA methylation and demethylation [59]. In cancer cells, *de novo* methylation in the promoter of anticancer

genes has been shown to repress gene expression. These aberrant methylations are considered cancer inducing mechanisms [60]. They are a good source of tumor markers [61, 62] and targets for chemotherapeutics [58].

4. BIOLOGICAL DISCOVERY OF *CIS*-REGULATORY ELEMENTS

Biological methods for discovery of *cis*-regulatory elements are generally designed to probe the interaction between protein and DNA. Earlier *in vitro* probes in this area include, but are not limited to, DNase I "footprinting" [63], chemical modification [64], bromouracil cross-linking [65] and drug cleavage [66, 67]. In these techniques, a protein is bound to a uniquely radio-labeled DNA fragment. The protein-DNA complex is then subjected to an enzymatic, chemical or photochemical treatment which either breaks the DNA backbone directly or modifies the DNA so that its backbone can subsequently be broken by alkali [68]. After removal of protein, the labeled DNA is denatured and electrophoresed in a polyacrylamide sequencing gel. An autoradiograph of this gel shows a pattern of the end-labeled DNA fragments, which correspond to breaks in the DNA backbone.

The enzymatic and chemical treatments of the above techniques for inducing DNA strand cleavage is very hard to perform without perturbing native protein-DNA interactions in living cells. In this regard, Becker and Wang [69] developed a photochemical method to probe the protein-DNA interaction, called "photofootprinting". Light is used to probe protein-DNA interaction *in vivo* and *in vitro*. This technique is based on the observation of a UV photoproduct formed through distortion of the double helix, which is caused by the binding of a protein [69, 70]. This technique was subsequently improved by introducing a better sequencing technique [71, 72], and by utilizing the thermostable DNA polymerase in a primer extension assay [73].

Discovery of formaldehyde-mediated DNA-protein cross-linking has enabled the development of a chromatin immunoprecipitation (ChIP) technique. Formaldehyde produces DNA-protein cross-links both *in vitro* and *in vivo* (see [74] and refs therein) and at the same time displays virtually no reactivity toward free double-stranded DNA [75, 76]. Cells containing the DNA-protein cross-links are ruptured through sonication and the sheared chromatin is isolated. Antibodies against a protein of interest are used to selectively immunoprecipitate chromatin fragments. The cross-links are reversed and the specifically enriched DNA fragments are purified and analyzed through slot blot hybridization, quantitative PCR or southern blot [77, 78].

The above methods are not comprehensive because they examine only a handful of the promoters at a time. To address this, a genome-wide location method was developed to monitor protein-DNA interactions across the entire yeast genome [79]. This method combines a modified version [80] of the ChIP procedure described above with DNA microarray (chip) technology to probe immunoprecipitated DNA fragments, and subsequently named ChIP-chip [81] (Fig. **1**). Briefly, after reversal of the cross-links, the immunoprecipitation-enriched DNA is amplified and labeled with a fluorescent dye (Cy5) by using a ligation-mediated-polymerase chain reaction (LM-PCR). A sample of DNA that is not enriched by immunoprecipitation is also subjected to LM-PCR, but in the presence of a different fluorophore (Cy3). Both the enriched and unenriched pools of labeled DNA are hybridized to a single DNA microarray containing all intergenic sequences [79]. The major limitation in this technique is the definition of promoter regions for microarray design. TFBSs do not only appear in the intergenic regions, but also in other regions specified in Section 3.2. The initial microarray used CpG islands which tend to be enriched in promoter regions [82]. The second generation of microarray for ChIP-chip assays used specific oligonucleotides or DNA fragments derived from known promoter sequences and was effective in characterizing yeast transcriptional units [83]. Recently, the ChIP-chip technique was successfully applied to mammalian systems [84].

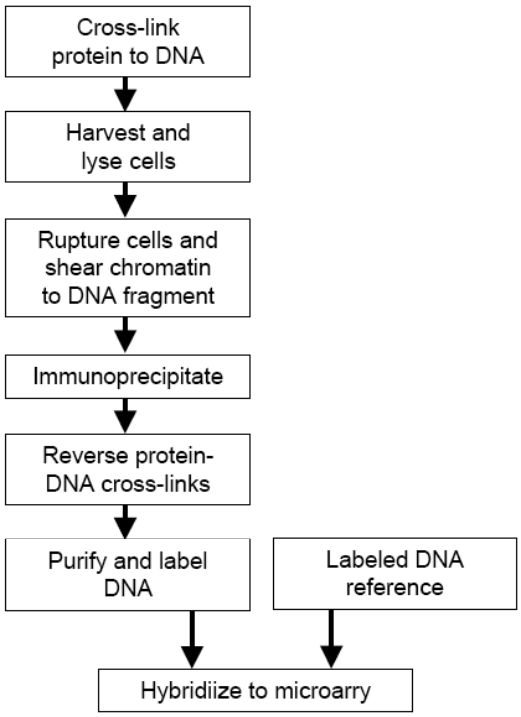

Fig. (1). Schematic procedure of the ChIP-chip technique.

Another drawback of ChIP-chip technology is that it requires a specific antibody against the TF of interest. Such an antibody may not be available. Also the condition and time for gene transcription and subsequent translation of the TF are not necessarily known. Therefore, such a TF may not be present under a given condition. The recently developed protein binding microarray (PBM) method avoids these two requirements [85, 86]. A DNA-binding protein (TF) of interest is expressed with an epitope tag, purified and then bound directly to a double stranded DNA microarray.

A combination of ChIP technology with SAGE (Serial Analysis of Gene Expression, [87]) and its modification has been successful [88, 89]. Briefly, after ChIP procedures and isolation of DNA fragments, the fragments are divided into two pools, tagged with different PCR adaptors and amplified through a LM-PCR. For these tagged genomic fragments with known chromosome locations, a procedure similar to SAGE (see [87] for details) is followed.

Another versatile method in this area is SELEX (systematic evolution of ligands by exponential enrichment) [90], also known as SAAB (selected and amplified binding sites) [91] or CASTing (cyclic amplification and selection of targets) [92]. Briefly, a purified protein is used to isolate high affinity binding sites through several rounds of *in vitro* selection and amplification. The strength of this method is its ability to isolate a small set of binding sites from a very large pool of random sequences. Selected DNA fragments are amplified through a polymerase chain reaction. One important aspect of this method is the separation of DNA-protein complexes from free DNA. Traditional methods of separation are gel mobility shift where DNA is radio-labeled [90, 93-96] and immunoprecipitation [92, 97, 98]. Recently, an affinity chromatography method was introduced for separation [99], and the entire SELEX procedure was automated [100].

Other biological methods for identification of protein-DNA interaction include nitrocellulose-binding assays [101], ELISA [102], southwestern blotting [103], reporter constructs [104] and luciferase reporter assays [105, 106]. These methods can only determine one interaction at a time; interested readers are referred to the original papers.

5. COMPUTATIONAL DISCOVERY OF *CIS*-REGULATORY ELEMENTS

There are many computational approaches to the discovery of *cis*-regulatory elements. Also, many reviews and evaluations of available software have been published over the past decade. Readers are referred to recent reviews [107-111] where further references can be found. Computational approaches can be divided into two classes based on the amount of knowledge about the *cis*-regulatory elements on which the TFs of interest are supposed to bind [112]. The first is, given a collection of known binding sites, develop a representation of these sites that can be used to find new sequence motifs and reliably predict where the additional binding sites occur. This class of methods is also applied to discover putative target genes of a TF (e.g. [113]). The second is, given a set of promoter sequences believed to contain binding sites for a TF but without knowing the sites' locations or the motif, discover the location of the sites on each sequence and a representation of the sites for the specificity of the TF. This process is also referred as *de novo* motif discovery.

5.1. Discovery of Known or Partially Known *cis*-Regulatory Elements

Biological observation of the DNA-protein interactions has enabled identification of sequence motifs to which TFs bind (Sections 3, 4). Many TFs are able to bind to motifs with alternative nucleotides at one or more positions in a motif. Through mutagenesis, such alternative nucleotides in a specific motif position can be identified. As more TFBSs become available, one can get more information through the alignment of all known sites that bind a specific TF. A *consensus* is defined to specify binding of the TF to DNA. Initially, the term *consensus* referred to a sequence that can describe most example sites for a given TF. Nowadays, the usage of *consensus* has been loosened to include degenerate site motifs that match all example sites closely, but not necessarily exactly.

A consensus can contain one or more degenerate positions to describe the specificity of a TF, but does not contain precise information about the relative likelihood of identifying alternative nucleotides at different positions of a motif. In most applications, a *positional weight matrix* (PWM), also named *frequency matrix*, *position-specific score matrix*, *position specific weight matrix, positional probability matrix* in the literature, is often more superior [112]. A PWM is usually obtained through aligning the subsequences (instances of TFBS) and describing the alignment with the frequency of each nucleotide in each column of the alignment. The result is a $4 \times m$ matrix, where m is the length of the subsequence (Fig. **2**). The PWM measures the likelihood that each nucleotide appears in each column. For example, A 4×5 PWM is derived through alignment of the six binding sites for yeast TF GCR1 (Fig. **2**). The likelihood is 1.00 for *C* to appear in position one and 0.00 for *A, G,* and *T* at the same position. Many PWMs are publicly available in databases such as TRANSFAC [115] and JASPAR [116].

>YAL038W	CTTCC
>YCR012W	CTTCC
>YCR012W	CTTCC
>YDR050C	CATCC
>YDR050C	CTTCC
>YHR174W	CATCC

Position	1	2	3	4	5
A	0.00	0.33	0.00	0.00	0.00
C	1.00	0.00	0.00	1.00	1.00
G	0.00	0.00	0.00	0.00	0.00
T	0.00	0.67	1.00	0.00	0.00
Consensus	C	W	T	C	C

Fig. (2). A set of binding site for yeast TF GCR1. Data from SCPD [114] represented with a PWM and a consensus.

The major issue with PWMs is how to pick the elements of a matrix to represent the sites. Information theory was proposed to describe the sites [117]. The information content at a position in a site was measured by a Kullback-Leibler distance (or relative entropy), which is defined as

$$I_i = \sum_b p_{ib} \log_2 \frac{p_{ib}}{f_b} \qquad (1)$$

where p_{ib} is the frequency of nucleotide b found at position i of the site, and f_b is the frequency of nucleotide b on sequences other than the sites (i.e. background) [118]. The traditional Shannon distance measure was also used as a special simplified version of equation (1) with the assumption that the frequencies of all 4 nucleotides in the background are identical:

$$I_i = H + \sum_b p_{ib} \log_2 p_{ib}$$

where $H = \log_2$(length of alphabet) = 2 [119]. The results of these distance measures can be depicted by using sequence

logos [120]. The information content of a site, with a simple assumption that the frequency at each position is independent, is the summation of information of each position in the site. For comparison among sites of different lengths, the information content of the site is usually normalized by the length (L) of the site, thus

$$I_{site} = \frac{1}{L}\sum_i I_i \qquad (2)$$

The information content tells how much the site differs from the background. Therefore, I_{site} is maximal if the site is well conserved and differs considerably from the background distribution. Comparison of information content between two matrices is usually based on the assumption that they are derived from same number of subsequences in their repetitive alignment. Caution should be taken when comparing two matrices that are derived from a different number of subsequences in the alignments. A larger number of subsequences could introduce more degenerate motifs on the one hand [121]. This will result in lower information content. Too few subsequences, on the other hand, would not represent the site well even it could have higher information content.

Recent research has highlighted that position-dependent models are more powerful than position-independent models like PWM described above and can significantly improve the accuracy of TFBS prediction. A hidden Markov model (HMM) architecture is well suited for representing profiles of multiple sequence alignments [122]. A high order HMM can model high-order positional dependency within a profile and the relationship among various profiles. In a HMM, a series of observations are described by a "hidden" stochastic process. For each column of the multiple alignments represented by a PWM, a "match" state is used to represent distribution of nucleotides in the column. An "insert" state and a "delete" state at each column allow for insertion of one or more nucleotides between the column and the next, or for deleting the consensus nucleotide. For details, please refer to the review articles [122-124] which provide descriptions of the architecture and parameters involved. The HMM algorithm has been extensively used for sequence research including searches for TFBSs.

Given a motif, either in the form of a consensus or a matrix, one is able to find new putative instances of the motif in the input sequences. For searching new putative instances through PWM represented either by the information theory or by HMM, each putative instance is scored based on its similarity to the PWM [117, 122]. One critical step in this process is the assessment of the motif quality and determination of a threshold in order to minimize the rate of false positive prediction. A standard classification test (see e.g., [125]) is usually performed to optimize both the threshold and the motif length by minimizing the classification error. More constraints, such as context information, genome-wide overrepresentation, and regional bias [126] are often used to reduce the rate of false positive prediction. Other constraints, such as orthologous gene search through phylogenetic footprinting, and modular *cis*-regulatory elements, are considered in the sections below (Sections 5.3 and 5.4). However, the introduction of constraints usually compromises sensitivity, losing the opportunity to predict a small proportion of the true motifs.

While searching for putative motifs, putative target genes of certain TFs are found (e.g. [113]). In this regard, discovery of certain pathway or disease related genes can be achieved through prediction of motifs in the promoter of the genes in question or through a genome-wide search of certain motifs.

5.2. Discovery of *De Novo cis*-Regulatory Elements

Discovery of a novel motif is usually based on statistical significance of a local alignment of input sequences and described with either a consensus or a PWM. The idea behind these methods is that each sequence in the input dataset contains one or more examples of the motif to be found, but the start offsets of the examples in each sequence are unknown. If this were known, subsequences of length m from each sequence starting at the known offsets could be aligned and a PWM or a consensus could be derived to represent this set of subsequences. The elements in a PWM constitute a motif model. The average frequency of each of the four nucleotides from the remainder regions of the input sequences constitutes the background model.

Information content has been applied to measure the statistical significance of a local alignment. Thus, optimization of the information content of a matrix for a fixed number of occurrences has been widely used in matrix-based motif discovery [127, 128]. There are several different approaches to estimate the motif parameters. Lawrence and Reilly [127] and Bailey and Elkan [129] used an expectation maximization (EM) algorithm [130], while Lawrence *et al.* [128] and Liu *et al.* [131] used a Gibbs sampling algorithm [132, 133].

The EM algorithms are named for their two iterative steps, expectation and maximization, which are repeated until a convergence criterion is satisfied. The expectation step evaluates the likelihood of each subsequence of length m to fit the PWM profile with respect to the background of the sequence. A likelihood value Z_{ij} is computed for each position i of input sequence j. The maximization step puts the best subsequence (with maximum likelihood value) from each input sequence j and builds a new alignment PWM profile. Then the expectation step is repeated with the new PWM profile. The result of the EM algorithm is influenced by the initial offset selection from each input sequence. Promoters usually contain multiple TFBSs. The MM algorithm was later developed to relax the assumption that each input sequence contains only one occurrence of the motif and was implemented in MEME software [129]. The MM algorithm estimates the start position systematically based on all subsequences through a mixture model, and applies the discovery-and-erase technique. When one session of EM is completed and a motif is found, all occurrences of this motif are "erased" from the input sequences. New motifs are found in subsequent sessions.

The EM-based algorithms (including the MM extension) are guaranteed to find a maximum, but they are sensitive to the initial start position and may be trapped by local maxima. Stochastic algorithms, such as Gibbs sampling [128], have been developed to overcome this problem and appear to be

successful. The main motivation of Gibbs sampling algorithms is to avoid premature convergence to local maxima. A subsequence of length m is selected randomly from each of the k input sequences. One of the k sequences is selected, either randomly or in specific order, and named S (Fig. **3**). A $4 \times m$ PWM profile M is built with the subsequences from the remaining k-1 input sequences. Similar to the EM algorithm, a model is built on M and the background probability of each nucleotide is calculated. The likelihood value of each possible m-mer (Z_x) from the sequence S is evaluated to represent how well each m-mer fits to the model. An m-mer is sampled stochastically according to the relative likelihood values ($Z_x / \sum_x Z_x$, [128]).

This relative likelihood value indicates that an m-mer that fits well is more likely sampled. This procedure is repeated for each sequence until all input sequences have been sampled. The whole procedure is repeated until a convergence criterion is satisfied.

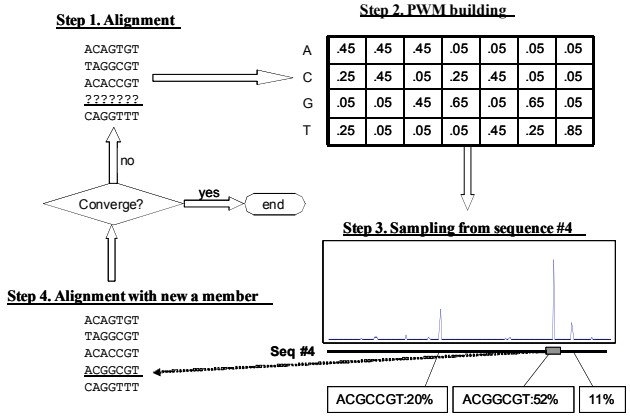

Fig. (3). Schematic description of Gibbs sampling algorithm. Note that a pseudocount (baseline probability) of 0.05 is applied to each of the four nucleotides to avoid probability of 0.00 when a nucleotide does not appear in a column of the alignment.

Theoretically, the relative information (Equation 2) is maximized only after an infinite number of cycles. Practically, this algorithm often converges very quickly [134]. One limitation of the algorithm is that it could be locked to a local maximum. Inserting another step into the algorithm can solve this problem. After a specific number of cycles, one can automatically shift (called a phase shift) all aligned subsequences to the left or right by a certain number of nucleotides.

The basic Gibbs sampling algorithm was designed for one ungapped motif per input sequence. However, the total number of TFBSs and the number of sites corresponding to each motif in an input sequence are unknown and vary among the input sequences. Motifs bound by homo- or hetero-dimers usually allow a number of non-conserved nucleotides in-between two conserved boxes. To consider these possibilities, the basic Gibbs sampling algorithm was first generalized to allow more than one type of motif per sequence and the widths of sites was inferred by using a fragmentation algorithm [135]. Subsequently, a fixed genome-wide nucleotide frequency was used as a background model; simultaneous searches of sites on both DNA strands and iterative masking techniques in searching multiple motifs were used to improve performance [136]. A higher-order Markov background model, modeling of gapped motifs and motifs with palindromic patterns [121, 137, 138], and a significance measure based on a motif score distribution estimated by a Monte Carlo method [137] were introduced to improve accuracy. Recursive sums over all possible alignments of $0 \le k \le K_{max}$ sites in a sequence was used to obtain Bayesian inference on the number of sites for each motif and the total number of sites in each sequence [139]. Recently, the Gibbs sampling algorithm was further modified to search for symmetrically structured and non-structured motifs in a set of unaligned DNA sequences [140]. As a consequence, the Gibbs sampling algorithm became one of the most popular methods in motif discovery.

Nevertheless, The Gibbs sampling algorithms do not guarantee reaching an optimal solution. This problem is solved through applying Bayesian optimization [141] and the recently implemented BioOptimizer [142], which takes as input both the sequence data and results from motif finding programs, such as Bioprospector [137], Consensus [143] or AlignACE [136], that implement algorithms described earlier. The output of BioOptimizer is a new set of predicted motif sites. Locations of each motif in the input sequences are indicated by a matrix **A** where each indicator $A_{ij} = 1$ if the motif site starts in position j of sequence i and 0 otherwise. Initially, the value of A_{ij} is unknown and a random indicator variable is considered based on an *a priori* probability. BioOptimizer then scans through each element (A_{ij}) of the matrix **A** and changes the indicator variable at each position to its opposite value only if the resultant scoring function is improved. A minor change in motif length (addition or deletion of nucleotides) is also performed only if the scoring function is improved. These changes are repeated until no further change to **A** is accepted. It is obvious that BioOptimizer is very much dependent on the initial matrix and, as with EM, can be easily trapped to a local maximum.

Both expectation maximization and Gibbs sampling algorithms are local search algorithms. A series of enumerative methods have emerged recently [144-147]. These methods search exhaustively for all possible combinations of nucleotides and select statistically the top few overrepresented motifs from a set of promoters. Subsequently, motif discovery is treated as a feature selection problem; a motif is treated as a feature of input promoter regions that discriminate the promoters from background sequences [148]. Instead of counting the frequency of overall occurrences of the motifs in a group of promoters, information about motif distribution in individual promoters is used to evaluate motif overrepresentation. Methods of this kind have been considered too slow due to the fact that their time complexity grows exponentially with the length of the motif. Given a motif length m, these methods need to evaluate 4^m candidate patterns before an optimal solution is found. However, by indexing the input sequences with a suffix tree, the execution time becomes exponential with respect to the number of allowed substitutions, which is usually small, instead of motif length [147]. These methods are guaranteed to find the globally most overrepresented motifs [145]. Some of these methods compare favorably with others with regard to accuracy [111].

In a recently developed matrix-based motif discovery method [149], a discriminating matrix enumerator (DME) is used instead of iterative sampling of subsequences. The DME exhaustively enumerates a discrete space of matrices and scores each matrix according to its relative overrepresentation (information content). The highest scoring matrices are then refined by using a local search procedure that optimizes the relative overrepresentation score (see [149] for details). When searching for multiple motifs, discovered motifs are "erased" from the sequences and the procedure is repeated.

Occurrences of a motif in a set of sequences are analogous to multiple occurrences of a *word* in a text. Therefore, a dictionary for such *words* can be established. Bussemaker *et al*. [150] first applied this analogy in their motif-finding method, MobyDick. Starting with nucleotide frequencies, one finds overrepresented dinucleotides and adds them to the dictionary, determines their probabilities, and continues to find larger overrepresented fragments of DNA. A composite fragment can be built by concatenating two or more short fragments [151]. The statistical significance of longer fragments is based on the probability of shorter fragments. Later Gupta and Liu [152] extended this dictionary model by allowing "stochastic words" and introduced a data augmentation procedure to find such words. This extension allows degenerate motifs represented by a probabilistic word matrix (e.g. Fig. **2**). A Gibbs sampling method is used to update the matrix [137, 152].

5.3. Discovering Modules of *cis*-Regulatory Elements

Expression of most eukaryotic genes is controlled by combinations of TFs binding to their corresponding DNA motifs known as "*cis*-regulatory modules" (CRM, [153, 154], Section 3.3). They are alternatively named *regulatory modules* [155-157], *promoter modules* [158], and *cis-element clusters* [159] in the literature. The modular organization of promoter functions was not extensively included in the computational modeling of transcription regulation until the past decade even though it has long been recognized from experimental work. Claverie and Sauvaget [160] first developed a method to detect a module of two distinct elements at a predefined distance and orientation in the promoters of heat-shock genes. Since there were neither compiled matrices nor reliable computational motif discovery tools available at that time, they directly encoded the two consensus sequences into their search patterns. Recently developed methods take advantage of existing databases of positional weight matrices, such as TRANSCompel [161], TRANSFAC [115] and JASPAR [116], and motif discovery tools, such as Gibbs sampling tools, which generate PWMs (Section 5.2). Most available approaches to CRMs discovery involve two primary computational steps. First, identify the existence of individual TFBSs from the input sequences using the methods described earlier (Sections 5.1 and 5.2) either through predicting novel motifs or using experimentally determined motif matrices from a database. Second, search for possible clusters of a predefined number of motifs in each module within a certain distance and orientation in a predefined sequence region [162, 163].

The individual motifs discovery processes for CRM are basically the same as those described in Sections 5.1 and 5.2. The main focus in CRM discovery is the combination step. Earlier methods generally combine individual motifs based on their physical proximity within a defined sequence window (e.g. 75-100 bp) and a distance correlation function [158, 164].

With *a priori* knowledge of motifs, multivariate logistic regression analysis [155, 156] and hidden Markov models [157, 159, 165] appear to deliver satisfactory results. Logistic regression analysis methods use PWMs to measure motif strength, but it is necessary to introduce an *ad-hoc* sequence window size. Motifs are considered together if they lie within a sequence window of a certain length, but the distance between motifs is not considered. Hidden Markov model methods consider both the strength of a motif and the distance between the motifs, but avoid the *ad hoc* window size and reduce the number of parameters to estimate in order to avoid the danger of over fitting [159]. Discrimination is greatly enhanced after the introduction of EM to concentrate weight on the relevant factors [151, 157, 159, 165] and phylogenetic conservation of the module sequence fragments in the scoring system [165, 166].

Without *a priori* knowledge of binding motifs, Zhou and Wong [167] applied a hierarchical mixture model and developed a Bayesian approach for simultaneous inference of CRMs and binding sites for a set of transcription factors by means of a Gibbs sampling approach. However, this hierarchical mixture approach inherits the problem of Gibbs sampling not guaranteeing a global optimal solution. Most recently, Gupta and Liu [163] first used the available database of transcription factors and a *de novo* motif finding algorithm to find a set of related candidate motifs, then applied the evolutionary Monte Carlo method [168] to iteratively select motif types that are likely members of a *cis*-regulatory module and a dynamic programming-based recursion to update the corresponding sites and parameters. Their EMCMODULE algorithm appears to out-perform other methods, such as *cis*Module [167] and Gibbs Module Sampler [154] and is probably the best of the currently available *de novo* CRM discovery tools. Although the algorithm does not guarantee to find a global optimal solution, results of multiple runs with different seeds indicate no noticeable difference over a wide range of prior settings [163].

The methods described in this section could potentially be applied to genome-wide search of CRM. Non-coding sequence regions of defined window size that conserved either within one species or across two or more species (see Section 5.4) could be first identified. The *cis*-elements found within each region are considered to perform the same function in regulating gene transcription.

5.4. Discovery of *cis*-Regulatory Elements and Modules Through Phylogenetic Footprinting

The computational methods for discovery of *cis*-regulatory elements reviewed in the previous sections essentially focus on the overrepresentation of these elements within promoters of a single genome. As many genomes are now available [169], comparative genomics can provide a powerful approach to the systematic discovery of functional DNA sequence elements in non-coding regions. As with the coding regions, nucleotides in the functional sequence

elements appear to have lower mutation rates than those in the non-functional regions and thus are more conserved across species [170]. Phylogenetic footprinting [171] refers to the identification of functional sequence signatures through comparison of orthologous genomic sequence regions across two or more species [109, 170]. This process has been incorporated into some recently developed algorithms in the discovery of *cis*-regulatory elements and resulted in a significant improvement in the specificity of prediction [155, 156, 172].

Two main approaches have been taken in the discovery of *cis*-elements through comparative genomics [170]. The first is to find motifs that are common from multiple orthologous sequences. Footprinter [173] and OrthoMEME [174] are two examples in this category. Footprinter takes as input a list of orthologous sequences and phylogenetic distances between these sequences or their corresponding species. To increase the number of input sequences, the system takes sequences from more than three species as well as paralogues. OrthoMEME takes promoter regions of orthologous genes from two species and looks for common motifs using the MEME approach.

Most programs in phylogenetic footprinting rely on either local or global sequence alignment. Local alignment tools, e.g. BLASTZ [175, 176], look for similarity in fragments between the compared sequences, while global alignment tools, e.g. LAGAN [177], look for similarity over the entire length of the compared sequences through progressive local alignments. Global alignment tools have a higher sensitivity, whereas local alignment tools have greater specificity. Therefore local alignment tools are often chosen for *cis*-regulatory element discovery from orthologous sequences [172]. There is a database for annotated orthologous genes at NCBI [178], but caution should be taken in directly using promoters of the orthologous genes from the database. Two orthologous coding regions might be highly conserved, but conservation of regulatory regions varies widely with particular genes [134] because the two species may use the orthologues differently [170]. An alignment step in *cis*-element discovery from orthologous genes is unavoidable.

The second approach is to globally align the promoter sequences of the orthologous genes, followed by identification of conserved windows. For example, EMnEM [179], which combines the mixture models of MEME with a probabilistic evolutionary model, takes a set of aligned sequences from different species and obtains a maximum likelihood estimate of both the motif matrix and the phylogeny. CompareProspector [180] takes window percentage identity values (WPID) from LAGAN alignments of orthologous sequences into the consideration and applies a Gibbs sampling procedure. Initial samples are taken only from highly conserved sequence regions (with high WPID values). WPID values are considered in the subsequent sampling process to weight the sampled site scores. PhyME [181] also takes into consideration the conserved windows from LAGAN alignments, and uses an EM algorithm to search for a motif that best explains the data. Similarly, PhyloGibbs considers the conserved windows from alignments, but uses Gibbs sampling approach to search for multiple motifs in parallel [182].

The probabilistic models, phylogenetic hidden Markov models (or Phylo-HMMs), consider not only the conservation at each site of a genome, but also the phylogenetic distance (see [183] and references therein). They treat the molecular evolution as a combination of two Markov processes, one operates in the *space* dimension (genomic location) and the other operates in the *time* dimension (branches of a phylogenetic tree). Earlier, Phylo-HMMs were used to improve phylogenetic models that allow variation among sites in the rate of substitution, to predict secondary structure and to detect recombination events. Most recently, Phylo-HMMs have been implemented to compute a score called phastCons [184] for highly conserved sequence elements including *cis*-regulatory elements.

Others used different algorithms to score highly conserved sequence elements. Margulies *et al.* [185] identified multi-species conserved sequences (MCSs) as blocks of highly conserved aligned sequences. Their methods weight scores by phylogenetic distance and adjust the estimate of significance by a neutral substitution rate. Elnitski *et al.* [186] introduced a regulatory potential (RP) score to distinguish regulatory regions from neutrally evolving DNA repeats. This method has been tailored to three-way alignments among human, mouse and rat sequences [187]. King *et al.* [188] compared the three scoring methods (phastCons, MCSs, and RP) and found they can correctly identify 50%-60% of non-coding sequences in the *HBB* gene complex as regulatory or non-regulatory. They also found that RP performed better than the other two methods.

Phylogenetic foot printing has been extensively applied to discovery of *cis*-regulatory modules. Non-coding sequences conserved between two or more related species are located genome-wide and treated as likely candidates for regulatory regions, usually ~500-1000 bp in length [189]. Then the individual *cis*-elements are discovered using the methods described in Sections 5.1, 5.2 and 5.3.

Many other tools (e.g. [189-193]) consider the phylogenetic conservation in *cis*-element discovery and have recently been reviewed [170, 181, 189, 194]. Incorporation of comparative genomics approaches into *cis*-element discovery greatly enhances the overall accuracy of prediction [181]. A database (CORG) [195] is established for phylogenetically conserved non-coding sequence blocks from the upstream regions of orthologous genes.

6. INCORPORATION OF MICROARRAY DATA

Microarray gene expression data can provide a genome-wide view of transcription regulation [196]. One can hypothesize that co-expressed genes may be co-regulated by a common TF or CRM. Some of the recent work in *cis*-element discovery takes advantage of both sequence and microarray data based on the premise that regulatory sequence elements should explain changes in gene expression patterns [197-199]. Typically, genes are clustered into disjoint groups based on their similarity in expression profiles over a number of experimental treatments (attributes). The promoter regions of the genes in each group are then analyzed for common sequence motifs that are conserved and/or overrepresented [198] using methods described in Section 5.

Segal *et al.* [199] developed a probabilistic graphical model that integrates both the gene expression measurements and DNA sequence data into a unified model. They take each cluster resulting from gene expression data as a module, and search for a common motif from the upstream of genes in the module. They then iteratively refine the model by moving member genes in and out of the module and through expectation maximization to optimize the extent to which the expression profile can be predicted transcriptionally by the motif profile. Park *et al.* [200] take a reverse order of the procedures; they tried to find whether the genes with similar promoter regions are in fact co-expressed based on microarray data. Bussemaker *et al.* [197] fit the gene expression ratio to a collection of sequence motifs, each of which contributes a fixed increment to the gene expression, and selects the most statistically significant motifs from a set of all oligomers up to a specified length, all dimers, and all groups thereof based on sequence alignment. Pilpel *et al.* [201] take a similar approach, but pay more attention to identifying synergistic motif combinations that control the gene expression profile. They first identify all genes containing each motif in their promoters, and then use the expression profiles of the genes whose promoters contain a particular motif A (or a motif combination) to evaluate the effect of A on gene expression. For each motif A, an expression coherence score (EC(A)) is calculated to measure the similarity of all genes containing A under different experimental conditions. A pair of motifs (A, B) is considered "synergistic" if the expression coherence score of genes containing both motifs (EC(AB)) is greater than that of genes containing either alone (EC(A/B) or EC(B/A), here "/" reads as "not"), that is

$$\mathrm{EC}(AB) > \mathrm{EC}(A/B) \text{ and } \mathrm{EC}(AB) > \mathrm{EC}(B/A).$$

In this way, a statistically significant motif combination is found through exploring the effect on gene expression profiles of adding or subtracting motif(s) from particular motif combinations. Zhu *et al.* [202] take this idea one step further by considering orientation and positional constraints in each motif combination. They also take into consideration of phylogenetic conservation and statistical overrepresentation. Starting from an anchor motif, their algorithm first discovers significantly enriched and phylogenetically conserved neighboring motif(s) in a defined sequence window (50 or 100 bp), and then examines the functional significance of their physical proximity through the assessment of similarity in expression profiles.

To study transcriptional co-regulation under different conditions, Ihmels *et al.* [203] treated both the co-regulated genes and the experimental conditions that trigger this co-regulation as a combination termed "transcription module" (TM) and developed an algorithm called the "signature algorithm" (SA, Fig. **4**). First, the SA receives a set of genes as input and identifies the experimental conditions under which the input genes are co-expressed most tightly. An average change in expression of the input genes under each condition is calculated as the "condition score". Only conditions with a large score are selected. Second, genes whose expression under the selected conditions is significant are chosen. Using the condition scores as weights, weighted average change in expression over the selected conditions is calculated as the "gene score". Only genes with a large score are selected. Finally, common motif(s) from each TM can be identified by using methods described in Section 5.

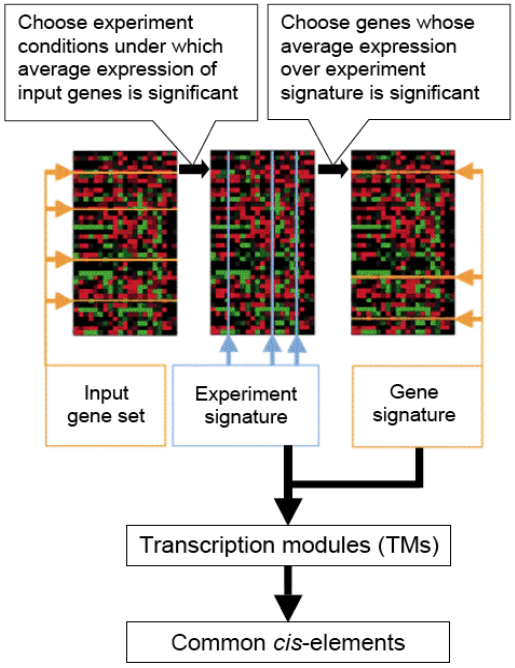

Fig. (4). Signature algorithm. Partially redrawn from [203].

One obvious problem in SA is that the output is dependent on the set of input genes. In this regard, Bergmann *et al.* [204] introduced an iterative version of SA (ISA) that takes a sufficiently large set of randomly selected genes (or conditions) as initial input to SA, and then takes the output of the previous SA as the input to the subsequent SA. The ISA iteratively refines the genes and conditions until they match a rigorous definition of a TM, which is a combination of the threshold for gene scores and the threshold for condition scores, known as the threshold coefficient. More recently, Ihmels *et al.* [205, 206] proposed a new scheme that provides a global decomposition of the expression data into a hierarchy of transcription units at various resolutions. This approach is suitable for cases where no *a priori* information is available, and can also be used to integrate external data in a natural way. It is obvious that the modules recovered by ISA depend strongly on the definition of a TM. Too rigorous a definition will result in excluding weaker TMs. To find all relevant TMs, the thresholds must be varied. Kloster *et al.* [207] developed the progressive ISA (PISA) to allow unsupervised identification of both large and small TMs through sequentially eliminating strong modules, so that weaker ones can be found.

ChIP-chip methods (Section 4) are popular for studying genome-wide protein-DNA interactions and transcription regulation. However, it can only map the probable protein-DNA interaction loci represented by the microarray, but not to the exact binding sites [208]. MDscan [208] uses the word enumeration and PWM updating (Section 5.2) to examine the ChIP-chip-selected sequences and search for the binding motifs. Most recently, Leung *et al.* [209] introduced a binding energy based motif finding algorithm (EBMF). They consider a scenario that multiple copies of a particular DNA fragment s_i are mixed with multiple copies of a particular TF. At the equilibrium state, some copies of the DNA

fragment are bound by the TFs while others are free. Using the binding reaction modeled by $TF + s_i \Leftrightarrow TF \bullet s_i$ [210], the average binding energy (e_i) between the TF and DNA is $e_i = -\ln(K_{eq})$, where the binding constant $K_{eq} = [TF \bullet s_i] / [TF][s_i]$, $TF \bullet s_i$ is the number of bound copies and $[TF][s_i]$ is the number of free copies. From yeast ChIP-chip data [79], they got the color ratio (*Cy5/Cy3*) of each sequence *i* as the binding constant K_{eq} to calculate the binding energy e_i, then take both the e_i values and sequences as input to the EBMF algorithm to look for binding motifs.

7. DISCUSSION

Identification of *cis*-regulatory elements is essential for deciphering gene regulatory machinery. Over the past three decades, significant progress has been achieved in the understanding of transcription factor interactions with DNA either through wet-lab experimentations, computational investigations, or a combination of both. The major progressions in this area are (1) the combination of chromatin immunoprecipitation techniques with high throughput oligo-nucleotide DNA microarray (ChIP-chip), which is then coupled with computational approaches, (2) protein binding microarray technology, (3) identification of *cis*-regulatory modules, (4) comparative genomic approaches, and (5) incorporation of gene expression profiles.

Yet, discovery of *cis*-regulatory elements has a long way to go. Many of the computational methods are developed based on statistical significance. Caution should be taken in the interpretation of results solely based those methods. Many TFBSs are not necessarily maximally overrepresented in the promoters as compared to genome-wide distribution or statistical randomness. Still, some are under-represented, such as the degenerate CRE motif TTACGTAA, but the signal can be revealed through a sub-promoter regional survey [126] and/or wet-lab experimentation.

Most methods developed so far are built based on yeast or other non-mammalian systems. Many of these methods do not work as well with mammalian systems. Genes in mammalian genomes are more dispersed with a greater proportion of intergenic sequences than those in yeast. Functional roles of the intergenic regions are rarely known even though some information has been revealed in recent years [211, 212]. Many *cis*-regulatory elements are dispersed in various regions including in 5' UTR, 3' UTR and the protein coding exons (Sections 3.2, 3.3). Strategies combining computational methods with microarray gene expression profiling and phylogenetic footprinting show signs of success for mammalian systems. Further development in this direction would enable a better understanding of *cis*-regulatory system in humans.

Use of structural information on proteins or related protein-DNA complexes has been seen in recent years. This information has come from either analyses of datasets of well-characterized protein-DNA interactions, computer modeling, or wet-lab experiments [17, 107]. An integrated approach that combines all sources of information such as phylogenetic footprinting, gene microarray, modularity of *cis*-elements, ChIP-chip, protein binding microarray, structural information, etc. will lead to a defined identification of *cis*-regulatory elements.

What makes discovery of *cis*-elements even harder is the fact that functions of some TFBSs are cell type or condition dependent. They are expressed and functional in one cell type at one condition but do not show any activity in another cell type or in the same cell type under a different condition. Also co-expressed genes are not necessarily co-regulated; genes whose promoters contain the same TFBSs do not necessarily have identical expression profiles. Groups defined by a common motif are not mutually disjoint. These problems have to be resolved at the systems biology level. Cellular biochemo-dynamic properties, which affect affinity between TFs and TFBSs and between TFs themselves, and secondary and tertiary structure of TFs, are important sources of information contributing to transcriptional regulation and should be considered in future research.

ACKNOWLEDGEMENTS

I wishes to thank Bob Orchard, Fazel Famili, Brandon Smith at NRC, Xuhua Xia at University of Ottawa and four anonymous reviewers for constructive comments and criticisms on the manuscript. There are many other wonderful works that are not cited in this article due to limited space. For this, I apologize. This is National Research Council Canada publication NRC 48464.

REFERENCES

[1] Ebright EH. RNA polymerase: structure similarity between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* **2000**; 304: 687-98.

[2] Carey M, Smale ST. Transcriptional regulation in eukaryotes: concepts, strategies and techniques. Cole Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2000.

[3] Werner T. Model for prediction and recognition of eukaryotic promoters. *Mamm Genome* **1999**; 10: 168-75.

[4] Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Ann Rev Biochem* **2003**; 72: 449-79.

[5] Pedersen AG, Baldi P, Chauvin Y, Brunak S. The biology of eukaryotic promoter prediction – a review. *Computer and Chemistry* **1999**; 23: 191-207.

[6] Davidson EH, Rast JP, Oliveri P, *et al*. A genomic regulatory network for development. *Science* **2002**; 295: 1669-78.

[7] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **2003**; 20: 1377-419.

[8] Butler JE, Kadonaga JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **2002**; 16: 2583-92.

[9] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Proc Natl Acad Sci* **1986**; 83: 4-8.

[10] Goldberg ML. Sequence analysis of *Drosophila* histone genes, PhD thesis, Stanford University, Stanford, CA 1979.

[11] Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P. Promoter sequences of eukaryotic protein-coding genes. *Science* **1980**; 209: 1406-14.

[12] O'Shea-Greenfield A, Smale ST. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J Biol Chem* **1992**; 267: 1391-402.

[13] Burke TW, Kadonaga JT. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **1996**; 10: 711-24.

[14] Kutach AK, Kadonaga JT. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol* **2000**; 20: 4754-64.

[15] Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **1998**; 12: 34-44.

[16] Evans R, Fairley JA, Roberts SGE. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev* **2001**; 15: 2945-9.

[17] Woychik NA, Hampsey M. The RNA polymerase II machinery: structure illuminates function. *Cell* **2002**; 108: 453-63.

[18] Suzuki Y, Tsunoda T, Sese J, *et al*. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res* **2001**; 11: 677-684.

[19] Ohler U, Niemann H. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trend Genet* **2001**; 17:56-60.

[20] Goodyer CG, Sogopolos G, Schwartzbauer G, Zheng H, Hendy GN, Menon PK. Organization and evolution of the human growth hormone receptor 5'-flanking region. *Endocrinology* **2001**; 142: 1923-34.

[21] Lewis BA, Kim TK, Orkin SH. A downstream element in the human-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci* **2000**; 97: 7172-7.

[22] Nakatani Y, Brenner M, Freese E. An RNA polymerase II promoter containing sequences upstream and downstream from the RNA start point that direct initiation of transcription from the same site. *Proc Natl Acad Sci* **1990**; 87: 4289-93.

[23] Nakatani Y, Horikoshi M, Brenner M, *et al.* A downstream initiation element required for efficient TATA box binding and *in vitro* function of TFIID. *Nature* **1990**; 384: 86-8.

[24] Locker J. Transcription factors. Academic Press, San Diego, CA 2001.

[25] Arnone ML, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **1997**; 124:1851-64.

[26] Simon J, Peifer M, Bender W, O'Connor M. Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. *EMBO J* **1990**; 9: 3945-56.

[27] Kammandel B, Chowdhury K, Stoykova A, Aparicio S, Brenner S, Gruss P. Distinct *cis*-essential modules direct the time–space pattern of the *Pax6* gene activity. *Dev Biol* **1999**; 205: 79-97.

[28] Nielsen LB, Kahn D, Duell T, Weier HUG, Taylor S, Young SG. Apolipoprotein B gene expression in a series of human apolipoprotein B transgenic mice generated with reca-assisted restriction endonuclease cleavage-modified bacterial artificial chromosomes. An intestine-specific enhancer element is located between 54 and 62 kilobases 5' to the structural gene. *J Biol Chem* **1998**; 273: 21800-7.

[29] Calhoun VC, Stathopoulos A, Levine M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila Antennapedia* complex. *Proc Natl Acad Sci* **2002;** 99: 9243-7.

[30] Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJC, Davidson EH. Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. *Dev Biol* **2002**; 146:148-61.

[31] Bamshad MJ, Mummidi S, Gonzalez E, *et al*. A strong signature of balancing selection in the 5' *cis*-regulatory region of CCR5. *Proc Natl Acad Sci* **2002**; 99: 10539-44.

[32] DiLeone RJ, Russell LB, Kingsley DM. An extensive 3' regulatory region controls expression of *Bmp5* in specific anatomical structures of the mouse embryo. *Genetics* **1998**; 148: 401-8

[33] Neznanov N, Umezawa A, Oshima RG. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J Biol Chem* **1997**; 272: 27549-57.

[34] Sandrelli F, Campesan S, Rossetto MG, *et al*. Molecular dissection of the 5' region of *no-on-transientA* of *Drosophila melanogaster* reveals *cis*-regulation by adjacent *dGpi1* sequences. *Genetics* **2001**; 157: 765-75.

[35] Lettice LA, Horikoshi T, Heaney SJH, *et al*. Disruption of a long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. *Proc Natl Acad Sci* **2002**; 99: 7548-53

[36] Wagner A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **1999**; 15: 776-84.

[37] Boehlk S, Fessele S, Mojaat A, *et al.* ATF and Jun transcription factors, acting through an Ets / CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. *Eur J Immunol* **2000**; 30: 1102-12.

[38] Werner T, Fessele S, Maier H, Nelson PJ. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J* **2003**; 17:1228-37.

[39] Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acid Res* **2003**; 31: 6016-26

[40] Terai G, Takagi T. Predicting rules on organization of *cis*-regulatory elements, taking the order of elements into account. *Bioinformatics* **2004;** 20: 1119-28.

[41] Yuh CH, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **1998**; 229: 1896-902.

[42] Klingenhoff A, Frech K, Werner T. Regulatory modules shared within gene classes as well as across gene classes can be detected by the same *in silico* approach. *In Silico Biol* **2000**; 1: 0020.

[43] Fessele S, Boehlk S, Mojaat A, *et al.* Molecular and *in silico* characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J* **2001**; 15: 577-9.

[44] Wang Q, Sigmund CD, Lin JJC. Identification of *cis* elements in the cardiac troponin T gene conferring specific expression in cardiac muscle of transgenic mice. *Circ Res* **2000**; 86: 478-84.

[45] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* **1987**; 196: 261-82.

[46] Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **1997**; 7: 399-406.

[47] Antequera F, Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci* **1993**; 90: 11995-9.

[48] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**; 409: 860-921.

[49] Venter JC, Adams MD, Mayers E, *et al.* The sequence of human genome. *Science* **2001**; 291: 1304-51.

[50] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**; 420: 520-62.

[51] Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci* **2002**; 99: 3740-5.

[52] Blake MC, Jambou RC, Swick AG, Kahn JW, Azizkhan JC. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol Cell Biol* **1990**; 10: 6632-41.

[53] Antequera F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* **2003**; 60: 1647-58.

[54] Deb-Rinker P, Ly D, Jezierski A, Sikorska M, Walker PR. Sequential DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. *J Biol Chem* **2005**; 280: 6257-60.

[55] Lande-Diner L, Cedar H. Silence of the genes-mechanisms of long-term repression. *Nat Rev Genet* **2005**; 6: 648-54.

[56] Bird A. DNA methylation patterns and epigenetic memory. *Gene Dev* **2002**; 16: 6-21.

[57] Weber M, Davies JJ, Wittig D, *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **2005**; 37: 853 – 62.

[58] Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **2004**; 429: 457-63.

[59] Robertson KD. DNA methylation and human disease. *Nat Rev Genet* **2005**; 6: 597-610.

[60] Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol* **2005**; 45: 629-56.

[61] Laird PW. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **2003**; 3: 253-66.

[62] Ushijima T, Okochi-Takada E. Aberrrant methylations in cancer cells: where do they come from? *Cancer Sci* **2005**; 96: 206-11.

[63] Galas D, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **1978**; 5: 3157-70.

[64] Siebenlist U, Simpson RB, Gilbert W. *E. coli* RNA polymerase interacts homologously with two different promoters. *Cell* **1976**; 20: 269-81.

[65] Ogata R, Gilbert W. Contacts between the *lac* Repressor and Thymines in the *lac* Operator. *Proc Natl Acad Sci* **1977**; 74: 4973-6.

[66] Ross W, Landy A, Kikuchi Y, Nash H. Interaction of int protein with specific sites on λ *att* DNA. *Cell* **1979**; 18: 297-307.

[67] Van Dyke MW, Hertzberg RP, Dervan PB. Map of distamycin, netropsin, and actinomycin binding sites on heterogeneous DNA: DNA cleavage-inhibition patterns with methidiumpropyl-EDTA·Fe(II). *Proc Natl Acad Sci* **1982;** 79: 5470-4.

[68] Maxam AM, Gilbert W. A New Method for Sequencing DNA. *Proc Natl Acad Sci* **1977**; 74: 560-64.

[69] Becker MM, Wang JC. Use of light for footprinting DNA *in vivo*. *Nature* **1984**; 309: 682-7.

[70] Ciarrocchi G, Pedrini AM. Determination of pyrimidine dimer unwinding angle by measurement of DNA electrophoretic mobility. *J Mol Biol* **1982**; 155: 177-83.

[71] Church GM, Gilbert W. Genomic Sequencing. *Proc Natl Acad Sci* **1984**; 81: 1991-5.

[72] Selleck SB, Majors J. Photofootprinting *in vivo* detects transcription-dependent change in yeast TATA boxes. *Nature* **1987**; 325: 173-7.

[73] Axelrod JD, Majors J. An improved method for photofootprinting yeast genes *in vivo* using taq polymerase. *Nucleic Acid Res* **1989**; 17: 171-83.

[74] Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: a probe for *in vivo* chromatin structures. *Proc Natl Acad Sci* **1985**; 82: 6470-4.

[75] McGhee JD, Von Hippel PH. Formaldehyde as a probe of DNA structure. I. Reaction with exocyclic amino groups of DNA bases. *Biochemistry* **1975**; 14: 1281-96.

[76] McGhee JD, Von Hippel PH. Formaldehyde as a probe of DNA structure. II. Reaction with endocyclic imino groups of DNA bases. *Biochemistry* **1975**; 14: 1297-303.

[77] Orlando V, Strutt H, Paro R. Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. *Methods* **1997**; 11: 205-14.

[78] Kuo MH, Allis CD. *In vivo* cross-linking and immunoprecipitation for studying dynamic protein: DNA associations in a chromatin environment. *Methods* **1999**; 19: 425-33.

[79] Ren B, Robert F, Wyrick JJ, *et al*. Genome-wide location and function of DNA binding proteins. *Science* **2000**; 290: 2306-9.

[80] Orlando V. Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **2000**; 25: 99-104.

[81] Horak CE, Snyder M. ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods In Enzymology* **2002**; 350: 469-83.

[82] Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ. Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol Cell Biol* **2001**; 21: 6820-32.

[83] Lee TI, Rinaldi NJ, Robert F, *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **2002**; 298: 799-804.

[84] Odom DT, Zizlsperger N, Gordon DB, *et al*. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **2004**; 303:1378-81.

[85] Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci* **2001**; 98: 7158-63.

[86] Mukherjee S, Berger MF, Jona G, *et al*. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **2004**; 36: 1331-9.

[87] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial Analysis of Gene Expression. *Science* **1995**; 270: 484-7.

[88] Impey S, McCorkle SR, Cha-Moistad H, *et al*. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **2004**; 119: 1041-54.

[89] Roh T, Ngau WC, Cui K, Landsman D, Zhao K. High-resolution genome-wide mapping of histone modifications. *Nat Biotch* **2004**; 22: 1013-6.

[90] Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **1990**; 249: 505-10.

[91] Blackwell TK, Weintraub H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* **1990**; 250: 1104-10.

[92] Wright WE, Binder M, Funk A. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Bio* **1991**; 11: 4104-10.

[93] Garner MM, Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res* **1981**; 9: 3047-60.

[94] Field DS, He Y, Al-Uzri AY, Stormo GD. Qautitative specificity of the Mnt repressor. *J Mol Biol* **1997**; 271: 178-94.

[95] Shimada T, Fujita N, Maeda M, Ishihama A. Systematic search for the Cra-binding promoters using genomic SELEX system. *Genes to Cells* **2005**; 10: 907-18.

[96] Beinoraviciute-Kellner R, Lipps G, Krauss G. *In vitro* selection of DNA binding sites for ABF1 protein from Saccharomyces cereviside. *FEBS* **2005**; 579: 4535-40.

[97] Sompayrac L. Danna KJ. Method to identify genomic targets for DNS binding proteins. *Proc Natl Acad Sci* **1990**; 87: 3274-8.

[98] Kinzler KW, Vogelstein B. Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Res* **1989**; 25: 3645–53.

[99] Liu J, Stormo GD. Combining SELEX with quantitative assays to rapid obtain accurate models of protein-DNA interactions. *Nucleic Acids Res* **2005**; 33: e141.

[100] Hybarger G, Bynum J, Williams RF, Valdes JJ, Chambers JP. A microfluidic SELEX prototype. *Anal Bioanal Chem* **2006**; 384: 191-8.

[101] Woodbury CP, Von Hippel PH. On the determination of deoxyribonucleic acid-protein interaction parameters using the nitrocellulose filter-binding assay. *Biochemistry* **1983**; 22: 4730-7.

[102] Choo Y, Klug A. A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nucleic Acids Res* **1993**; 21: 3341-46.

[103] Bowen B, Steinberg J, Laemmli UK, Weintraub H. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Res* **1980**; 8: 1–20.

[104] Hanes SD, Brent R. A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **1991**; 251: 426-30.

[105] Meighen EA. Molecular biology of bacterial bioluminescence. *Microbiol Rev* **1991**; 55: 123-42.

[106] Duanmu Z, Kocarek TA, Runge-Morris M. Transcriptional regulation of rat hepatic aryl sulfotransferase (SULT1A1) gene expression by glucocorticoids. *Drug Metab Dispos* **2001**; 29: 1130-5.

[107] Bulyk ML. Computational prediction of transcription-factor binding site locations. *Genome Biology* **2003**; 5: 201.

[108] Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, Van de Peer Y. Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. *Plant Physiol* **2003**; 132: 1162-76.

[109] Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genetics* **2004**; 5: 276-87.

[110] Pavesi G, Mauri G, Pesole G. *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform* **2004**; 5: 217-36.

[111] Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotech* **2005**; 23: 137-44.

[112] Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* **2000**; 16: 16-23.

[113] Pan Y, Pylatuik JD, Ouyang J, Famili A, Fobert PR. Discovery of functional genes for systemic acquired resistance in *Arabidopsis thaliana* through integrated data mining. *J Bioinf Comput Biol* **2004**; 2: 639-55.

[114] http://rulai.cshl.edu/SCPD/ Oct. 29, 2005.

[115] Matys V, Fricke E, Geffers R, *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **2003**; 31: 374-8.

[116] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **2004**; 32: D91–4.

[117] Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol* **1986**; 188: 415-31.

[118] Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* **1951**; 22: 79-86.
[119] Shannon CE. A mathematical theory of communication. *Bell System Tech J* **1948**; 27: 379-423, 623-56.
[120] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. **1990**; 18: 6097-100.
[121] Thijs G, Marchal K, Lescot M, *et al.* A Gibbs Sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *J Comput Biol* **2002**; 9: 447-64.
[122] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology: application to protein modeling. *J Mol Biol* **1994**; 235: 1501-31.
[123] Eddy SR. Hidden Markov models. *Curr Opin Struct Biol* **1996**; 6: 361-5.
[124] Eddy SR. Profile hidden Markov models. *Bioinformatics* **1998**; 14: 755-63.
[125] Fukunaga K. Introduction to Statistical Pattern Recognition (2nd Ed). Academic Press, San Diego, CA 1990.
[126] Pan Y, Smith B, Fang H, Famili FA, Sikorska M, Walker PR. Selection of putative *cis*-regulatory motifs through regional and global conservation. In: Proceedings of the 2004 IEEE Computational System Bioinformatics Conference (CSB2004), 16-19 August 2004, Stanford, CA, USA 2004; 684-5.
[127] Lawrence CE, Reilly AA. 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.
[128] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignments. *Science* **1993**; 262: 208-14.
[129] Bailey EL, Elkan C. Fitting a mixture model by expectation maximization to discovery motifs in biopolymers. In: Proceeding of the Second International Conference on intelligent System for Molecular Biology. AAAI Press 1994; 28-38.
[130] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data *via* the EM algorithm. *J Roy Stat Soc Ser B* **1977**; 39: 1-38.
[131] Liu JS, Lawrence CE, Neuwald A. Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *J Amer Statist Assoc* **1995**; 90: 1156-70.
[132] Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* **1984**; 6: 721-41.
[133] Gelfand A, Smith A. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* **1990**; 85: 398-409.
[134] Zhang MQ. Computational methods for promoter recognition. In: Jiang T, Xu Y, Zhang MQ Eds, Current Topics in Computational Molecular Biology. MIT Press, Cambridge, Massachusetts 2002; 249-68.
[135] Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* **1995**; 4: 1618-32.
[136] Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **2000**; 296: 1205-14.
[137] Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* **2001**; 127-38.
[138] Thijs G, Lescot M, Marchal K, *et al.* A higher order background model improves the detection of regulatory elements by Gibbs Sampling. *Bioinformatics* **2001**; 17: 1113-22.
[139] Thompson W, Rouchka EC, Lawrence CE. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res* **2003**; 31: 3580-5.
[140] Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* **2005**; 21: 2240-5.
[141] Jensen AT, Liu XS, Zhou Q, Liu JS. Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Stat Sci* **2004**; 19: 188-204.
[142] Jensen AT, Liu JS. BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics* **2004**; 20: 1557-64.
[143] Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **1999**; 15: 563-77.
[144] van Helden J, André Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **1998**; 281: 827-42.
[145] Sinha S, Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* **2002**; 30: 5549-60.
[146] Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **2001**; 17: S207-14.
[147] Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **2004**; 32: W199-203.
[148] Sinha S. Discriminative motifs. *J Comput Biol* **2003**; 10: 599-615.
[149] Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci* **2005**; 102: 1560-5.
[150] Bussemaker HJ, Li H, Siggia ED. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci* **2000**; 97:100096-100.
[151] Rajewsky N, Vergassola M, Gaul U, Siggia ED. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **2002**; 3: 30.
[152] Gupta M, Liu JS. Discovery of conserved sequence patterns using a stochastic dictionary model. *J Am Stat Asso* **2003**; 98: 55-65.
[153] Berman BP, Nibu Y, Pfeiffer BD, *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* **2002**; 99: 757-62.
[154] Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. Decoding human regulatory circuits. *Genome Res* **2004**; 14: 1967-74.
[155] Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **1998**; 278: 167-81.
[156] Krivan W, Wasserman WW. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* **2001**; 11: 1559-66.
[157] Bailey TL, Noble WS. Searching for statistical significant regulatory modules. *Bioinformatics* **2003**; 19: ii16-25.
[158] Klingenhoff A, Frech K, Quandt K, Werner T. Functional promoter modules can be detected by formal modules independent for overall nucleotide sequence similarity. *Bioinformatics* **1999**; 15: 180-6.
[159] Frith MC, Hansen U, Weng Z. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* **2001**; 17: 878-89.
[160] Claverie JM, Sauvaget I. Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Comput Appl Biosci* **1985;** 1: 95-104.
[161] Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* **2002**; 30: 332-4.
[162] Kreiman G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res* **2004**; 32: 2889-900.
[163] Gupta M, Liu JS. *De novo cis*-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci* **2005**; 102: 7079-84.
[164] Quandt K, Grote K, Werner T. GenomeInspector: a new approach to detect correlation patters of elements on genomic sequences. *Comput Appl Biosci* **1996**; 12: 405-13.
[165] Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory module. *Bioinformatics* **2003**; 19: i292-301.
[166] Philippakis AA, He FS, Bulyk ML. ModuleFinder: a tool for computational discovery of *cis* regulatory modules. *Pac Symp Biocomput* **2005**; 519-30.
[167] Zhou Q, Wong WH. CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci* **2004**; 101: 12114-9.
[168] Liang F, Wong YH. Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Stat Sinica* **2000**; 10: 317-42.
[169] http://www.ebi.ac.uk/genomes/ Oct 29, 2005.

[170] Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazone eukaryotes. *Nat Rev Genetics* **2003**; 4: 251-62.

[171] Tagle D, Koop B, Goodman M, Slightom J, Hess D, Jones R. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **1988**; 203: 439-455

[172] Dickmeis T, Müller F. The identification and functional characterisation of conserved regulatory elements in developmental genes. *Brief Funct Genomics Proteomics* **2005**; 3: 332-50.

[173] Blanchette M, Tompa M. Discovery of regulatory elements by comparative method for phylogenetic footprinting. *Genome Res* **2002**; 12: 739-48.

[174] Prakash A, Blanchette M, Sinha S, Tompa M. Motif discovery in heterogeneous sequence data. *Pac Symp Biocomput* **2004**; 248-59.

[175] Schwartz S, Zhang Z, Frazer KA, *et al.* PipMaker—a Web server for aligning two genomic DNA sequences. *Genome Res* **2000**; 10: 577-86.

[176] Schwartz1 S, Kent WJ, Smit A, *et al*. Human–mouse alignments with BLASTZ. *Genome Res* **2003**; 13: 103-7.

[177] Brudno M, Do CB, Cooper GM, *et al*. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **2003**; 13: 721-31.

[178] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene Oct 29, 2005.

[179] Moses AM, Chiang DY, Eisen MB. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput* **2004**: 324-35.

[180] Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res* **2004**; 14: 451-8.

[181] Sinha S, Blanchette M, Tompa M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **2004**; 5: 170.

[182] Siddharthan R, Siggia ED, van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **2005**; 1: 534-56.

[183] Siepel A. Haussler D. Phylogenetic hidden Markov models. In: Nielsen R Ed, Statistical Methods in Molecular Evolution. Springer, New York 2005; 325–351.

[184] Siepel A, Bejerano G, Pedersen JS, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **2005**; 15:1034-50.

[185] Margulies EH, Blanchette M, NICS comparative Sequencing Program, Haussler D, Green ER. Identification and characterization of multi-species conserved sequences. *Genome Res* **2003**; 13: 2507-18.

[186] Elnitski L, Hardison RC, Li J, *et al*. Distinguishing regulatory DNA from neutral sites. *Genome Res* **2003**; 13: 64-72.

[187] Kolbe D, Taylor J, Elnitski L, *et al*. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* **2004**; 14: 700-7.

[188] King DC, Talor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. Evaluation of regulatory potential and conservation scores for detecting cis regulatory modules in aligned mammalian genome sequences. *Genome Res* **2005**; 15: 1051-60.

[189] Grad YH, Roth FP, Halfon MS, Church GM. Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*. *Bioinformatics* **2004**; 20: 2738-50.

[190] Sharan R, Ovcharenko I, Ben-Hur A, Karp RM. CRÈME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **2003**; 19: i283-91.

[191] Sandelin A, Wasserman WW, Lenhard B. ConSite: Web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res* **2004**: 32: W249-52.

[192] Berezikov E, Gurywv V, Plasterk RHA, Cuppen E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res* **2004**: 14: 170-8.

[193] Corcoran DL, Feingold E, Benos PV. FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res* **2005**; 33: W442-6.

[194] Siggia ED. Computational methods for transcriptional regulation. *Curr Opin Genet Dev* **2005**; 15: 214-21.

[195] Dieterich C, Wang H, Ratertschak K, Luz H, Vingron M. CORG: a database for Comparative Regulatory Genomics. *Nucleic Acids Res* **2003**; 31: 55-7.

[196] Segal E, Shapira M, Regev A, *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **2003**; 34: 166-76.

[197] Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. *Nat Genet* **2001**; 27: 167-71.

[198] Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell **2004**; 117: 185-98.

[199] Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **2003**; 19: i273-82.

[200] Park PJ, Butte AJ, Kohane IS. Comparing expression profiles of genes with similar promoter regions. *Bioinformatics* **2002**; 18: 1576-84.

[201] Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **2001**; 29: 153-9.

[202] Zhu Z, Shendure J, Church GM. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res* **2005**; 15: 848-55.

[203] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **2002**; 31: 370-7.

[204] Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E* **2003**; 67: 031902.

[205] Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisia*e. *Nat Biotech* **2004**; 22: 86-92.

[206] Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **2004**; 20: 1993-2003.

[207] Kloster M, Tang C, Wingree NS. Finding regulatory modules through large-scale gene-expression data analysis. *Bioinformatics* **2005**; 21: 1172-9.

[208] Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein−DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotech* **2002**; 20: 835-9.

[209] Leung HCM, Chin FYL, Yiu SM, Rosenfeld R, Tsang WW. Finding motifs with insufficient number of strong binding sites. *J Comput Biol* **2005**; 12: 686-701.

[210] Klotz IM. Introduction to biomolecular energetics: including ligand-receptor interactions, Academic Press, London 1986.

[211] He L, Hannon GJ. MicroRNAs: Small RNAs with a big role in gene regulation. *Nat Rev Genet* **2004**; 5:522-31.

[212] Shabalina SA, Spiridonov NA. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol* **2004**; 5: 105.