Naturwissenschaften (2003) 90:156–166
DOI 10.1007/s00114-003-0409-4

REVIEW ARTICLE

Wyeth W. Wasserman · William Krivan

In silico identification of metazoan transcriptional regulatory regions

Published online: 27 March 2003

Abstract Transcriptional regulation remains one of the most intriguing and challenging subjects in biomedical research. The catalysis of transcription is a clear example of multiple proteins interacting to orchestrate a biological process, offering a starting point for the study of biological systems. Transcriptional regulation is viewed as one of the principal mechanisms governing the spatial and temporal distribution of gene expression, thus the field of transcriptional regulation provides a natural stage for quantitative studies of multiple gene systems. Building on the body of focused experimental studies and new genomics-driven data, computational biologists are making significant strides in accelerating our understanding of the transcriptional regulatory process in metazoan cells. Recent advances in the computational analysis of the interplay between factors have been fueled by well-defined computational methods for the modeling of the binding of individual transcription factors. We present here an overview of advances in the analysis of regulatory systems and the fundamental methods that underlie the recent developments.

Introduction

At the foundation of metazoan cell biology is the precisely regulated process of transcription, which generates an enormous variety of RNA molecules and consequently a great diversity in protein forms. During the stages of development, spatial and temporal control of transcription is essential for differentiation, which results in diverse cell types with specialized functions. In order to maintain tight control of this process, a diverse array of biochemical mechanisms has evolved, which target the initiation, splicing, and localization of RNA within cells. The initiation of transcription, as the first step of the process, has received significant attention from both experimental and computational biologists. Many of the biochemical mechanisms underlying transcription initiation have been identified and characterized, including chromatin-mediated gene accessibility, the regulated recruitment of the RNA polymerase machinery by transcription factors (TFs), and the targeting of the initiation of transcription to specific nucleotides within each gene (Fig. 1). Research within computational biology has addressed these three areas with varying degrees of success.

Within the scope of this review, it is not possible to cover the enormous breadth of experimental research that has shaped the computational approaches. Where possible the experimental literature will be highlighted in order to provide descriptive examples of discoveries, ideas, and the new methods that will shape the flow of data in the future. For overviews of the experimental work, we recommend a review compendium (e.g. the "*Chromosomes and expression mechanisms*" issue of *Current Opinion in Genetics and Development*), as well as specific topical reviews addressing topics such as: the general machinery (basal) utilized by most genes (Lemon and Tjian 2000), the importance of protein interactions (Wolberger 1999), the modular structure of regulatory regions (Blackwood and Kadonaga 1998; Davidson 2001), the evolution of regulatory controls (Tautz 2000), and the central role of chromatin structure (Cremer et al. 2000; Cremer and Cremer 2001). It should be noted that this review is restricted to transcription mediated by RNA polymerase II (pol-II), which is responsible for the transcription of essentially all protein-coding genes.

At the intersection of computation and biology, it is often difficult for scientists to communicate, due to the complexity and inconsistency of the vocabulary. For the purposes of this report, we will attempt to restrict the

W. W. Wasserman (✉)
Centre for Molecular Medicine and Therapeutics (CMMT),
University of British Columbia,
950 West 28th Avenue, Vancouver, BC, V5Z 4H4, Canada
e-mail: wyeth@cmmt.ubc.ca
Fax: +1-604-8753819

W. Krivan
ZymoGenetics, 1201 Eastlake Avenue East,
Seattle, WA 98102, USA

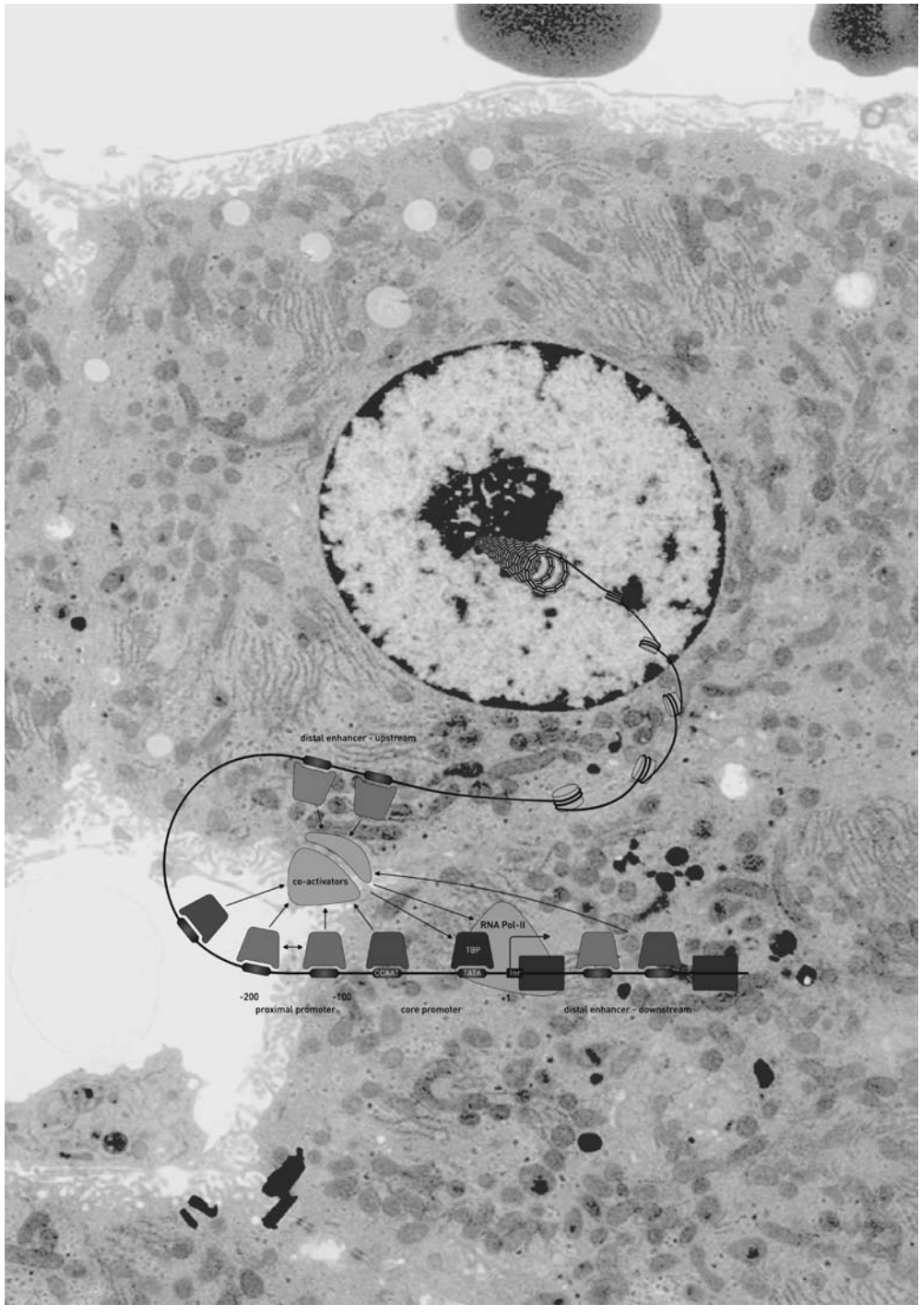

Fig. 1 Overview of the regulated transcription of a gene. As illustrated, groups of transcription factors (*TFs*) bind to sets of transcription factor binding sites (*TFBSs*) within or adjoining genes to activate, amplify, or repress gene expression. Within the figure are examples of numerous regulatory regions, including regions proximal to the promoter, distal regions both upstream and downstream of the transcription start site (*TSS*). The interaction between TFs and the basal transcriptional machinery is partially mediated by co-activating proteins. For working definitions of many of the terms refer to Table 1

transcription-specific vocabulary to a small set of terms describing the biochemistry summarized in Fig. 1 (see Table 1).

Computational advances in the study of transcription have ranged widely, but can be grouped loosely into three overlapping categories: identification of properties associated with regulatory sequences (including the properties of conservation observed in the comparative analysis of orthologous gene sequences), construction and analysis of quantitative models for the binding to DNA of individual TFs, and the identification of combinations of transcription factor binding sites (TFBSs) likely to be associated with regulatory processes.

We provide a brief overview of the past advances, as well as highlighting the new directions that are likely to shape the near future in these areas.

It is noteworthy, however, that the bulk of the computational methods in all three areas are based on *descriptions* of regulatory regions rather than on an *understanding* of the fundamental molecular mechanisms

Table 1 Restricted vocabulary

Term	Definition
Chromatin	association of histone proteins with DNA
Core promoter	region where the RNA polymerase initiates transcription
Distal	location of regulatory elements outside the promoter
Position weight matrix (PWM)	tool for the quantitative characterization of sequence-based properties of transcription factor binding sites
Proximal promoter region	adjacent to the core promoter, approx. –200 to +100 relative to the transcription start site
Regulatory module	short (30–500 bp) region of DNA containing combinations of functional TFBS
Regulatory region	general term referring to one or more regulatory modules
Transcription factor (TF)	protein capable of activating, amplifying, or repressing gene expression; present in complex with DNA
Transcription factor binding site (TFBS)	short (6–20 bp) DNA segment which can be bound by a TF (often referred to as element or cis-acting site)
Transcription start site (TSS)	position where transcription is likely to be initiated

(Claverie 2000). This is not surprising in the light of the complexities of the workings of eukaryotic transcriptional regulation and the lack of understanding thereof. Despite important discoveries about the transcriptional control apparatus in eukaryotes, our knowledge about the details of the regulation of individual genes remains dramatically incomplete (Lemon and Tjian 2000). As chromatin structure presents the most important and arguably the most difficult problem in the computational analysis of transcriptional regulation (Kadonaga 1998), we will highlight some of the constraints which have hindered advancements, and the emerging data which may allow researchers to bridge the chasm between *descriptive* bioinformatics and *hypothesis-driven* computational biology.

Fundamentals

On the most basic level, the mechanism of transcriptional gene regulation is orchestrated by TFs binding to specific segments of DNA. The individual TFs interact with target sites in the DNA to activate, amplify, or repress gene expression. Each TF (or set of closely related TFs) has characteristic binding properties, including the pattern and width of its DNA-binding sites, and the energy with which the target sequences are bound (Stormo and Fields 1998). Computational approaches for the discovery, description, and modeling of the individual binding sites have been well defined by a series of significant studies.

Motif models for the binding of described transcription factors

Individual TFs are known to bind functionally to sequences with diverse sequence characteristics. The set of experimentally identified TFBSs for any given TF can usually be aligned to identify a few strict sequence requirements (presumably direct contact points between the TFs and DNA) and a range of "preferences." Given sufficient binding site examples, numerous studies have demonstrated that quantitative matrix-based approaches can be highly effective in generating accurate models. This subject has been extensively reviewed (Werner 1999; Stormo 2000; Ohler and Niemann 2001), and we will restrict our comments to a few highlights.

From the body of literature, there are several key terms to recognize. A set of known binding sites can be aligned and the frequency of individual nucleotides at each position counted to generate a *count matrix*. Figure 2a shows a count matrix for the NF-κB TF. Here we describe a standard approach to the construction of motifs, in which the individual nucleotide positions are considered to be independent (Stormo 2000). The assumption of statistical independence of the positions is reflected in multiplicative quantities associated with the individual positions. In order to work with additive values from each column, it is preferable to convert count matrices to a logarithmic scale for computational analysis. Many research groups use subtle variations of the conversion function based on the number of representative binding sites that contributed to the frequency matrix. The resulting log-converted matrix has a variety of names; most commonly it is called a *position weight matrix* (PWM, Fickett 1996a; Fig. 2b). Recently, several studies of TFs for which large collections of binding sites are available have suggested that subtle improvements in predictive performance can be achieved by modeling higher-order interactions between positions (Udalova et al. 2002; Bulyk et al. 2002; Roulet et al. 2002; Benos et al. 2002). For visualization of the binding targets of individual TFs, a convenient sequence *logo* (Schneider and Stephens 1990) format has been developed (Fig. 2c). A sequence logo builds on Shannon's theory of information (Shannon 1948) to convert a frequency matrix to an information content matrix, in which each position of the binding profile has a maximum information content of 2 bits.

There are three key observations we would like to highlight from the vast body of literature on the construction of binding profiles. First, Stormo and Fields elegantly demonstrated the link between information theory-based models and thermodynamic binding energy to suggest that, in the best cases, PWMs produce scores correlated with the binding energies of TFs (Stormo and

(a)

Pos	1	2	3	4	5	6	7	8	9	10
A	0	0	1	25	19	7	1	2	2	0
C	0	0	0	0	13	1	2	17	35	36
G	38	38	37	13	1	3	2	0	0	0
T	0	0	0	0	5	27	33	19	1	2

(b)

Pos	1	2	3	4	5	6	7	8	9	10
A	-2.8	-2.8	-2.1	1.3	0.9	-0.4	-2.1	-1.6	-1.6	-2.8
C	-2.8	-2.8	-2.8	-2.8	0.4	-2.1	-1.6	0.8	1.7	1.8
G	1.8	1.8	1.8	0.4	-2.1	-1.3	-1.6	-2.8	-2.8	-2.8
T	-2.8	-2.8	-2.8	-2.8	-0.8	-1.4	1.7	0.9	-2.1	-1.6

(c)

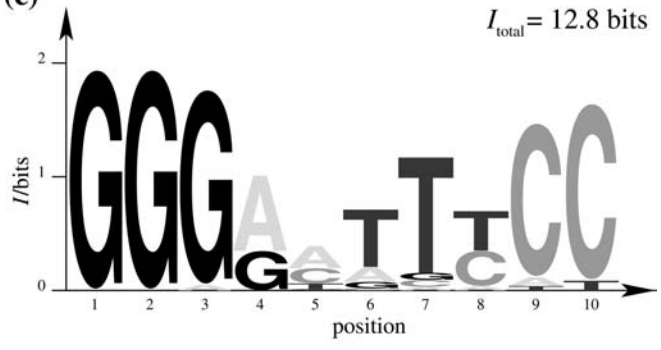

Fig. 2a–c Representations of NF-κB binding sites. **a** Count matrix for NF-κB resulting from an alignment of 38 experimentally verified functional binding sites (B. Lenhard, A. Sandelin, L. Mendoza, N. Jareborg and W.W. Wasserman, unpublished). **b** Position weight matrix (*PWM*). The PWM elements are the logarithms (base 2) of the frequency counts divided by expected counts. A corrective term is added to the counts that reflects the finite sample size (in this example 38) and avoids logarithms of zero ("pseudocounts", cf. Fickett 1996a). **c** Visualization of the count matrix in the form of a sequence logo, showing conservation as well as variation. The height of a stack of letters at each position represents the information content at that position, and the relative sizes of the letters are proportional to the nucleotide frequencies at that position of the alignment

Fields 1998). Second, Tronche et al. (1997) showed that the majority of potential HNF-1 binding sites predicted by a PWM model were indeed bound by the protein in vitro. Finally, Fickett demonstrated that a binding model for Mef2 predicted binding sites on the order of once every 5,000 basepairs (bp) in the human genome (Fickett 1996a). Given the limited subset of genes responsive to Mef2 and the average gene size, this prediction rate indicates there is a poor correlation between *in silico* predictions and sites with function in vivo. Despite these limitations, it can be advantageous to identify potential TFBSs that regulate a gene of interest. Databases of compiled motif models and sequence analysis tools are available to screen genomic sequences for potential TFBSs (see Table 2).

Motif discovery

Sufficient experimental data for the construction of robust matrix models are only available for a small subset of TFs. [For example, 108 non-redundant models for TFs of multicellular eukaryotes are available within the ConSite matrix collection (B. Lenhard, A. Sandelin, L. Mendoza, N. Jareborg and W.W. Wasserman, unpublished).] As these data are laborious to obtain, many future advances are likely to come from ab initio motif finding. Motif discovery has been approached from a variety of perspectives (Ohler and Niemann 2001), which are based on the compilation of sets of sequences known to share characteristic regulatory control mechanisms. Such sets may be composed of regulatory regions of co-regulated genes identified by large-scale expression analysis (e.g. microarrays) or orthologous regulatory regions for a single gene from multiple species. In both cases, the discovery of the motif is based on the expectation that the patterns characteristic of the TFBS will be over-represented in the sequence collection compared with a reference background of non-coding sequence.

Numerous groups have developed *exhaustive* algorithms based on the calculation of the statistical significance for all oligomer frequencies and reporting overrepresented oligomers (Brazma et al. 1998; van Helden et al. 1998; Bussemaker et al. 2000). There are three general limitations with oligo-based approaches: (1) computational time for the analysis of long patterns is prohibitive, (2) the output is a list of oligomers, whereas matrices are more descriptive [current research is addressing the extension of the algorithm described in Bussemaker et al. (2000) to matrices (Hao Li, personal communication)], and (3) TFs often tolerate considerable variation in their binding sites which can reduce the sensitivity of oligo-based analysis [see Ohler and Niemann (2001) for a discussion]. Another set of successful approaches, which circumvents these problems, is based on the extraction of matrix-binding models by performing local multiple sequence alignments of candidate sites. Examples are expectation maximization (EM) methods (Lawrence and Reilly 1990; Bailey and Elkan 1995) and Gibbs sampling (Lawrence et al. 1993). Several recent refinements of the Gibbs sampling algorithm have been developed: correction for biased local sequence characteristics (Workman and Stormo 2000), detection of multiple patterns (GuhaThakurta and Stormo 2001), and analysis of discontinuous or heterogeneous patterns of information content to more accurately model TFBSs (Wasserman et al. 2000). A new generation of approaches is emerging, based upon the observation that many structurally related TFs bind to similar target sequences (Xing et al. 2003; A. Sandelin and W.W. Wasserman, unpublished).

Table 2 Suggested Internet resources

Program	URL
Introductory sites and tutorials	
ITP online: Introduction to molecular biology	http://online.itp.ucsb.edu/online/infobio01/stormo/
ITP online: DNA–protein interactions	http://online.itp.ucsb.edu/online/infobio01/stormo4/
Sequence data sources	
EPD	http://www.epd.isb-sib.ch/
S/MARt DB	http://Transfac.gbf.de/SMARtDB/
NIH Comparative Vertebrate Genome Sequencing Project	http://www.nisc.nih.gov/open_page.html?projects/zooseq.html
Promoter finding/CpG islands	
Berkeley TSS site	http://www.fruitfly.org/seq_tools/promoter.html
PromoterInspector	http://www.genomatix.de/software_services/software/PromoterInspector/PromoterInspector.html
FirstEF	http://rulai.cshl.org/tools/FirstEF/
Motif scans of sequences	
Transfac/ModelInspector	http://transfac.gbf.de/programs/modelinspector/modelinspector.html
Tess	http://www.cbil.upenn.edu/tess/
ConSite	http://www.phylofoot.org
Motif discovery	
Integrated expression/motif analysis	htpp;//www.esat.kuleuven.ac.be/~dna/BioI/Software.html
AlignACE	http://arep.med.harvard.edu/mrnadata/mrnasoft.html
Gibbs sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html
Co-Bind	http://Ural.wustl.edu/~dg/
Modules	
TransRegio	http://www.phylofoot.org.html
CISTER	http://sullivan.bu.edu/~mfrith/cister.shtml
FastM	http://www.gsf.de/biodv/fastm.html
Phylogenetic footprinting	
VISTA/AVID	http://www-gsd.lbl.gov/VISTA
BALSA	http://bayesweb.wadsworth.org/cgi-bin/bayes_align12.pl
PIPMaker	http://bio.cse.psu.edu/pipmaker/
FootPrinter	http://abstract.cs.washington.edu/~blanchem/FootPrinterWeb/FootPrinterInput.pl
LAGAN	http://lagan.stanford.edu

Motif discovery has been extremely successful in conjunction with microarray analysis of expression for yeast (Roth et al. 1998), but motif discovery in human regulatory sequences has been problematic. In yeast, regulatory regions are typically contained in compact segments spanning 200–500 bp upstream of functional open reading frames, while the regulatory sequences of multicellular eukaryotes are located within regions that span tens of kilobases (Simpson et al. 1997). This enormous size of regions potentially containing regulatory elements in conjunction with the low binding specificity of TFs (resulting in a "weak" pattern) results in a large number of false predictions in the analysis of genes from metazoan species. However, the search space can be dramatically reduced by selecting predicted patterns likely to have sequence-specific functions by a process termed "phylogenetic footprinting" to identify regions preferentially conserved over evolution [reviewed in Fickett and Wasserman (2000); see also Levy et al. (2001)]. In particular, human–rodent comparisons have proven a valuable resource for the identification of functional regulatory elements (Wasserman et al. 2000; Levy and Hannenhalli 2002). Figure 3 illustrates the patterns of conservation for the human and hamster cholesterol 7α-hydroxylase (CYP7A1) genes. A recent technique uses a combination of conservation measures and criteria such as the statistical significance of individual sites and clustering of TFBSs (Levy and Hannenhalli 2002).

Phylogenetic footprinting can be applied to gene-specific sets of orthologous sequences for a diverse set of species to identify functional binding elements. This has been widely used with the abundant microbial genomes (McCue et al. 2001; Tan et al. 2001; McGuire et al. 2000; Rajewsky et al. 2002a; McCue et al. 2002; reviewed in Stormo and Tan 2002). This idea has also been successfully developed for eukaryotes with an exhaustive method (Blanchette et al. 2000). [See Blanchette and Tompa (2002) for a recent pattern discovery method that incorporates information about phylogenetic relationships among the input sequences.] The analysis of multiple orthologous gene sequences is of growing importance and interest. The impact of comparative sequence analysis is demonstrated by a study of the interleukin locus (Loots et al. 2000), in which a coordinating regulatory region for the interleukin 4, 5 and 13 genes was found by comparison of 1 Mb of non-coding human and mouse sequences.

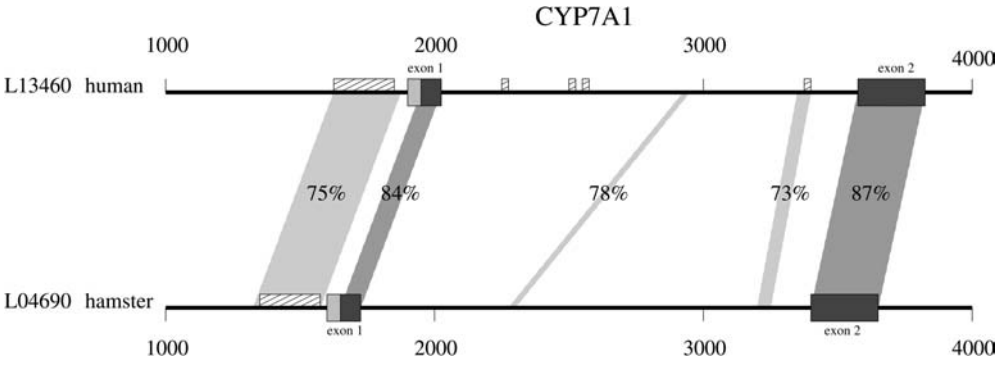

Fig. 3 Phylogenetic footprinting. Comparison of the human and hamster genomic sequences containing the promoter and the first two exons of the CYP7A1 gene. Non-coding/coding portions of exons are shown in *light/dark gray*. Conserved regions are shown as *gray areas* between the sequences. The *streaked rectangles* depict regions containing documented regulatory elements (Krivan and Wasserman 2001 and references therein)

Future directions

As the individual algorithms have reached a level of maturity, integrated analysis systems are emerging which combine multiple techniques (e.g. Thijs et al. 2002a, 2002b; Loots et al. 2002; B. Lenhard, A. Sandelin, L. Mendoza, N. Jareborg and W.W. Wasserman, unpublished). There still are many avenues for continued algorithmic improvements and the development of novel techniques. By incorporating phylogenetic footprinting directly into the modeling process, it may be possible to build a new generation of motif discovery algorithms. Alternatively, our growing knowledge of TFs may enable the improved characterization of binding sites. For instance, past advances have been made in the discrimination of palindromic sequences characteristic of many TFBSs (Lawrence and Reilly 1990; McCue et al. 2001) (e.g. NF-κB in Fig. 2). In addition to the use of sequence information, growing knowledge of the structural properties of transcription factors will impact computational models.

Prediction of regulatory regions

Motif models in isolation are not sufficient for identification of functional binding sites within metazoan genomes. While the TFBS motif models for the binding of individual TFs can accurately reflect the binding properties of the protein in vitro, identification of biologically functional sequences within a genome requires more complex models. Recent advances in bioinformatics for the identification of regulatory regions likely to have a function in vivo have addressed both the sequence properties of regulatory regions and, most importantly, cooperation between TFs.

Three recent advances elucidating the characteristic biochemical properties of regulatory regions are particularly important. First, the demonstration of the cooperative nature of TFs and the presence of multiple binding sites in locally dense clusters (Arnone and Davidson 1997; Blackwood and Kadonaga 1998; Davidson 2001; Pearce et al. 1998) has provided fertile ground for computational studies. Second, the observation that transcription start sites (TSSs) for a significant portion of genes are preferentially localized within CpG islands (Bird 1987; Gardiner-Garden and Frommer 1987) has belatedly altered computational approaches to the promoter-finding challenge. Finally, the preferential conservation of orthologous regulatory sequences over the course of evolution is enhancing the specificity of predictions of many methods (Wasserman et al. 2000).

Core promoters

Diverse approaches with varied results have been applied to identify the specific segments of genes within which transcription initiates. These "core" promoter regions contain sequences upon which pol-II complexes are assembled and the TSS at which extension is initiated. Core promoters support the binding and positioning of the TATA-binding protein (TBP) to a region approximately 30 bp upstream of the start of transcription for many genes, often called the TATA box. "TATA box" refers to the nucleotide pattern best defined by AT-rich motif models for these regions (Bucher 1990; Kraus et al. 1996). The emphasis on the TATA motif in computational analysis has declined in recent years, as numerous groups have demonstrated that TBP and the entire basal complex can be positioned accurately in the absence of TATA-like sequences (Smale 1997). Other suitable target sites may be functionally substituted for the −30 region, such as well-defined initiation regions or downstream signals. In short, what was once viewed as a "general" model for core promoters has grown into a continuous range of promoter models.

The decreased emphasis on the role of TATA-boxes in the experimental literature is paralleled by a recent redefinition of the problems associated with the identification of promoter sequences. Early computational algorithms emphasized the precise identification of TSSs

within genomic sequences (reviewed in Pedersen et al. 1999). In a comprehensive review of the core promoter finding algorithms, Fickett and Hatzigeorgiou (1997) indicated that the "signal" provided by core promoters is insufficient to allow accurate TSS prediction based solely on local sequence features. It was recommended that future approaches split the efforts into two equally difficult problems: the identification of regions likely to contain core promoters and the precise specification of transcription initiation within such regions. Significant recent advances have built on this proposed division, as efforts to identify regions likely to contain core promoters have improved the specificity of predictions. The first algorithms, focused on the discrimination of promoter-containing regions, demonstrated the benefits of the dual approach, providing the first tools with sufficient specificity to motivate the analysis of long genomic sequences (Scherf et al. 2001). While not originally recognized as such, subsequent analysis demonstrated that the increase in specificity could almost entirely be attributed to the detection of CpG islands (Hannenhalli and Levy 2001). A recently described algorithm produced significantly improved performance by fusing detection of CpG islands with analysis of splicing signals consistent with the 3' edges of first exons (Davuluri el al. 2001). Another recently published method is based on detecting a combination of a TATA signal and a G-C-rich region (Down and Hubbard 2002). While the technical details vary between the algorithms, there are two clear observations: (1) promoter-containing regions associated with CpG islands can be predicted with high specificity, and (2) promoters situated outside CpG islands, potentially two out of every three promoters, have proven resilient to detection algorithms. With regard to (2), considerable improvement of specificity for sequences not characterized by high G-C content may be gained from novel methods that explicitly use information about exon positions derived from transcript sequences (Liu and States 2002).

Modeling of regulatory modules

Although both experimental and computational work has long emphasized the importance of individual TFBSs, a new paradigm has been taking root for the interpretation of regulatory regions as clusters of TFBSs. While this view has existed in the biological literature for a long period, an important review clearly presented the case for emphasizing clusters in both experimental and computational analyses (Arnone and Davidson 1997). The burgeoning of many recent advances in the detection of regulatory regions is based on this paradigm shift (Michelson 2002; Halfon and Michelson 2002).

The analysis of combinations of TFBSs has taken two forms: (1) rule-based architectures and (2) detection of segments within genomic sequences containing overrepresentations of potential binding sites. The rule-based architectures build on from studies of composite response elements (Pearce et al. 1998), in which adjoining or overlapping binding sites are known to functionally interact with specific, defined spacing requirements. Composite site computational models for two adjoining TFBSs were applied to combinations of Mef2 and MyoD binding sites (Fickett 1996b) and more recently to the analysis of binding sites of members of the E2F family of TFs in cell cycle genes (Kel et al. 2001). A recent study examined a broad set of possible pairings based on annotated sites in proximity (Hannenhalli and Levy 2002). A system for the analysis of such regulatory modules has been developed for the rule-based detection of pairs of specific TFBSs (Klingenhoff et al. 1999).

Restriction to pairs of binding sites fails to reflect the more general clusters of binding sites identified in laboratory studies (Arnone and Davidson 1997). For most cases in biology, the data are insufficient to establish rigid rules for the architectures of regulatory regions. In one exceptional case, extensive rule-based modeling was used to detect retroviral LTR sequences (Frech et al. 1997). The success in this case was most likely enhanced by the strict spatial constraints imposed upon LTR sequences by the packaging requirements of the virus and the direct evolutionary links between the target sequences. For more general cases in which regulatory sequences may have evolved independently, rule-based approaches are difficult to develop, at least in part due to sparse data. Several projects have used a large set of binding profiles (both matrix and string-based models) to identify statistically significant clusters of binding sites (Crowley et al. 1997; Wagner 1999). In order to identify regions with a tissue-specific biological function, an algorithm was developed that tackles the problem using explicit information about (1) individual TFBSs and (2) their clustering in tissue-specific modules. The method was applied to muscle-specific regulatory regions (Wasserman and Fickett 1998), and, in combination with phylogenetic footprinting, to liver-specific genes (Krivan and Wasserman 2001). In the latter case, only those predictions are reported which fall within regions of significant conservation between human and rodent sequence, reducing the number of false positive predictions. New approaches to the clustered sites problem have been reported (Frith et al. 2001, 2002; Berman et al. 2002; Halfon et al. 2002; Rajewsky et al. 2002b) which offer flexibility in analyzing regulatory regions that may contain multiple binding sites for the same or different TFs.

Despite these developments, none of the module analysis methods addresses the biochemical reality of transcriptional regulation. In order to gain greater insights into the biochemical mechanisms that drive transcription, we will need increasing amounts of data produced by detailed analysis of regulatory sequences for more genes. With this knowledge, future models can be expected to use more flexible combinations of binding sites with increased attention to spacing and nearest neighbor constraints. "Phylogenetic footprinting" has primarily been utilized as a tool to increase the specificity of predictions (Loots et al. 2002; B. Lenhard, A. Sandelin, L.

Mendoza, N. Jareborg and W.W. Wasserman, unpublished), but future efforts should benefit from incorporating conservation analysis into the modeling process at earlier stages.

The validation of computational methods is an important point that has to be addressed in order to assess the value of *in silico* predictions. One possibility is the validation of methods based on existing biological knowledge about utilized test data sets. However, this is not unproblematic, since experiment-based annotations are likely to be dramatically incomplete (Loots et al. 2002). Therefore, the validation by computational means is often indispensable (Crowley et al. 1997; Wasserman et al. 2000). More reliable and informative, but also laborious and expensive, are de novo studies addressing the in vivo validation of computational predictions (Berman et al. 2002; Markstein et al. 2002; Halfon et al. 2002). Ultimately, the measurement of performance of any predictive method should be carefully evaluated to determine whether unrealistic bias is present.

Regulated access: chromatin and transcription

One of the trivia highlights of cellular biology is that the length of human DNA if stretched out end-to-end comes to approximately 1 m. The size of eukaryotic nuclei is on the order of 1 μm, indicating that DNA is highly compacted in order to fit into the nucleus. The compaction is principally performed through the association of histone proteins with DNA in a complex known as chromatin. The fundamental role of chromatin in the transcriptional regulatory process (Wu and Grunstein 2000; Kornberg and Lorch 1999; Kadonaga 1998; Wolffe and Guschin 2000) and the consequent need for the incorporation of higher-dimensional DNA structure information into computational models of transcriptional regulation have long been recognized. Little is known about the chromatin structure of specific regions of DNA beyond the local nucleosome level (Polach and Widom 1996) and consequently there is no widely accepted model for the spatial arrangement of nucleosomes. A large number of critical biochemical questions remain to be answered. The positioning of nucleosomes within regulatory regions is not widely understood (Polach and Widom 1995, 1996; Shim et al. 1998) and little is known as to how TFs gain access to DNA-binding sites in the presence of chromatin (Anderson and Widom 2000). A profound understanding of chromatin structure and its dynamics (McNally et al. 2000), which needs to be incorporated into computational models of gene regulation, is missing at the present time.

Given our lack of understanding of the biochemical processes of chromatin-mediated regulation, opportunities for the creation of meaningful computational approaches have been sparse. Nevertheless, numerous researchers have attempted to produce descriptive and predictive tools for the analysis of chromatin. Efforts have been made to model the physical properties of DNA based on local sequence composition (Ohler et al. 2001). Levitsky et al. (2001) developed a computational method for scoring "nucleosome forming potential." Higher concentrations of segments with elevated "nucleosome forming potential" were detected in promoter regions. This was interpreted as an indication that nucleosome positioning is an important factor in the regulation of gene expression. Several groups have attempted to build predictive models for "scaffold/matrix-attached regions" (S/MAR elements, reviewed in Bode et al. 2000). S/MARs are proposed to be regulatory elements that associate with components of the nuclear matrix and thereby affect both chromatin organization and gene expression. A recent attempt to predict S/MARs in large genomic sequences is described in Frisch et al. (2002).

Research addressing the spatial organization of the nucleus into discrete pockets (Cremer and Cremer 2001; Jackson 1997), has led to intriguing computational contributions. As opposed to cytoplasmic structures, nuclear pockets are not delineated by membranes (Dundr and Misteli 2001), suggesting that other forces must be in effect to preserve their integrity. One hypothesis is that this structural regulation of nuclear architecture is mediated by a set of proteins that create and maintain the pockets. Recent computational modeling supports a passive (indirect) mechanism (Cremer et al. 2000). Complementing experimental data, the computational model suggests that the transcription of genes can maintain accessible regions, while intervening segments of non-transcribed chromosomes are compacted.

One avenue for the analysis of chromatin organization addresses the colocalization of genes with similar expression characteristics. Recent studies have identified significant correlation between the expression patterns of adjacent genes for genes with selective breadth of expression in yeast (Kruglyak and Tang 2000; Cohen et al. 2000), worms (Roy et al. 2002), and flies (Spellman and Rubin 2002; Boutanaev et al. 2002). For humans, similar results have been observed for genes expressed in the cardiovascular system (Dempsey et al. 2001) and highly expressed genes (Caron et al. 2001). For the later, a second study correcting for tandem gene duplications and excluding data from cancerous tissues suggested that the observed clusters were linked by breadth rather than magnitude of expression (Lercher et al. 2002). From the body of studies, there is sufficient motivation to develop computational approaches for the analysis of regulatory regions based on an expectation of correlated expression for genes in close proximity.

Ultimately, a large source of reliable data must be compiled for the development of mature models of chromatin and its dynamical role in the process of transcriptional regulation. At the sequence level there are promising results emerging for accessibility data derived from microarray-based chromatin immunoprecipitation studies in yeast (Ren et al. 2000; Iyer et al. 2001). Perhaps of equal importance in ultimately addressing the structural properties of chromatin, data about the spatial organization of chromatin within the nucleus are begin-

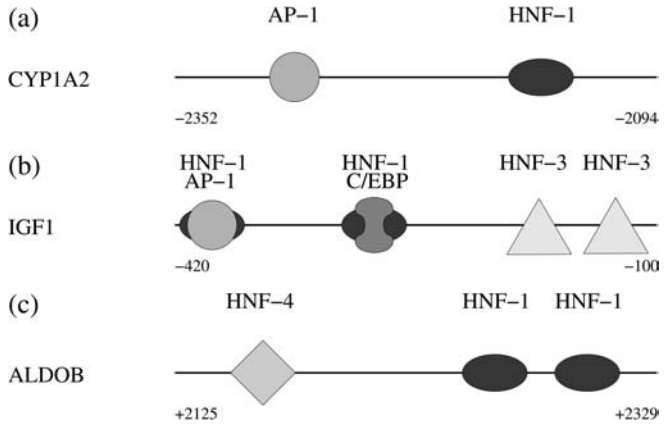

Fig. 4 Regulatory modules exist in distal enhancers, promoters, and introns. A positive enhancer element of the human CYP1A2 gene is shown in example (*a*). The element is located from –2,352 to –2,094 relative to the TSS and contains functional binding sites for the TFs AP-1 and HNF-1 (Chung and Bresnick 1997). Functional binding sites contained in the promoter of the rat insulin-like growth factor 1 (*IGF1*) gene are shown in example (*b*) (Zhu et al. 2000; Fournier et al. 2001). Note that the two sites on the upstream side of the region have been verified to bind more than one factor. Example (*c*) depicts the enhancer of the rat Aldolase B gene located in the first intron (Gregori et al. 1998)

ning to emerge (Skalnikova et al 2000; Boyle et al. 2001). Knowledge gathered from the implementation of these new approaches is vital for the elucidation of chromatin-mediated effects and will constitute an integral part of future computational approaches to the modeling of in vivo gene regulation. Recent advances in genomics suggest a brighter future and may provide opportunities for computational approaches to contribute to the scientific discovery process in the earliest stages.

In silico analysis of transcriptional regulation

By combining improved methods for the analysis of regulatory modules (see Fig. 4) and nuclear organization with enhanced resources for comparative genome sequence analysis, it is possible to advance our understanding of the transcriptional programs played out in the nucleus of metazoan cells. Computational biologists should seek opportunities to contribute to the formation of improved models of transcriptional regulation. To build a foundation for advances extending beyond the level of primary sequence analysis, computational biologists must identify future requirements for genome-scale experimental data. In deciphering this next generation of novel data, computational biologists will contribute to our understanding of the molecular mechanisms of transcriptional regulation.

Acknowledgements We are grateful for suggestions from David A. Adler, James L. Holloway, Charles E. Lawrence, Richard H. Price, Gary D. Stormo, Marissa Vignali, and the members of the Wasserman research group. We are indebted to Carol Sattler for providing the electron micrograph in the background of Fig. 1, which was created by Edward J. Andrews (SomeLabDesign, Seattle, WA, USA). We also thank David A. Adler for help with Fig. 1.

References

Anderson JD, Widom J (2000) Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol 296(4):979–987

Arnone MI, Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124(10):1851–1864

Bailey TL, Elkan C (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 21:51–83

Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 30(20):4442–4451

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci USA 99(2):757–762

Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet 3(12):342–347

Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. Science 281:61–63

Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 12(5):739–748

Blanchette M, Schwikowski B, Tompa M (2000) An exact algorithm to identify motifs in orthologous sequences from multiple species. Proc Int Conf Intell Syst Mol Biol 8:37–45

Bode J, Benham C, Knopp A, Mielke C (2000) Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). Crit Rev Eukaryot Gene Expr 10(1):73–90

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI (2002) Large clusters of co-expressed genes in the *Drosophila* genome. Nature 420(6916):666–669

Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet 10(3):211–219

Brazma A, Jonassen I, Vilo J, Ukkonen E (1998) Predicting gene regulatory elements in silico on a genomic scale. Genome Res 8(11):1202–1215

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212:563–579

Bulyk ML, Johnson PL, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 30(5):1255–1261

Bussemaker HJ, Li H, Siggia ED (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci USA 97(18):10096–10100

Caron H, Schaik B van, Mee M van der, Baas F, Riggins G, Sluis P van, Hermus MC, Asperen R van, Boon K, Voute PA, Heisterkamp S, Kampen A van, Versteeg R (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291(5507):1289–1292

Chung I, Bresnick E (1997) Identification of positive and negative regulatory elements of the human cytochrome P4501A2 (CYP1A2) gene. Arch Biochem Biophys 338(2):220–226

Claverie JM (2000) From bioinformatics to computational biology. Genome Res 10:1277–1279

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data

reveals chromosomal domains of gene expression. Nat Genet 26(2):183–186
Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2(4):292–301
Cremer T, Kreth G, Koester H, Fink RH, Heintzmann R, Cremer M, Solovei I, Zink D, Cremer C (2000) Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture. Crit Rev Eukaryot Gene Expr 10(2):179–212
Crowley EM, Roeder K, Bina M (1997) A statistical model for locating regulatory regions in genomic DNA. J Mol Biol 268(1):8–14
Davidson EH (2001) Genomic regulatory systems: development and evolution. Academic Press, San Diego
Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. Nat Genet 29(4):412–417
Dempsey AA, Pabalan N, Tang HC, Liew CC (2001) Organization of human cardiovascular-expressed genes on chromosomes 21 and 22. J Mol Cell Cardiol 33(3):587–591
Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12(3):458–461
Dundr M, Misteli T (2001) Functional architecture in the cell nucleus. Biochem J 356(2):297–310
Fickett JW (1996a) Quantitative discrimination of MEF2 sites. Mol Cell Biol 16:437–441
Fickett JW (1996b) Coordinate positioning of MEF2 and myogenin binding sites. Gene 172(1):GC19–32
Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. Genome Res 7(9):861–878
Fickett JW, Wasserman WW (2000) Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 11:19–24
Fournier B, Gutzwiller S, Dittmar T, Matthias G, Steenbergh P, Matthias P (2001) Estrogen receptor (ER)-α, but not ER-β, mediates regulation of the insulin-like growth factor I gene by antiestrogens. J Biol Chem 276(38):35444–35449
Frech K, Danescu-Mayer J, Werner T (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol 270(5):674–687
Frisch M, Frech K, Klingenhoff A, Cartharius K, Liebich I, W (2002) In silico prediction of scaffold/matrix attachment regions in large genomic sequences. Genome Res 12(2):349–354
Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17(10):878–889
Frith MC, Spouge JL, Hansen U, Weng Z (2002) Statistical significance of clusters of motifs represented by position-specific scoring matrices in nucleotide sequences. Nucleic Acids Res 30(14):3214–3224
Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196(2):261–282
Gregori C, Porteu A, Lopez S, Kahn A, Pichard AL (1998) Characterization of the aldolase B intronic enhancer. J Biol Chem 273(39):25237–25243
GuhaThakurta D, Stormo GD (2001) Identifying target sites for cooperatively binding factors. Bioinformatics 17(7):608–621
Halfon MS, Michelson AM (2002) Exploring genetic regulatory networks in metazoan development methods and models. Physiol Genomics 10(3):131–143
Halfon MS, Grad Y, Church GM, Michelson AM (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res 12(7):1019–1028
Hannenhalli S, Levy S (2001) Promoter prediction in the human genome. Bioinformatics 17 [suppl 1]:S90–S96
Hannenhalli S, Levy S (2002) Predicting transcription factor synergism. Nucleic Acids Res 30(19):4278–4284
Helden J van, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281(5):827–842
Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409(6819):533–538
Jackson DA (1997) Chromatin domains and nuclear compartments: establishing sites of gene expression within eukaryotic nuclei. Mol Biol Rep 24(3):209–220
Kadonaga JT (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. Cell 92(3):307–313
Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J Mol Biol 309(1):99–120
Klingenhoff A, Frech K, Quandt K, Werner T (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics 15(3):180–186
Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome: fundamental particle of the eukaryote chromosome. Cell 98(3):285–294
Kraus RJ, Murray EE, Wiley SR, Zink NM, Loritz K, Celembiuk GW, Mertz JE (1996) Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. Nucleic Acids Res 24:1531–1539
Krivan W, Wasserman WW (2001) A predictive model for regulatory sequences directing liver-specific transcription. Genome Res 11(9):1559–1666
Kruglyak S, Tang H (2000) Regulation of adjacent yeast genes. Trends Genet 16(3):109–111
Lawrence CE, Reilly A (1990) An EM algorithm for the identification and characterization of common sites in unaligned biopolymers sequence. Proteins 7:41–51
Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214
Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev 14(20):2551–2569
Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31(2):180–183
Levitsky VG, Podkolodnaya OA, Kolchanov NA, Podkolodny NL (2001) Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. Bioinformatics 17(11):998–1010
Levy S, Hannenhalli S (2002) Identification of transcription factor binding sites in the human genome sequence. Mamm Genome 13(9):510–514
Levy S, Hannenhalli S, Workman C (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 17(10):871–877
Liu R, States DJ (2002) Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. Genome Res 12(3):462–469
Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136–140
Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res 12(5): 832–839
Markstein M, Markstein P, Markstein V, Levine MS (2002) Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc Natl Acad Sci USA 99(2):763–768
McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE (2001) Phylogenetic footprinting of transcrip-

tion factor binding sites in proteobacterial genomes. Nucleic Acids Res 29(3):774–782
McCue LA, Thompson W, Carmack CS, Lawrence CE (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res 12(10):1523–1532
McGuire AM, Hughes JD, Church GM (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res 10(6):744–757
McNally JG, Muller WG, Walker D, Wolford R, Hager GL (2000) The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. Science 287(5456):1262–1265
Michelson AM (2002) Deciphering genetic regulatory codes: a challenge for functional genomics. Proc Natl Acad Sci USA 99:546–548
Ohler U, Niemann H (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet 17(2):56–60
Ohler U, Niemann H, Liao GC, Rubin GM (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics 17 [suppl 1]:S199–S206
Pearce D, Matsui W, Miner JN, Yamamoto KR (1998) Glucocorticoid receptor transcriptional activity determined by spacing of receptor and non-receptor DNA sites. J Biol Chem 273:30081–30085
Pedersen AG, Baldi P, Chauvin Y, Brunak S (1999) The biology of eukaryotic promoter predictiona review. Comput Chem 23(3–4):191–207
Polach KJ, Widom J (1995) Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. J Mol Biol 254(2):130–149
Polach KJ, Widom J (1996) A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J Mol Biol 258(5):800–812
Rajewsky N, Socci ND, Zapotocky M, Siggia ED (2002a) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res 12(2):298–308
Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002b) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. BMC Bioinformatics 3(1):30
Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science 290(5500): 2306–2309
Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16(10):939–945
Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol 20(8):831–835
Roy PJ, Stuart JM, Lund J, Kim SK (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature 418(6901):975–979
Scherf M, Klingenhoff A, Frech K, Quandt K, Schneider R, Grote K, Frisch M, Gailus-Durner V, Seidel A, Brack-Werner R, Werner T (2001) First pass annotation of promoters on human chromosome 22. Genome Res 11(3):333–340
Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18(20):6097–6100
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–56
Shim EY, Woodcock C, Zaret KS (1998) Nucleosome positioning by the winged helix transcription factor HNF3. Genes Dev 12:5–10
Simpson ER, Michael MD, Agarwal VR, Hinshelwood MM, Bulun SE, Zhao Y (1997) Cytochromes P450 11: expression of the CYP19 (aromatase) gene: an unusual case of alternative promoter usage. FASEB J 11(1):29–36
Skalnikova M, Kozubek S, Lukasova E, Bartova E, Jirsova P, Cafourkova A, Koutna I, Kozubek M (2000) Spatial arrangement of genes, centromeres and chromosomes in human blood cell nuclei and its changes during the cell cycle, differentiation and after irradiation. Chromosome Res 8(6):487–499
Smale ST (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim Biophys Acta 1351(1–2):73–88
Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol 1(1):5
Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16(1):16–23
Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23:109–113
Stormo GD, Tan K (2002) Mining genome databases to identify and understand new gene regulatory systems. Curr Opin Microbiol 5(2):149–153
Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD (2001) A comparative genomics approach to prediction of new members of regulons. Genome Res 11(4):566–584
Tautz D (2000) Evolution of transcriptional regulation. Curr Opin Genet Dev 10(5):575–579
Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouzé P, De Moor B, Marchal K (2002a) INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics 18(2):331–332
Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y (2002b) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9(2):447–464
Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J Mol Biol 266(2):231–245
Udalova IA, Mott R, Field D, Kwiatkowski D (2002) Quantitative prediction of NF-kappa B DNA-protein interactions. Proc Natl Acad Sci USA 99(12):8167–8172
Wagner A (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics 15(10):776–784
Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278(1):167–181
Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human–mouse genome comparisons to locate regulatory sites. Nat Genet 26:225–228
Werner T (1999) Models for prediction and recognition of eukaryotic promoters. Mamm Genome 10(2):168–175
Wolberger C (1999) Multiprotein–DNA complexes in transcriptional regulation. Annu Rev Biophys Biomol Struct 28:29–56
Wolffe AP, Guschin D (2000) Review: chromatin structural features and targets that regulate transcription. J Struct Biol 129(2–3):102–122
Workman CT, Stormo GD (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Proc Pac Symp Biocomputing 2000:467–478
Wu J, Grunstein M (2000) 25 years after the nucleosome model: chromatin modifications. Trends Biochem Sci 25(12):619–623
Xing EP, Wu W, Karp RM (2003) Capturing characteristic structural features for motif detection using a hierarchical Bayesian Markovian model. Genome Biol
Zhu JL, Kaytor EN, Pao CI, Meng XP, Phillips LS (2000) Involvement of Sp1 in the transcriptional regulation of the rat insulin-like growth factor-1 gene. Mol Cell Endocrinol 164:205–218