APPLIED BIOINFORMATICS FOR THE IDENTIFICATION OF REGULATORY ELEMENTS

Wyeth W. Wasserman[*‡] *and Albin Sandelin*[§]

The compilation of multiple metazoan genome sequences and the deluge of large-scale expression data have combined to motivate the maturation of bioinformatics methods for the analysis of sequences that regulate gene transcription. Historically, these bioinformatics methods have been plagued by poor predictive specificity, but new bioinformatics algorithms that accelerate the identification of regulatory regions are drawing disgruntled users back to their keyboards. However, these new approaches and software are not without problems. Here, we introduce the purpose and mechanisms of the leading algorithms, with a particular emphasis on metazoan sequence analysis. We identify key issues that users should take into consideration in interpreting the results and provide an online training example to help researchers who wish to test online tools before taking an independent foray into the bioinformatics of transcription regulation.

The creation of diverse cell types from an invariant set of genes is governed by biochemical processes that regulate gene activity. As the initial step of gene expression, transcription — one of the most widely studied processes in cell and molecular biology — is central to regulatory mechanisms. Transcription is shaped by the interactions between transcription factors (TFs) that bind *cis*-regulatory elements in DNA, additional co-factors and the influence of chromatin structure (FIG. 1). *Trans*-acting proteins that control the rate of transcription at the level of the individual gene bind crucial *cis*-regulatory sequences[1]. A full understanding of the interplay between *trans*-factors and *cis*-sequences would transform biological research, providing the means to interpret and model the responses of cells to diverse stimuli. Computational methods for the identification of *cis*-regulatory sequences that are associated with genes have long been sought owing to the arduous laboratory procedures required to identify them.

Deciphering the regulatory control mechanisms that govern gene expression might enable simplified interpretation of the complex data that now flood our computers. Ultimate success would produce a comprehensive map of the regulatory networks of each organism[2]. The reality, in all likelihood, is that the complex mixture of regulatory mechanisms that control the cellular concentrations of RNA will lead such efforts not to a single map, but rather to the creation of additional layers of large and complex data sets, the deciphering of which will require computational methods. The mastery of the entire network of gene regulation therefore remains a distant hope and aspiration. For the focused researcher, however, there are powerful and improving methods to identify regulatory sequences that control the rate of transcription initiation of specific genes of interest. For these researchers who strive to understand gene regulation in a targeted manner, bioinformatics methods can greatly accelerate their studies.

Although nearly all mature bioinformatics methods for the analysis of regulatory sequences address the initiation of transcription, other mechanisms that control gene expression should not be neglected. Regulation of any specific gene might occur at any point in the progression of transcripts into functional proteins (for example, splicing or protein modification)[1]. Characterizing the mechanisms that govern the initiation of transcription

**Centre for Molecular Medicine and Therapeutics and British Columbia Women's and Children's Hospitals, 3018–950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada.*
‡Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V5Z 4H4, Canada.
§Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, SE–171 77, Stockholm, Sweden.
Correspondence to W.W.W.
e-mail: Wyeth@cmmt.ubc.ca
doi:10.1038/nrg1315

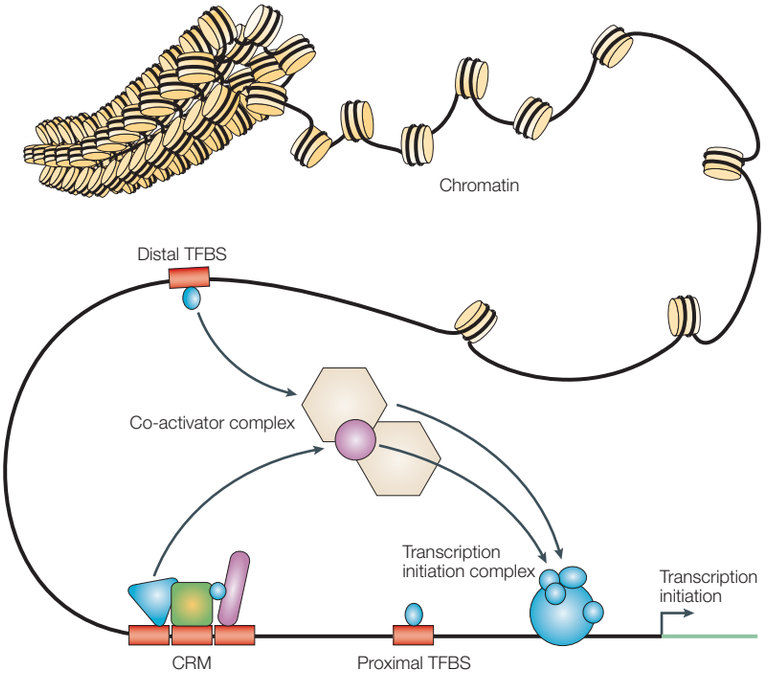


Figure 1 | **Components of transcriptional regulation.** Transcription factors (TFs) bind to specific sites (transcription-factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific binding TFs is highly dependent on the three-dimensional structure of chromatin.

does not reveal the entire picture. There is only partial correlation between transcript and protein concentrations[3]. Nevertheless, the selective transcription of genes by RNA polymerase-II under specific conditions is crucially important in the regulation of many, if not most, genes, and the bioinformatics methods that address the initiation of transcription are sufficiently mature to influence the design of laboratory investigations.

Below, we introduce the mature algorithms and online resources that are used to identify regions that regulate transcription. To this end, underlying methods are introduced to provide the foundation for understanding the correct use and limitations of each approach. We focus on the analysis of *cis*-regulatory sequences in metazoan genes, with an emphasis on methods that use models that describe transcription-factor binding specificity. Methods for the analysis of regulatory sequences in sets of co-regulated genes will be addressed elsewhere. We use a case study of the human skeletal muscle troponin gene *TNNC1* to demonstrate the specific execution of the described methods. A set of accompanying online exercises provides the means for researchers to independently explore some of the methods highlighted in this review (see online links box). Because the field is rapidly changing, emerging classes of software will be described in anticipation of the creation of accessible online analysis tools.

ORTHOLOGY
Two sequences are orthologous if they share a common ancestor and are separated by speciation.

PHYLOGENETIC FOOTPRINTING
An approach that seeks to identify conserved regulatory elements by comparing genomic sequences between related species.

MACHINE LEARNING
The ability of a program to learn from experience — that is, to modify its execution on the basis of newly acquired information. In bioinformatics, neural networks and Monte Carlo Markov Chains are well-known examples.

Identification of regions that control transcription

An initial step in the analysis of any gene is the identification of larger regions that might harbour regulatory control elements. Several advances have facilitated the prediction of such regions in the absence of knowledge about the specific characteristics of individual *cis*-regulatory elements. These tools broadly fall into two categories: promoter (transcription start site; TSS) and enhancer detection. The methods are influenced by sequence conservation between ORTHOLOGOUS genes (PHYLOGENETIC FOOTPRINTING), nucleotide composition and the assessment of available transcript data.

Functional regulatory regions that control transcription rates tend to be proximal to the initiation site(s) of transcription. Although there is some circularity in the data-collection process (regulatory sequences are sought near TSSs and are therefore found most often in these regions), the current set of laboratory-annotated regulatory sequences indicates that sequences near a TSS are more likely to contain functionally important regulatory controls than those that are more distal. However, specification of the position of a TSS can be difficult. This is further complicated by the growing number of genes that selectively use alternative start sites in certain contexts. Underlying most algorithms for promoter prediction is a reference collection known as the 'Eukaryotic Promoter Database' (EPD)[4]. Early bioinformatics algorithms that were used to pinpoint exact locations for TSSs were plagued by false predictions[5]. These TSS-detection tools were frequently based on the identification of TATA-box sequences, which are often located ~30 bp upstream of a TSS. The leading TATA-box prediction method[6], reflecting the promiscuous binding characteristics of the TATA-binding protein, predicts TATA-like sequences nearly every 250 bp in long genome sequences.

A new generation of algorithms has shifted the emphasis to the prediction of promoters — that is, regions that contain one or more TSS(s). Given that many genes have multiple start sites, this change in focus is biochemically justified.

The dominant characteristic of promoter sequences in the human genome is the abundance of CpG dinucleotides. Methylation plays a key role in the regulation of gene activity. Within regulatory sequences, CpGs remain unmethylated, whereas up to 80% of CpGs in other regions are methylated on a cytosine. Methylated cytosines are mutated to adenosines at a high rate, resulting in a 20% reduction of CpG frequency in sequences without a regulatory function as compared with the statistically predicted CpG concentration[7]. Computationally, the CG dinucleotide imbalance can be a powerful tool for finding regions in genes that are likely to contain promoters[8].

Numerous methods have been developed that directly or indirectly detect promoters on the basis of the CG dinucleotide imbalance. Although complex computational MACHINE-LEARNING algorithms have been directed towards the identification of promoters, simple methods that are strictly based on the frequency of CpG dinucleotides perform remarkably well at correctly predicting regions that are proximal to or that contain the

sites of transcription initiation[8]. Two leading methods — Eponine[9] and FirstEF[10] — use divergent approaches. FirstEF finds regions in genes with higher concentrations of CG dinucleotides than the local C and G concentrations would suggest. It subtly improves performance by restricting predictions to those regions that contain or are followed by a predicted 3′-splice site, thereby indicating the presence of a first exon. Eponine uses a NEURAL NETWORK model that analyses the over- and under-representation of longer oligonucleotide sequences. As Eponine's strand prediction is based on the identification of a TSS, which is an unreliable step, predictions of promoter orientation are not reliable. There is increasing evidence to indicate that promoters are bidirectional[11], signifying that the inability of bioinformatics methods to accurately predict promoter orientation is a by-product of biochemistry.

It is important to bear in mind that not all transcription-initiation sites are proximal to CpG islands and that the association between CpG dinucleotides and promoters is not present in all organisms. As only ~60% of human promoters are situated proximally to CpG islands[12], alternative approaches are required to identify a substantial portion of promoters. In our experience, the identification of promoter regions that lack CpG islands requires the use of transcript data. Recurrent alignment of the 5′ edges of ESTs and/or full-length cDNAs can be indicative of promoter locations. New mRNA cap-cloning techniques have overcome some of the technical limitations in generating full-length cDNAs[13]. The most direct means for users to access transcript data is through genome browsers[14]. Although human intuition can be remarkably adept at identifying sets of cDNAs that terminate at approximately the same position, there are emerging bioinformatics methods to quantitatively assess the significance of the observed transcript ends (REF. 15, and H. Sui and W.W.W., unpublished observations). The DBTSS database provides access to transcript-based TSS assignments for human and mouse genes[16].

A new source of data has the potential to place even greater emphasis on the interpretation of transcript data. Cap analysis of gene expression (CAGE) is a cap-cloning technique that has been extended with a SAGE-like procedure to cleave the initial 5′ 20 nucleotides of full-length cDNAs[17]. These oligomers are subsequently ligated into long polymers and sequenced. Generation of these CAGE tags from transcripts that are derived from diverse tissues promises not only to facilitate improved promoter prediction, but also to provide insights into tissue-specificity.

NEURAL NETWORK
A machine-learning technique that simulates a network of communicating nerve cells.

CAGE
(Cap analysis of gene expression). The high-throughput sequencing of concatamers of DNA tags that are derived from the initial nucleotides of 5′ mRNA.

SAGE
(Serial analysis of gene expression). A method for quantitative and simultaneous analysis of a large number of transcripts; short sequence tags are isolated, concentrated and cloned; their sequencing reveals a gene-expression pattern that is characteristic of the tissue or cell type from which the tags were isolated.

Phylogenetic footprinting

Sequence similarity that results from selective pressure during evolution is the foundation for many bioinformatics methods[18,19]. For the prediction of transcription-factor binding sites (TFBSs), sequence similarity is primarily manifested in the process known as phylogenetic footprinting (reviewed in REF. 18). Under the assumption that mutations within functional regions of genes will accumulate more slowly than mutations in regions without sequence-specific function, the comparison of sequences from orthologous genes can indicate segments that might direct transcription. The completion of several eukaryotic genome sequences[20–24] has motivated the creation of a new set of alignment, analysis and visualization methods to discern conserved segments. The initial studies emphasized pairwise comparisons of sequences that are separated by 50–70 million years of evolution (for example, human–rodent)[25,26]. In its current form, phylogenetic footprinting can reveal genomic regions that are likely to regulate gene expression with a limited chance of bypassing functionally important sequences. In the most successful cases, phylogenetic footprinting can pinpoint important regulatory regions with sufficient clarity to motivate targeted validation experiments.

A key assumption in the application of phylogenetic footprinting is the implicit hypothesis that the regulation of orthologous genes will be subject to the same regulatory mechanisms in different species. Although generally correct over moderate evolutionary distances, an investigator should consider whether there is evidence that supports or contradicts this implicit assumption. Alignment-based phylogenetic footprinting methods are relevant for orthologous genes from species with appropriate evolutionary divergence. Pairwise alignment comparison of promoters from closely related species, such as human–chimpanzee, generally provide little benefit, as the sequences closely resemble each other, whereas promoters from widely divergent species (primate–fish) can show no detectable similarity[26]. The rate of evolutionary events in promoters is different for genes within the same organism; so, in some cases, it is most productive to compare sequence pairs from more diverged species. For instance, genes that are important in early embryonic development can require comparisons as extreme as 450–500 million years apart (that is, primate–fish) to reveal regulatory regions[27,28]. The selective pressure that results in the high retention of sequences in well-studied cases — exemplified by Hox clusters — has been linked to chromatin structure or unknown mechanisms that allow coordinated regulation of clusters of genes[29].

There are three components to the existing phylogenetic footprinting algorithms: defining suitable orthologous gene sequences for comparison, aligning the promoter sequences of orthologous genes and visualizing or identifying segments of significant conservation.

Although retained function is not inherent to the definition of orthology, for the purpose of phylogenetic footprinting, the assumption is made that orthologous genes are under common evolutionary pressures. Defining orthologues is complicated by the duplication and/or deletion of genes during evolution — it is sometimes difficult to reliably select suitable sets of sequences for study. Bioinformatics resources that provide broadly related orthologues between species include COGs/KOGs[30], HOPs[31] and HomoloGene[32].

Once suitable sequences are obtained, they must be aligned to identify segments of similarity. There are two broadly used algorithms for such alignments: one that targets short segments of similarity and the other an optimal description of similarity across an entire pair of

Table 1 | **Selected web-based resources for gene regulation bioinformatics***

Resource name	URL	Reference
Promoter predictions		
Eponine	http://www.sanger.ac.uk/Software/analysis/eponine	9
FirstEF	http://rulai.cshl.edu/tools/FirstEF	10
DBTSS	http://dbtss.hgc.jp/index.html	16
Transcription-factor binding profile databases		
TRANSFAC®	http://www.gene-regulation.com/pub/databases.html#transfac	61
JASPAR	http://jaspar.cgb.ki.se	59
Transcription-factor binding site predictions		
Match™	http://www.gene-regulation.com/pub/programs.html#match	95
ConSite	http://phylofoot.org/consite	26
rVista	http://rvista.dcode.org	37
Transcription-factor module predictors		
MSCAN	http://tfscan.cgb.ki.se/cgi-bin/MSCAN	72
Cluster Buster	http://zlab.bu.edu/cluster-buster/cbust.html	68
CRÈME	http://creme.dcode.org/	85
Alignment of non-coding genome sequences		
PipMaker	http://bio.cse.psu.edu	38
LAGAN	http://lagan.stanford.edu	34
AVID	http://baboon.math.berkeley.edu/mavid	60
Data visualization		
Sockeye	http://www.bcgsc.ca/gc/bomge/sockeye	41
rVista	http://rvista.dcode.org	37
Genome browsers		
UCSC Genome Browser	http://genome.ucsc.edu	14
Ensembl	http://www.ensembl.org	96
Orthology resources		
COGs/KOGs	http://www.ncbi.nlm.nih.gov/COG	30
EGO (formerly TOGA)	http://www.tigr.org/tdb/tgi/ego	97
Orthostrapper	http://orthostrapper.cgb.ki.se	98
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene	32

*The list is not exhaustive.

LOCAL ALIGNMENT
The detection of local similarities between two sequences.

GLOBAL ALIGNMENT
The alignment of two sequences over their full length.

NEEDLEMAN–WUNSCH ALGORITHM
A commonly used algorithm in bioinformatics that produces a global alignment of two sequences. The term 'global' refers to alignments across the entirety of the sequences. The algorithm returns an optimal alignment, in which 'optimal' refers to the highest possible score under a specific scoring system. The algorithm is computationally demanding, restricting its direct application to sequences of modest length.

sequences. For the former, the BLASTZ[33] algorithm identifies short segments of exact identity and constructs LOCAL ALIGNMENTS by extending the analysis from the edges of each seed. A large set of these local alignments can be displayed in a format known as PIPs (percent identity plots), which more accurately delineate the edges of similar subsegments than window-based conservation plots.

The alternative method generates a single, near-optimal alignment across the entire length of the orthologous gene sequences. In the case of LAGAN[34], a widely used algorithm of this type, short local alignments are generated (similar to the seeds produced by BLASTZ) to establish related sub-segments. Subsequently, a GLOBAL ALIGNMENT is produced using the NEEDLEMAN–WUNSCH ALGORITHM[35]. The choice to use global alignments introduces the assumption that important functional sequences will remain collinear over evolution (in the same order and orientation along the gene). A recent extension of the LAGAN algorithm circumvents this particular problem by identifying blocks of sequence that are likely to have undergone inversions (shuffle-LAGAN[36]). Several similar algorithms that, in our opinion, perform comparably well are listed in TABLE 1.

The global alignment tools generally have difficulty with duplicated segments, producing results that indicate that one of the copies of a duplicated sequence is not conserved. The BLASTZ local alignment method circumvents such problems, but the failure to consider collinearity in functional elements might result in a decreased ability to identify subtle similarities in weakly conserved segments between well-conserved blocks (although this has not been conclusively demonstrated).

Once an alignment or set of alignments is defined, various tools are available to assist in the interpretation of the data. The VISTA browser[37] presents a graph of nucleotide identity within a sliding window along a pairwise alignment. Similarly, PipMaker[38] displays BLASTZ results in an intuitive presentation. Although graphical display is useful, computational analysis of observed conservation patterns is essential for the analysis of long sequences.

A new method analyses the patterns of nucleotide identity in subregions of the alignment and classifies conserved regions as coding or regulatory[39]. This 'regulatory potential' algorithm is based on the pattern of observed identical nucleotides. For instance, coding regions tend to vary at the third codon position and have insertion/deletion (indels) lengths that are multiples of three. Alternatively, regulatory sequences tend to have more frequent indels and variations occur in distinct blocks that are separated by segments of high similarity. The method, implemented as a HIDDEN MARKOV MODEL, is not broadly available, but represents a class of analysis that is likely to become increasingly important as more genome sequences become available.

With the emergence of diverse genome sequences[40], some of the limitations of pairwise analysis methods have become apparent. Multiple sequence alignment methods, enhanced visualization tools and a new class of statistical analysis methods will be required to identify and interpret patterns that are restricted to a branch of a species tree. The mLAGAN[34] alignment algorithm seems to be well suited to the alignment challenge. Once an alignment is created, however, the analysis of multiple sequences is problematic. The basis for determining the significance of local similarity within a branch of a species tree remains to be established for large-scale analyses. For visualization, the new Sockeye package[41] creates dynamic, three-dimensional graphics that allows users to create a virtual phylogenetic footprinting landscape. For the impatient scientist awaiting appropriate tools for multiple-sequence phylogenetic footprinting, it seems that mLAGAN alignments visualized in Sockeye represent the near-term solution.

Modelling sequence-specific binding

TFs generally have distinct preferences towards specific target sequences. Given a set of known binding sites, it is possible to construct a model to describe the target sequence properties that can be used to predict potential binding sites in genomic sequences. The problem is twofold: it is necessary to select an appropriate way to model binding preferences on the basis of experimental data and to develop methods to apply the models to find functional TFBSs in promoter sequences.

Several assumptions underlie the most prevalent methods for TFBS prediction. The one that is most likely to be violated is that each TF binds independently to its target. In specific terms, we assume that binding is not influenced by the content of adjoining sequences and the proximity of other proteins. This is fundamentally incorrect, as combinatorial interactions of multiple factors that bind to multiple sites are essential for the specific regulation of gene transcription[2]. Such combinatorial requirements have been demonstrated for genes ranging from the *endo16* regulatory network in sea urchin[42] to the β-globin cluster in human[43]. The above assumption results in a severe limitation — the inability to specifically distinguish between sites that have a functional role *in vivo* and sites that exert no function. Owing to an extremely high rate of false-positive predictions of TFBSs[44], specificity is usually measured in terms of the rate of predictions. This rate varies for each TF binding model and is influenced by model parameters, but the application of most models with standard settings will report TFBSs in the range of 1/500–1/5000 bp. Take, for instance, a model for the binding of myoD, a muscle-specific TF, that predicts one binding site in approximately every 500 bp[45]. Applying this model to the human genome produces ~10^6 predictions of binding sites, of which ~10^3 are likely to be functional. The high number of false predictions is not however, as Tronche[46] demonstrated, simply a result of inadequate model frameworks — predicted sites are bound readily by TFs *in vitro*. In fact the methods do detect potential binding sites, albeit not necessarily those of functional importance. By most accounts, the three orders of magnitude difference between true and false predictions is intolerable, resulting in what we choose to term the FUTILITY THEOREM — that essentially all predicted TFBSs will have no functional role. Fortunately, there are biologically motivated approaches to overcome this 1000-fold excess of false predictions.

To understand both the strengths and weaknesses of the current methods to model TF binding, it is necessary to understand not only the theory behind the models, but also the scoring methodology and the limitations that are imposed by the available experimental data. The construction of models for predicting binding sites for TFs is limited by the limited abundance of valid *cis*-regulatory elements. Such target sites are generally defined by arduous laboratory analysis of promoters that involves deletion mapping and, eventually, mutagenesis of regulatory sequences. Owing to the tolerance of TFs for significant variation between target sequences, multiple sites are required to construct a model. As an example, we can consider the myocyte enhancer factor 2 (MEF2) for which two of the known binding sites are conserved at only 7 out of 14 positions (BOX 1).

There are two distinct approaches for generating binding-site collections for a specific TF, each with its own caveats. Functional regulatory elements that are defined from genes are sparse, but, for a subset of TFs, a sufficient number have accumulated to indicate the diversity of possible binding sites. Alternatively, high-throughput selection procedures can be performed, in which pools of random DNA sequences are mixed with a TF and those that are preferentially bound are recovered and sequenced[47], or fluorescently labelled proteins are directly bound to arrays of potential binding sites[48]. Based on a comparison with binding sites that were defined in functional *in vivo* assays, sites for a prokaryotic TF detected with *in vitro* SELEX assays were not fully representative[49]. Despite the potential for a partial binding profile, a new generation of high-throughput methods is generating collections of thousands of sites that will facilitate the creation of useful binding models[50].

Consensus sequences can be used to represent the properties of known binding sites. The binding sites for a factor are aligned together and a consensus nucleotide letter is assigned to represent the nucleotide composition in each column. Although the use of consensus sequences provides better representation than a single

HIDDEN MARKOV MODEL (HMM). A probabilistic model for the recognition of patterns in DNA or protein sequences. HMMs represent a system as a set of discrete states and as transitions between those states. Each transition has an associated probability, which can be readily derived from training sets, such as alignments of known examples of a pattern. HMMs are valuable because they enable a search or alignment algorithm to be built on firm probabilistic bases.

FUTILITY THEOREM
The authors' assertion that essentially all predicted transcription-factor (TF) binding sites that are generated with models for the binding of individual TFs will have no functional role.

SELEX
(Systematic evolution of ligands by exponential amplification). A set of laboratory procedures for the identification of representative sets of ligands for a protein. In the case of DNA-binding proteins, the protein is mixed with a pool of double-stranded oligonucleotides that contain a random core of nucleotides flanked by specific sequences. The protein in complex with bound DNA is recovered and the ligands are subsequently amplified by PCR. The recovered oligonucleotides are sequenced and analysed to reveal the binding specificity of the protein.

Box 1 | Building models for predicting transcription-factor binding sites

The first step towards building models for predicting transcription-factor (TF) binding sites involves data collection. To illustrate the process, we use MEF2 as an example.

Data collection

A set of experimentally validated MEF2-binding sites was collected from the literature and aligned (**a**). The sequence variability of the collection of binding sites strongly affects the downstream models for predicting additional sites. Note the diversity between the sites; for instance, only 50% of the nucleotides are identical between sites one and eight.

Model building

Consensus sequence model: a consensus sequence is defined by selecting a degeneracy nucleotide symbol for each position (column) in the alignment (**b**). Unusual binding sites can have an extreme effect on the consensus (see, for example, site eight).

Position frequency matrix

To more accurately reflect the characteristics at each position, a matrix that contains the number of observed nucleotides at each position is created (**c**). For instance, the first column in the alignment (**a**) consists of no As, three Cs, two Gs and three Ts, therefore resulting in the corresponding first matrix column {0,3,2,3}.

Position weight matrix

The frequency matrix is usually converted to a position weight matrix (PWM) using a formula (BOX 2, equation 2) that converts normalized frequency values to a log-scale (**d**). PWMs are also known as position-specific scoring matrices (PSSMs, pronounced 'possums'). Using a matrix model, a quantitative score for any DNA sequence can be generated by summing the values that correspond to the observed nucleotide at each position (**e**). For large and representative collections of binding sites, the scores are proportional to binding energies[51].

Sequence logo

The specificity in each column of the alignment can be measured in terms of INFORMATION CONTENT[92]. A sequence logo scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position (**f**; BOX 2, equation 4). Sequence logos enable fast and intuitive visual assessment of pattern characteristics.

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

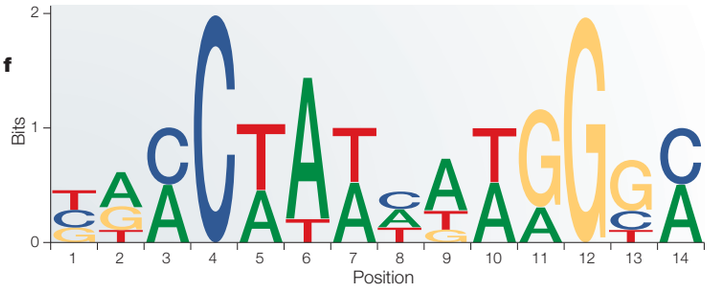
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

e Site scoring

1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

f



sequence and lends itself to fast visual comparisons, it fails to reflect the quantitative characteristics of TF binding. Consensus sequences confer an information loss from the original data, as binding bias towards one of the possible nucleotides is not reflected in the model (BOX 1).

Position weight matrix (PWM) profiles provide quantitative descriptions of the known binding sites for a TF[51]. Based on an alignment of all known sites, the total number of observations of each nucleotide is recorded for each position, producing a position frequency matrix (PFM; see BOX 1). A normalized PFM, in which each column adds up to a total of one, is a table of probabilities for observing each nucleotide at each position.

The matrix framework enables us to assign a quantitative score to any sequence to identify potential binding sites. It is helpful to visualize a profile model as a 'machine' that analyses a string of nucleotides (of the same length as the profile). The calculation of the probability of observing a certain sequence is straightforward: it is simply the product of the relevant nucleotide probabilities in each position in the profile.

For efficient computational analysis, the PFM must be converted to a log-scale. To eliminate null values before log-conversion, and in part to correct for small samples of binding sites, a sampling correction, known as PSEUDOCOUNTS, is added to each cell of the PFM (BOX 2). The specific formula for the pseudocount correction varies widely between software applications[52]. In our formulation, pseudocount values are defined as the square root of the number of sites that contribute to the model. Additionally, the genome nucleotide distribution is taken into account in the conversion (BOX 2). The final log-scale matrix is referred to as a PWM. A quantitative score for a potential site is produced by summing the relevant nucleotide PWM values, analogous to the calculation of the probability of observing the site, as discussed above (BOX 1 and 2). For longer sequences, the PWM is slid over the sequence in 1-bp increments, evaluating each possible binding site (on both strands).

INFORMATION CONTENT
A measure of nucleotide conservation in a position, based on information theory.

PSEUDOCOUNT
The sample correction that is added when assessing the probability to correct for small sample sizes (that is, few binding sites).

The PWM scores are directly related to the binding energy of the DNA–protein interaction[51,53]. So, the PWM representation can be viewed both as a statistical and as an energy-based model.

There are two additional assumptions to consider. Current matrix models for binding-site prediction are based on the assumption that a nucleotide at one position has no effect on the likelihood of a nucleotide being observed at an adjoining position. For a few cases in which large data collections have been generated to richly define binding, advanced models that incorporate higher-order interactions between positions have proved more effective[53–55]. However, the improved specificity of the models has been modest, indicating that the simpler, position-independent matrix models are adequate in most cases[56]. The second assumption is that TFs have strict spatial requirements in their binding sites that preclude variable spacing. For some TFs, such as a subset of the nuclear receptor family[57], variable spacing is allowed, rendering standard PWMs inappropriate for TFBS prediction. Specialized models, such as one for the transcription factor CTF[58], have been created to model binding for some of these cases.

Prediction of functional binding sites

Internet-based software tools have been implemented to screen DNA sequences with databases of matrix models. Although the TRANSFAC database and associated search tools are broadly used, the futility theorem holds that the resulting site predictions will not be functional *in vivo* despite a strong likelihood that the TF would bind to the sequence *in vitro*. This discrepancy between the *in vivo* and *in vitro* predictive accuracy indicates that additional properties must specify the function of regulatory sequences.

Two complementary observations of the characteristics of regulatory sequences have motivated substantial improvements in the prediction of functional binding sites. First, the previously indicated observation of sequence conservation in regulatory regions can be extended to enhance the predictive specificity with matrix models. Second, gene regulation that is mediated by cooperative interactions between TFs that bind to clusters of sites within *cis*-regulatory modules (CRMs) can be captured in computational algorithms to improve performance.

User-orientated tools have emerged that combine matrix-based site predictions with phylogenetic footprinting. In general, these tools require pairs of orthologous gene sequences. As mentioned above, programs such as LAGAN can generate global progressive alignments [34]. Fixed-length windows that exceed a defined sequence-identity threshold in the alignment are classified as conserved. A database of PWM binding profiles, such as JASPAR[59], is used to predict binding sites within the conserved regions, with the most stringent methods restricting reported predictions to TFBSs that are present at corresponding positions in the alignment of the orthologous sequences. The results are usually represented graphically as conservation plots (BOX 3).

Box 2 | **Formulae linked to methods for the analysis of regulatory sequences**

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation:
$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion:
$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

$p(b)$ = background probability of base b; $p(b,i)$ = corrected probability of base b in position i; $W_{b,i}$ = PWM vaue of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3).

Evaluation of sequences:
$$S = \sum_{i=1}^{w} W_{l_i,i} \quad (3)$$

l_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

Information content calculation:
$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \quad (4)$$

D_i = information content in position i; $p(b,i)$ = corrected probability of base b in position i

The performances of the available algorithms that couple TFBS prediction with phylogenetic footprinting have been similar. The rVista[37] service, which uses the AVID[60] alignment program, the TRANSFAC[61] database and the VISTA visualization package[37], was assessed on a collection of 21 functional binding sites for complexes of AP1 and NF-AT from genes in the cytokine gene cluster. The ConSite[26] service, which uses the ORCA alignment program, the JASPAR database of binding profiles and a web interface powered by the TFBS perl modules[59,62] for the Perl programming language, was assessed on a reference collection of more than 100 functional binding sites for a wide range of TFs from genes distributed across the human genome. Although performance depends on settings, both systems eliminate ~90% of predictions while retaining ~70–80% of experimentally validated sites. Therefore, the combination of phylogenetic footprinting and PWM searches applied to orthologous human and mouse gene sequences reduces the rate of false predictions by an order of magnitude with modest reduction in sensitivity. Two recent reports have indicated conservation of only ~50% of human/mouse regulatory sites[63,64]. Certain TFs, such as Sp1 and C/EBP, bind to target sequences that vary widely. Evolutionary pressure to retain such binding sites is minimal owing to the high likelihood that alternative sites will be available within a regulatory region. This indicates that there might be two subtypes of TFBS: highly selected sites that rarely occur by chance and auxiliary sites that are available by convenience. This hypothesis implies that phylogenetic footprinting methods will be well suited for binding sites for TFs with greater binding specificity.

The proliferation of complete genome sequences has created opportunities for a variation on phylogenetic footprinting in which multiple sequences from closely related species are used. The comparison of multiple primate sequences was used originally to identify short, highly conserved binding sites in globin genes[65]. In this variant on footprinting, subsequently called 'phylogenetic shadowing'[66], multiple sequence alignments are

Box 3 | **Coupling binding-site prediction with phylogenetic footprinting**

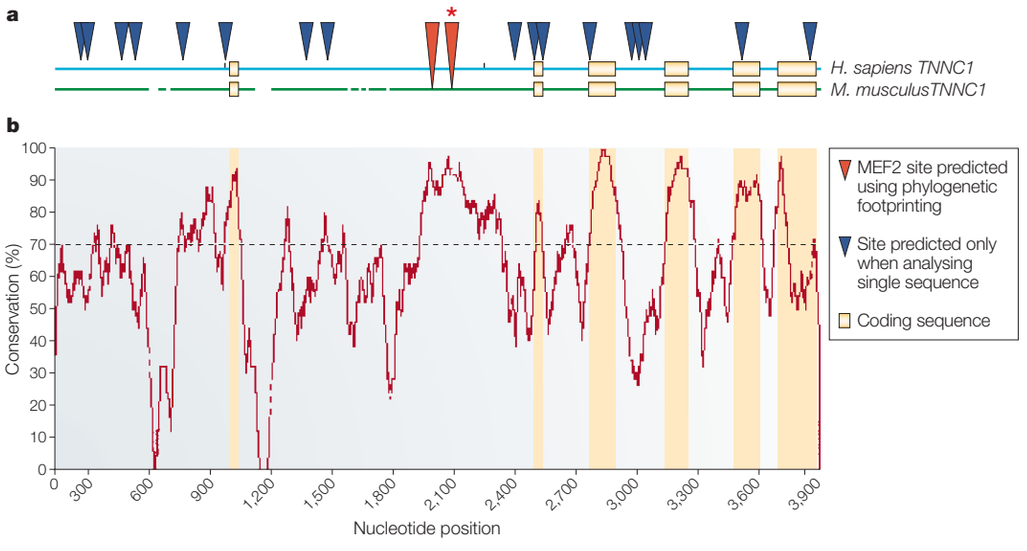
To illustrate the power of cross-species comparison to eliminate spurious predictions of binding sites, we analysed the promoter of the human *TNNC1* gene using the MEF2 position weight matrix (PWM).

Method

We did the analysis with and without restricting predicted sites to segments of high sequence conservation with the mouse orthologue. We used the ConSite system[24] for the analysis, which aligns the input gene sequences and analyses them with user-specified PWMs (settings: window size = 50 bp, conservation cutoff 70%, relative score threshold 72%). The system returns an illustration of the gapped alignment (a) and a conservation plot (b). The exon locations that are predicted from the user-specified cDNA sequence are marked in yellow.

Interpretation

Predicted binding sites for MEF2 are displayed in blue (no conservation constraint) and red (conserved): some subregions, despite being non-coding, are conserved; the largest of them is located in the first intron. Of the two predicted sites in this region, one is experimentally verified to be functional *in vivo* (marked with *)[93,94]. Consistent with the futility theorem, most predicted binding sites that were generated in the analysis of the single human sequence (blue) are spurious. Cross-species comparisons can substantially improve the specificity of predictions, eliminating up to 90% of false predictions[26]. Careful consideration should be given to whether the regulation of the gene is likely to be similar in the species analysed.



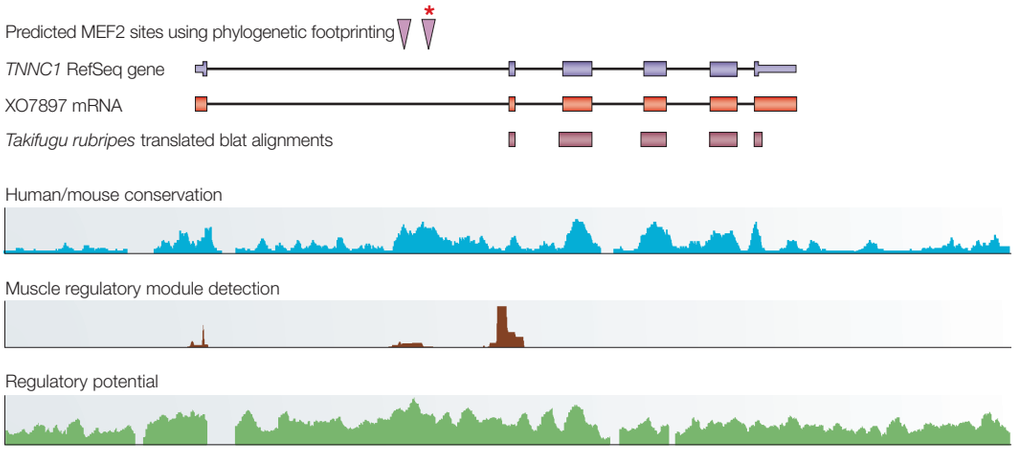
Box 4 | **Analysis of *cis*-regulatory modules**

Combinatorial interactions generate functional specificity

Biochemical specificity of transcription is generated in the cell nucleus by combinatorial interactions between transcription factors[2]. Bioinformatics approaches that are based on the detection of such combinations of binding sites (termed *cis*-regulatory modules; CRMs) can produce predictions of substantially better specificity than analysis of isolated sites. A muscle-specific CRM predictor[67], which uses matrix models for SRF, MEF2, TEF1 and MyoD/Myf, was used to analyse the muscle-specific TNNC1 promoter on chromosome 3p21. Results are shown superimposed with human genome browser annotations for exons and sequence conservation with mouse. The gene is predicted to contain three potential muscle CRMs (three brown peaks in muscle regulator region detection graph). The experimentally confirmed module (marked with *), containing the previously described MEF2-binding site (BOX 3), corresponds to a weaker module prediction. The region predicted to have the highest muscle CRM potential is not situated within a conserved region and its function has not been directly analysed.

Combining predictions from independent methods

Supplemental data can and should be considered when evaluating predictions. The conservation track between human and mouse complements the phylogenetic footprinting results to predict the functionally confirmed CRM. The regulatory potential track[39], discussed in the main text, provides additional support in favour of the known module. Phylogenetic footprinting, module-based detection and regulatory potential estimation all agree on the regulatory importance of the first intron. A quantitative method for weighing the influence of these measures remains to be established.



analysed to identify short invariant blocks of sequence. The method is useful for analysing genes, such as apolipoprotein(a)[66], which emerged or obtained new functions during the evolution of primates. However, it is unclear how much impact it will have on improving the performance of binding-site predictions. Assessment of the statistical significance of observed shadows indicates that the method has broad utility (J. McAuliffe, personal communication), but a comprehensive analysis remains to be produced.

Returning to the three orders of magnitude deficiency in specificity that is expressed in the futility theorem, an order of magnitude reduction in false-positive predictions, although appreciated, is insufficient to circumvent the rate of false predictions made by PWMs. Phylogenetic footprinting methods are, ultimately, a 'crutch' used in bioinformatics that reflects our underlying naive understanding of the biochemical mechanisms of gene regulation. Within a cell, TFs do not check an evolutionary index to determine whether a site is suitable for binding. So, increasingly, there is focus on the creation of bioinformatics algorithms that more directly reflect the biochemical mechanisms that regulate gene transcription.

Efforts to incorporate biochemical knowledge into regulatory sequence-prediction algorithms have focused on the identification of regions in genes with statistically significant combinations of binding sites for biologically-linked sets of TFs (BOX 4). These methods for detection of CRMs have evolved rapidly within a few years and fall into two classes: trained and untrained methods. Trained approaches use machine-learning techniques to identify characteristics of known regulatory modules that can be used to accurately detect sequences with similar properties. Untrained methods are based on the statistical likelihood of detecting observed combinations of predicted TFBSs within a specified segment of a sequence.

For a few richly studied cell types, there is a relatively abundant set of experimentally defined regulatory sequences, which is sufficient to direct expression of a reporter gene in a cell-specific pattern. One large subset of methods for human CRM prediction is based on a curated collection of regulatory regions that direct gene expression selectively to skeletal muscle (most often to C2C12 cells in culture that have differentiated into myotubes). The original analysis of the muscle CRMs used logistic regression analysis with a vector of five

PWM-generated scores that were obtained with profiles for the five TFs associated with skeletal muscle expression[67]. Compared with the rate of predictions of individual TFBSs, the focus on CRMs eliminated ~99% of false TFBS predictions while retaining 60% of functional regions. The initial version of this method allowed flexible spacing between sites and weighted predictions towards key classes of binding sites, but did not allow for multiple binding sites for the same TF to contribute to the predictions. A subsequent algorithm uses hidden Markov models to circumvent this limitation, improving specificity a further twofold[68]. Similar studies with sets of genes expressed selectively in hepatocytes[69] or in response to inflammation[70] demonstrated the broad applicability of the trained models in diverse biological contexts. In the analysis of hepatocyte regulation, the CRM analysis was coupled to phylogenetic footprinting, eliminating 99.9% of predictions while retaining ~50% of known CRMs. In the best cases, such integration can overcome the constraints that are expressed by the futility theorem.

Trained methods place emphasis on predicted sites for key TFs, whereas untrained methods allow the identification of significant combinations of sites in the absence of extensive reference collections of functional modules. Current data constraints limit most users to the untrained methods that focus on the significance of observed concentrations of sites for one or more TFs. In these cases, biological knowledge that highlights such a set of TFs as being potentially involved in regulating transcription in a specific context is generally available. For instance, in certain genes that are expressed during pattern formation in the fly embryo, large clusters of binding sites for homeobox and zinc-finger TFs were qualitatively detected[71]. Efforts to establish methods for statistically assessing the significance of the combination of sites have proliferated (for example, MSCAN[72], MCAST[73] and ModuleScanner[74]). Such assessment is challenging because the non-random properties of chromosomal DNA can lead to the identification of erroneous regions. Most methods attempt to model the regional properties of sequences and assign significance to observed combinations of sites on the basis of the local characteristics of nucleotide composition. Such methods are prone to identifying short local segmental duplications[72], in which identical TFBS predictions are conjugated into statistically significant, but biologically meaningless, chains. In the case of MSCAN, the untrained model identifies false CRMs at a rate approximately fourfold higher than the best trained methods. Recent advances with models for HOMOTYPIC CLUSTERS of binding sites in fly genes have modelled positional interactions between sites to achieve specificity rates of up to 50% correct predictions[75,76].

To detect CRMs, sufficient data must be available to accurately model the binding specificity of each contributing TF. In fact, such data are sparse and the availability and quality of PWMs sharply restricts the application of CRM analysis. For cases in which a PWM for a crucial TF is not available, it is often suitable to substitute a binding profile for a TF from the same structural class. Applicable to such cases, a set of general binding profiles has recently been developed for structural classes that show consistent binding specificity[77].

The analysis of CRMs can generate predictions of sufficient specificity to motivate detailed laboratory studies. However, the limited knowledge of binding sites for many TFs and reference collections of known CRMs precludes the wide application of cluster analysis.

Emerging methods

The population of computational biologists who explore regulatory sequence analysis is growing exponentially. Although much work remains to be done to optimize proven methods such as phylogenetic footprinting, important new directions are being actively explored. Some of these promising new methods are likely to influence the field in the near future.

The availability of genome sequences from more diverse species will influence all aspects of regulatory analysis. Taking the lead from research with bacterial[78] and yeast[79] genomes, a new class of phylogenetic footprinting, which we term 'regulog' analysis, will become widely used in understanding the regulation of human genes. This procedure identifies predicted binding sites that are statistically overrepresented in sets of promoters from orthologous genes from widely diverged species. In such cases, although regulatory mechanisms have been retained, regulatory sequences have diverged to the extent that global alignments cannot be generated. These regulatory regions are screened with collections of PWMs to identify classes of overrepresented binding sites. Such methods are starting to emerge for sets of co-expressed human genes[80,81], but will eventually be applied to sets of distantly related orthologues.

Similarly, the pool of diverse genomes will increase the challenge of aligning known regulatory modules. Given a regulatory region in a human gene, for instance, one would like to determine whether a similar region is present in a distant orthologue. Given the lack of similarity at the nucleotide level, new methods will be required to align predicted binding sites. An early method in this direction used strict spacing rules and site requirements to detect modules[82]. A new method is more flexible[83] because it aligns motif matches instead of individual nucleotides. Substantial effort will be required to develop a method that incorporates confidence weighting that emphasizes the functionally confirmed TFBSs.

CRM models are improving rapidly. A new generation of BAYESIAN CRM models will sharply improve predictive performance by incorporating interactions between pairs of factors (Thompson *et al.*, in preparation). Early bioinformatics efforts indicated that certain pairings of TFBS types could be identified, which sharply improve predictions[45]. Recent efforts have returned to this theme to demonstrate the possibility of identifying significant correlations between site types[84,85]. Full CRM models that incorporate these relationships and couple them to phylogenetic footprinting might enable the accurate computational prediction of human regulatory networks.

HOMOTYPIC CLUSTER
A cluster of similar transcription-factor (TF) binding sites, often binding the same TF.

BAYESIAN [METHOD]
A statistical method of combining the likelihood with additional information to produce an overall estimate of the strength of a piece of evidence.

There is increasing evidence that the relationship between TF structures and binding specificity can be resolved. In an important demonstration of the idea, models were generated for the binding specificity of zinc-finger TFs on the basis of the amino acids in the protein–DNA interface. Such models reliably predict binding specificity, indicating that the gap between linear DNA sequence analysis and protein structural analysis can be traversed[55].

The analysis of regulatory sequences has been significantly improved through the analysis of sequence evolution (phylogenetic footprinting) and combinatorial interactions between TFs (CRM analysis). It is likely that the next breakthrough will depend on interpreting the unaddressed regulatory system in the cell nucleus — the chromatin structure[86,87]. Despite the fact that some early bioinformatics pioneers have attempted to construct algorithms related to chromatin effects[88], progress has been extremely slow. Although data remain sparse, there is increasing hope that new methods, such as chromatin immunoprecipitation microarrays[89] and new biochemical insights (for example, into the characteristics of insulator sequences[90,91]), can enhance our understanding of regulation of gene expression.

1. Alberts, B (ed.). *et al. Molecular Biology of the Cell* 4th edn (Garland Science, New York, 2002).
2. Davidson, E. H. *Genomic regulatory systems: development and evolution* (Academic, San Diego, 2001).
3. Greenbaum, D., Jansen, R. & Gerstein, M. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18**, 585–596 (2002).
4. Schmid, C. D., Praz, V., Delorenzi, M., Perier, R. & Bucher, P. The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res.* **32**, D82–D85 (2004).
5. Fickett, J. W. & Hatzigeorgiou, A. G. Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878 (1997).
Demonstrated the poor performance of promoter-prediction software. Led to a shift from predicting specific transcription start sites, and towards prediction of regions that are likely to contain a TSS.
6. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578 (1990).
7. Antequera, F. Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci.* **60**, 1647–1658 (2003).
8. Hannenhalli, S. & Levy, S. Promoter prediction in the human genome. *Bioinformatics* **17** (Suppl. 1), S90–S96 (2001).
9. Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
10. Davuluri, R. V., Grosse, I. & Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nature Genet.* **29**, 412–417 (2001).
11. Adachi, N. & Lieber, M. R. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807–809 (2002).
12. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
13. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
14. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
15. Liu, R. & States, D. J. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res.* **12**, 462–469 (2002).
16. Suzuki, Y., Yamashita, R., Sugano, S. & Nakai, K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.* **32**, D78–D81 (2004).
17. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* (2003).
Introduces a new method for the identification of TSS on the basis of improved laboratory methods for the generation of full-length cDNAs. The data generated from this method will be important for the identification of alternative promoters.
18. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4**, 251–262 (2003).
19. Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I. & Hardison, R. C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* **13**, 1–12 (2003).
20. Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
21. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
22. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
23. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
24. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
25. Levy, S. & Hannenhalli, S. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome* **13**, 510–514 (2002).
26. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2**, 13 (2003).
Demonstrates that phylogenetic footprinting can eliminate an order of magnitude of false-positive transcription-factor binding-site predictions, in exchange for a modest sensitivity decrease.
27. Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G. & Pfeifer, D. Comparative genomics of the SOX9 region in human and *Fugu rubripes*: conservation of short regulatory sequence elements within large intergenic regions. *Genomics* **78**, 73–82 (2001).
28. Aparicio, S. *et al.* Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl Acad. Sci. USA* **92**, 1684–1688 (1995).
29. Santini, S., Boore, J. L. & Meyer, A. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. *Genome Res.* **13**, 1111–1122 (2003).
30. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
31. Storm, C. E. & Sonnhammer, E. L. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* **13**, 2353–2362 (2003).
32. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32**, D35–D40 (2004).
33. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
34. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
One of the best progressive alignment algorithms for global genome sequence alignment that facilitates phylogenetic footprinting.
35. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
36. Brudno, M. *et al.* Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19** (Suppl. 1), I54–I62 (2003).
37. Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**, 832–839 (2002).
38. Elnitski, L. *et al.* PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics* **80**, 681–690 (2002).
39. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13**, 64–72 (2003).
A new method to classify functions of conserved regions as regulatory or coding on the basis of the pattern of identical nucleotides.
40. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
A first look at methods to analyse large sets of orthologous eukaryotic gene sequences.
41. Montgomery, S. B. *et al.* Sockeye: A 3D environment for comparative genomics. *Genome Res.* (in the press).
42. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
One of several papers by Davidson that constructs the argument that genes are regulated by composite interactions of transcription factors that interact with locally dense clusters of binding sites.
43. Palstra, R. J. *et al.* The β-globin nuclear compartment in development and erythroid differentiation. *Nature Genet.* **35**, 190–194 (2003).
44. Fickett, J. W. Quantitative discrimination of MEF2 sites. *Mol. Cell Biol.* **16**, 437–441 (1996).
45. Fickett, J. W. Coordinate positioning of MEF2 and myogenin binding sites. *Gene* **172**, GC19–GC32 (1996).
46. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**, 231–245 (1997).
Demonstration that matrix-based profiles for the prediction of transcription-factor binding sites accurately predict *in vitro* binding.
47. Pollock, R. & Treisman, R. A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res.* **18**, 6197–6204 (1990).
48. Bulyk, M. L., Gentalen, E., Lockhart, D. J. & Church, G. M. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnol.* **17**, 573–577 (1999).
49. Shultzaberger, R. K. & Schneider, T. D. Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.* **27**, 882–887 (1999).
50. Roulet, E. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnol.* **20**, 831–835 (2002).
51. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
An excellent explanation of the relationship between scores that are produced by binding-site profiles and binding energy.
52. King, O. D. & Roth, F. P. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.* **31**, e116 (2003).
53. Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).
54. Udalova, I. A., Mott, R., Field, D. & Kwiatkowski, D. Quantitative prediction of NF-κ B DNA-protein interactions. *Proc. Natl Acad. Sci. USA* **99**, 8167–8172 (2002).
55. Barash, Y., Elidan, G., Friedman, N. & Kaplan, T. in *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology* (eds Vingron, M., Istrail, S., Pevzner, P. and Waterman, M.) 28–37 (ACM, New York, 2003).
56. Benos, P. V., Bulyk, M. L. & Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **30**, 4442–4451 (2002).
Summary of several key papers that demonstrate that matrix profiles provide reasonable predictions of binding sites in most cases.
57. Owen, G. I. & Zelent, A. Origins and evolutionary diversification of the nuclear receptor superfamily. *Cell. Mol. Life Sci.* **57**, 809–827 (2000).
58. Roulet, E. *et al.* Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J. Mol. Biol.* **297**, 833–848 (2000).
59. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).

60. Bray, N., Dubchak, I. & Pachter, L. AVID: a global alignment program. *Genome Res.* **13**, 97–102 (2003).
61. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
62. Lenhard, B. & Wasserman, W. W. TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135–1136 (2002).
63. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
64. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
An examination of the patterns of sequence evolution in regulatory regions. Surveys the genetic consequences of changes in binding sites.
65. Tagle, D. A. *et al.* 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.* **203**, 439–455 (1988).
One of several papers from the group that, to the best of our knowledge, established the phrase 'phylogenetic footprinting'.
66. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
67. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
68. Frith, M. C., Li, M. C. & Weng, Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **31**, 3666–3668 (2003).
69. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566 (2001).
Demonstration that coupling module predictions with phylogenetic footprinting can result in reliable predictions of regulatory sequences.
70. Liu, R., McEachin, R. C. & States, D. J. Computationally identifying novel NF-κ B-regulated immune genes in the human genome. *Genome Res.* **13**, 654–661 (2003).
71. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
72. Johansson, O., Alkema, W., Wasserman, W. W. & Lagergren, J. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* **19** (Suppl. 1), I169–I176 (2003).
73. Bailey, T. L. & Noble, W. S. Searching for statistically significant regulatory modules. *Bioinformatics* **19** (Suppl. 2), II16–II25 (2003).
74. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of *cis*-regulatory modules. Bioinformatics **19** (Suppl. 2), II5–II14 (2003).
75. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3**, 30 (2002).
An excellent algorithm for the detection of locally dense clusters of transcription-factor binding sites, particularly orientated towards large clusters of sites for a single factor.
76. Lifanov, A. P., Makeev, V. J., Nazina, A. G. & Papatsenko, D. A. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* **13**, 579–588 (2003).
77. Sandelin, A. & Wasserman, W. W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.* (in the press).
78. Gelfand, M. S., Novichkov, P. S., Novichkova, E. S. & Mironov, A. A. Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform.* **1**, 357–371 (2000).
79. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
80. Aerts, S. *et al.* Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.* **31**, 1753–1764 (2003).
81. Vadigepalli, R., Chakravarthula, P., Zak, D. E., Schwaber, J. S. & Gonye, G. E. PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. *Omics* **7**, 235–252 (2003).
82. Klingenhoff, A., Frech, K., Quandt, K. & Werner, T. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**, 180–186 (1999).
83. Berezikov, E., Guryev, V., Plasterk, R. H. & Cuppen, E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.* **14**, 170–178 (2004).
84. Kel-Margoulis, O. V., Ivanova, T. G., Wingender, E. & Kel, A. E. Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac. Symp. Biocomput.* 187–198 (2002).
85. Sharan, R., Ovcharenko, I., Ben-Hur, A. & Karp, R. M. CRÈME: a framework for identifying *cis*-regulatory modules in human–mouse conserved segments. *Bioinformatics* **19** (Suppl. 1), I283–I291 (2003).
86. Felsenfeld, G. Quantitative approaches to problems of eukaryotic gene expression. *Biophys. Chem.* **100**, 607–613 (2003).
87. O'Brien, T. P. *et al.* Genome function and nuclear architecture: from gene expression to nanoscience. *Genome Res.* **13**, 1029–1241 (2003).
88. Levitsky, V. G., Podkolodnaya, O. A., Kolchanov, N. A. & Podkolodny, N. L. Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics* **17**, 998–1010 (2001).
89. Shannon, M. F. & Rao, S. Transcription: of chips and ChIPs. *Science* **296**, 666–669 (2002).
90. Gerasimova, T. I. & Corces, V. G. Chromatin insulators and boundaries: effects on transcription and nuclear organization. *Annu. Rev. Genet.* **35**, 193–208 (2001).
91. West, A. G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes Dev.* **16**, 271–288 (2002).
92. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
93. Christensen, T. H., Prentice, H., Gahlmann, R. & Kedes, L. Regulation of the human cardiac/slow-twitch troponin C gene by multiple, cooperative, cell-type-specific, and MyoD-responsive elements. *Mol. Cell Biol.* **13**, 6752–6765 (1993).
94. Parmacek, M. S. *et al.* A novel myogenic regulatory circuit controls slow/cardiac troponin C gene transcription in skeletal muscle. *Mol. Cell Biol.* **14**, 1870–1885 (1994).
95. Kel, A. E. *et al.* MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).
96. Clamp, M. *et al.* Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42 (2003).
97. Lee, Y. *et al.* Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12**, 493–502 (2002).
98. Hollich, V., Storm, C. E. & Sonnhammer, E. L. OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics* **18**, 1272–1273 (2002).

Acknowledgements

W.W.W. is supported by a grant from the Canadian Institutes of Health Research.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink

endo16 | *TNNC1*

FURTHER INFORMATION

Accompanying online exercises:

http://www.phylofoot.org/NRG_testcases

Access to this interactive links box is free online.