Published online July 19, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 12 ***3585–3598***
doi:10.1093/nar/gkl372

SURVEY AND SUMMARY

Computational identification of transcriptional regulatory elements in DNA sequence

Debraj GuhaThakurta*

Research Genetics Division, Rosetta Inpharmatics LLC (a wholly owned subsidiary of Merck & Co., Inc.), 401 Terry Avenue North, Seattle, WA 98109, USA

Received March 5, 2006; Accepted April 27, 2006

ABSTRACT

Identification and annotation of all the functional elements in the genome, including genes and the regulatory sequences, is a fundamental challenge in genomics and computational biology. Since regulatory elements are frequently short and variable, their identification and discovery using computational algorithms is difficult. However, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements. The availability of complete genome sequence from multiple organisms, as well as mRNA profiling and high-throughput experimental methods for mapping protein-binding sites in DNA, have contributed to the development of methods that utilize these auxiliary data to inform the detection of transcriptional regulatory elements. Progress is also being made in the identification of *cis*-regulatory modules and higher order structures of the regulatory sequences, which is essential to the understanding of transcription regulation in the metazoan genomes. This article reviews the computational approaches for modeling and identification of genomic regulatory elements, with an emphasis on the recent developments, and current challenges.

INTRODUCTION

The identification of genomic regulatory elements is an important but unsolved problem in genome annotation (1). Of the ~5% of mammalian genome that is estimated to be under evolutionary selection pressure, less than a third is coding (2,3). The remaining portion is believed to be composed of untranslated regions, non-coding genes, chromosomal structural elements and regulatory elements that control a variety of biological processes including gene expression, translation, chromosomal replication and condensation. However, little is currently known about the vast array of these regulatory elements. Whereas the number of coding genes in many of the sequenced organisms can now be reasonably estimated, there is no clear estimate of the number of functional regulatory elements in these genomes, especially in the metazoa. In eukaryotes ranging from the nematodes to flies to the mammals, the number of coding genes is similar (2,4–6), and it is now thought that organismal complexity may be attributed to phenomena such as alternative splicing, DNA rearrangement and increased number of transcriptional regulatory elements as well as transcription factors (TFs) which regulate gene expression (7). The identification of *cis*-regulatory elements controlling gene expression, and characterization of their interaction with the respective TFs, thus lie not only at the very heart of understanding of the network of gene interactions but also of explaining the origins of organismal complexity and development.

Genomic regulatory elements are frequently represented by DNA motifs. As such these representations are general and can be used to describe any class of short DNA sequence elements. However the theories for weight matrix model, which is a common way to represent a collection of DNA elements, are based on the biophysical considerations of protein–DNA interactions (8–11) (as described in the following section). Therefore, here we primarily discuss the transcriptional regulatory elements, more specifically the DNA sites that are bound by the TFs.

It is estimated that there are ~2000 TFs in the mammalian genomes (2,5) and ~1000 in the flies and worms (7). However, only for a minority of the TFs (~900 in human, 700 in mouse, 200 in *Drosophila* and 100 in *Caenorhabditis elegans*) is there currently any known information on binding sites or interacting protein partners (12). DNA-binding site models are available for ~500 vertebrate TFs, and <5000 genomic sites are known in all vertebrates in fewer than 3000 genes (12). Based on the information available from a few of the well-studied genes (13), it appears that the total number of such sites in the multicellular genomes could be at least an order of magnitude higher than the number of coding genes, i.e. in the order of hundreds of thousands or more.

*Tel: +1 206 802 6430; Fax: +1 206 802 6377; Email: debraj_guhathakurta@merck.com

Thus, our knowledge of the TFs, and especially their binding sites and regulated genes, is severely limited at this present time.

Computational methods for modeling and identification of DNA regulatory elements have been developed over the past two and a half decades (14–20). Orthogonal information from comparative genomics or co-regulation at the transcriptional level have also been integrated into these methods to identify *cis*-regulatory sites (21–25). More recently, methods have been developed (26–32) to analyze composite regulatory elements, i.e. modules consisting of multiple DNA sites bound by the regulatory factors (33). All of these methods have been valuable in expanding our limited knowledge of regulatory elements in the genome.

Here we discuss the progress made in the computational identification of genomic regulatory elements, recent advances, the utility of orthogonal data, existing challenges in the field and some selected examples where these methods have been successfully applied to discover novel functional elements. Rather than exhaustively covering the literature, we focus on the key concepts. Significant progress has also been made in the experimental characterization of regulatory sequences, a discussion of which is beyond the scope of this article, and the reader is referred elsewhere for further information (34–43).

REPRESENTATION AND SEARCH OF DNA REGULATORY ELEMENTS

Recurrent motifs in a collection of DNA sites are most commonly modeled by sequence patterns (also called regular expressions), or by position weight matrices (PWMs, also called profiles and position-specific scoring matrices, PSSMs). The sequence patterns are simply strings over the 4-letter alphabets [A,C,G,T] that form the DNA. In order to capture variation in a specific position, the degenerate IUPAC nucleic acid codes (44) (http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html, nomenclature for incompletely specified bases in nucleic acid sequences) are used (Figure 1). An L-long sequence motif can also be represented by a $4 \times L$ matrix with weights giving the frequency of the four DNA bases (or the logarithm, see below) in each of the L positions (18,45,46) (Figure 2). It is worth noting here that some DNA-binding site motifs are bipartite, i.e. have two halves that are sometimes separated by a spacer element in between. Such bipartite motifs can often be palindromic, e.g. 5′-CGGnnnnnnnnnnCCG-3′, the binding site for yeast TF, Gal4 (47).

```
W = A or T
S = C or G
R = A or G
Y = C or T
K = G or T
M = A or C
B = C, G, or T (not A)
D = A, G, or T (not C)
H = A, C, or T (not G)
V = A, C, or G (not T)
N = A, C, G, or T
```

Figure 1. The IUPAC (International union of pure and applied chemistry) code for representing degenerate nucleotide sequence patterns.

The pattern and PWM representations serve as complementary approaches and have been used widely since the 1980s. The DNA sequence patterns are simpler in representation, and advantageous in terms of exact enumeration of their significance in the genome using statistical methods (explained in detail later). The PWMs are able to capture information on the variability of a collection of DNA sites in a quantitative manner, which is not possible with the DNA patterns. However when detecting significant PWMs in DNA sequences, heuristic methods have to be used instead of exhaustive enumeration. Perhaps the most important practical utility of these models is their application in scanning DNA sequences for new regulatory element candidates. For this, one needs the appropriate motif models representing the regulatory elements, a statistical framework to score sites and determine their significance, and suitable thresholds to minimize false positives and false negatives. These issues are discussed below in more detail.

Currently there are two comprehensive and curated databases containing information on TFs binding site profiles (12,48). JASPAR (48) contains a smaller set that is non-redundant (i.e. each TF has only one profile), while TRANSFAC (12) contains multiple profile models for some TFs. In addition to the above generic databases, other organism-specific databases exist that host transcriptional regulation data (47,49–51) (http://arep.med.harvard.edu/dpinteract/). Profiles are biased by the observed sites on which they are built. Therefore if a large enough sampling of sites is not available, they may not capture all possible variations of the functional sites. This tends to produce false negative calls on new sites that do not match well with the previously characterized ones. The binding sites for structurally related TFs are often similar, and in such cases building familial binding profiles instead of profiles for each TF may be useful (52).

In both patterns as well as PWM representations, the significance of a particular motif is given by a measure of statistical surprise (or likelihood) for the motif given the data. In the case of patterns, significance can be calculated given the distribution of all occurrences of patterns using standard statistical procedures (20,53). This has been used in several software packages that discover over-represented patterns from input sequences (54–58). In the methods using weight matrix models, the measure of significance is commonly given by the information content (IC, also called relative entropy) (8,9,11):

$$I(p) = \sum_{j=1}^{L} \sum_{i=\mathrm{A}}^{\mathrm{T}} \mathrm{f}_{i,j} \log \frac{f_{i,j}}{P_i}, \qquad \mathbf{1}$$

where I(p) is the IC for the PWM representing a pattern p, L is the pattern length, i is the index of a base {range A through T} at position j of the PWM, $f_{i,j}$ is the frequency of base i at position j of the PWM, and P_i is the probability of observing that base in the data. Based on biophysical models of protein–DNA binding, it has been shown that the contributions of the individual positions of a site to the

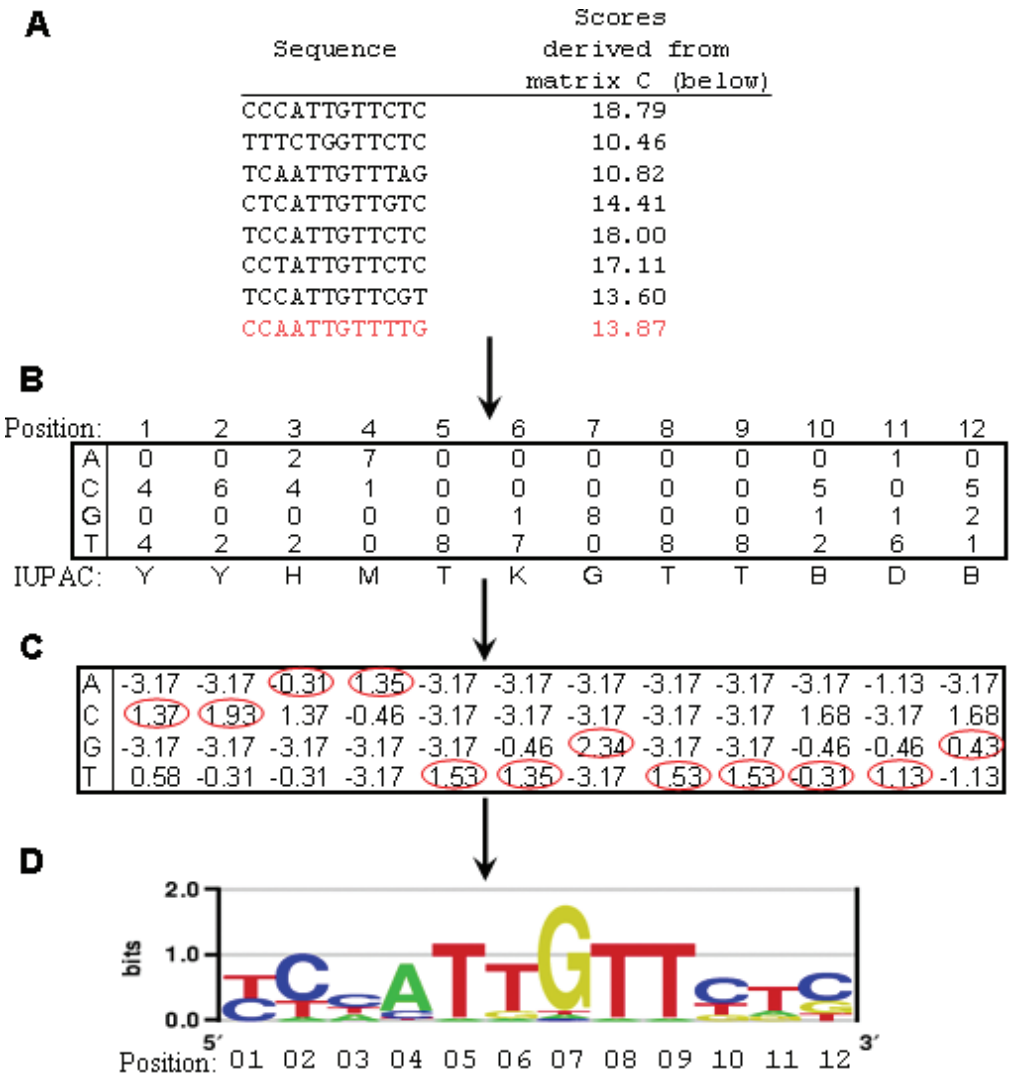

Figure 2. (**A**) The collection of eight known Rox1-binding sites taken from SCPD (47). Scores of the sites are according to the PWM described in (C). (**B**) Alignment matrix and IUPAC representation of the eight Rox1-binding sites. The cells represent the number of times a base *i* is observed at position *j* in the alignment of sites. The frequencies, $f_{i,j}$, of base *i* at position *j* of the binding sites can be obtained by dividing the values in the cells of the alignment matrix by the total number of sites, e.g. $f_{C,1} = f_{T,1} = 4/8 = 0.5$. (**C**) PWM for scoring sequences. Each weight is given by $\log_2(f_{i,j}/P_i)$ (see text), where P_i is the probability of observing the base *i* in the data; here we have taken $P_A = P_T = 0.32$, and $P_C = P_G = 0.18$ (corresponding to the *S.cerevisiae* genome). A pseudocount of 1 was added to the alignment before deriving the weights. This matrix was used to score the sites in A. As an example, the score of the site in red (sequence CCAATTGTTTTG, score 13.87) is given by the summation of the scores that are circled in red. Note that the scores of the two consensus sequences, CCCATTGTTCTC and TCCATTGTTCTC are different because $P_C \neq P_T$. (**D**) Sequence logo representation (187) of the alignments, visually showing the IC and conservation at each of the alignment positions. The IC of this matrix is 11.3 bits or 7.83 nats (Equation 1).

total binding free energy of the TF may be given by $\log(f_{i,j}/P_i)$ (8,11,46), which are often used as the weights (or log-odds scores) in a PWM (Figure 2C). The IC is therefore the weighted average of the binding energies from each of the sites represented in the matrix, and lower IC indicates higher variability (consequently lower specificity) in the sites. Statistical methods for computing the *P*-value of IC have been defined (59,60). Several algorithms use the IC measure to identify optimal motifs from input sequences (59,61–64).

Given a DNA motif, searching sequences for candidate sites is straightforward; but distinguishing real sites from artifactual ones is difficult. Consequently the signal-to-noise ratio in such searches is often small. The problem of distinguishing true versus false sites arises from the fact that TF DNA-binding sites are usually degenerate and many subsequences may match a given motif. The situation is illustrated with the collection of eight known sites for a *Saccharomyces cerevisiae* TF, Rox1 [taken from the SCPD database (47)] (Figure 2). Even though there is a preferred base at 11 of the 12 positions, at 8 of those more than one base is tolerated (in a larger collection of sites, even these 4 conserved positions may show variations). Only two of the known binding sites correspond to the consensus pattern, (T/C)CCA-**TTGTT**CTC (conserved positions are shown in boldface), so a search of the genome with this consensus pattern would miss many functional sites. On the other hand, the IUPAC representation of the sites, YYHMTKGTTBDB,

captures the variation of all the sites but is also likely to find a large number of false positives and non-functional sites in a genome-wide search.

Given a sequence and a PWM, one can score any subsequence in that input. Suppose the length of the PWM is L and the weight of base i at position j is $w_{i,j}$, the score of a subsequence (s), when aligned to the PWM, is then given by the following equation:

$$\text{Score}_s = \sum_{j=1}^{L} \sum_{i=\text{A}}^{\text{T}} w_{i,j} \cdot S_{i,j}, \quad \mathbf{2}$$

where $x_{\text{i,j}} = 1$ if base i occurs at position j of subsequence s and 0 otherwise. This simple scheme is commonly used (59,65,66), and assumes an additive contribution from each position towards the score. Methods have been described to calculate the P-values of these scores based on their distribution (59,65,67). Similar to DNA patterns, site predictions with the PWMs can suffer from high false positive rates if motifs are degenerate. If base composition of a genome is assumed to be random (which is not an accurate assumption, but helps us to discuss the following point in an approximate but simplified way), the Rox1 PWM (Figure 2C), which has an IC of 11.2 bits, would be expected to have a site every 2500 nt [for details see (9)], or ~4700 total sites in the yeast genome. There are motifs for which the IC is even lower, leading to a much higher number of possible sites. For example, a collection of 13 known binding sites for UASH (47) has an IC of only 8.5 bits, giving ~32 000 potential sites in the yeast genome (or 1 every ~370 nt). Thus, the problem of distinguishing real from false (biologically non-functional) sites is an important but challenging one.

In addition to the motif degeneracy, there are other biological issues relating to the predictions of functional TF-binding sites in the genome, which are discussed below. Although there are limitations, the PWM models work fairly well in representing the specificity of DNA-binding sites and predicting TF-binding probability to a given site (11,18). Some TFs are by nature moderately or poorly specific in their DNA binding and achieve higher specificity only in the context of other binding partners. One also has to remember that the chromatin structure and DNA methylation play important roles in gene regulation (68–70). Large portions of the chromosomal DNA are sequestered by histones forming part of the nucleuosomal structure, and are therefore not accessible for binding by the TFs. DNA methylation can inhibit interaction of the regulatory proteins with cognate DNA sites and also influence the chromatin structure. While doing genome-wide searches for putative binding sites using a motif model, one typically does not know the chromosomal regions that are open for the regulatory proteins to bind, or (with a few exceptions) the binding partners for a given TF. A blind search of the entire genome without such information usually returns a large number of sites, many of which would probably bind to the TF (strongly or weakly, depending on the match of the sites to the model and the specificity of the model) if the DNA sequences were open for binding, but are biologically non-functional *in vivo*. Genomic sequences that play an active role in transcriptional regulation, such as the promoters, may often be outside of the nucleosome structure, or at least available a part of the time due to chromatin remodeling (71). So the rate of non-functional site predictions in these sequences is likely to be lower than other parts of the genome. However, without the information on DNA availability, methylation status or other binding partners, the binding site predictions with individual models are unlikely to achieve the same level of specificity that are achieved *in vivo* by the TFs.

Despite the fact that for individual PWM models determining thresholds for distinguishing real versus non-functional sites is difficult, computational approaches exist that address this issue. One approach is based on the IC of the PWMs. A sample-size adjusted IC may be defined as the true IC (Equation 1) minus the IC expected from an arbitrary alignment of an equal number of random sites. The PATSER program (59) (ftp://ftp.genetics.wustl.edu/pub/stormo/Consensus/) uses a default cutoff score for which the log_e(probability) (i.e. the probability of observing a score greater or equal to the cutoff) equals the sample-size adjusted IC. Another approach is based on prediction rates on sequences that contain known sites and on sequences that are likely not to contain sites (66,72). False negative rates are computed from predictions made on the experimentally characterized TF-binding sites, and false positive rates may be computed, for example, from predictions in exons, where the number of regulatory elements is expected to be low (66). Given these prediction rates, the selection of the appropriate cut-off score depends largely on the user's objectives. A third approach has been to use a distribution of scores for a PWM and use one or two standard deviations below the mean score as a threshold for filtering out low-scoring sites (73), but retaining most of the true positive ones. Recently, Djordjevic *et al*. (10) have described a support vector machine (SVM) method for representation of DNA motifs based on thermodynamic principles of protein–DNA binding. The biophysical treatment and optimization of the SVM automatically provides a threshold to distinguish the binding energies of the real sites versus those that are likely to be false. This natural and objective thresholding is one of the strengths of their approach and it provides a way to avoid the use of the more subjective or user-defined thresholds that are frequently employed.

Since many of the TFs bind DNA in the context of other TFs, where information is available about the organization of multiple DNA-binding sites or the binding partners for a given TF, it can be utilized to reduce the rate of false positive predictions. Even when individual PWMs tend to be fairly non-specific, searching for co-occurrence of binding sites that form regulatory modules has been shown to be an effective approach to increasing prediction specificity without losing sensitivity (74).

Finally, we discuss one limitation of the additive PWM models in representing DNA sites, and recent developments which address this issue. As seen from Equations 1 and 2, the simple PWM approach assumes independent contributions from each position within a DNA site towards the binding free energy of the TF (mononucleotide model). This has been shown not to hold true in several situations (37,75,76) where sufficient binding site data are available

for detailed analysis. In such cases, higher-order models, i.e. those that consider interdependence between positions within a site, are better representations and give more accurate predictions when searching for candidate sites (10,76–79). Most of the higher-order models (10,75–81) use di-nucleotide interactions, which is a reasonable compromise between the mononucleotide PWMs and the more complete models with interactions between multiple (more than two) positions. This is because (i) the large number of characterized binding sites that is needed to build complex models (with increased number of parameters) is rarely available for any particular TF (12) and (ii) in a few cases where sufficient experimental data are available (37,75), interdependencies between adjacent nucleotide positions appear to be the most significant of the within-site interactions.

From the DNA-binding site data available for a small number of TFs, it has been estimated that ~25% of sites may show significant within-site positional correlations (76). Even in cases where intra-site interactions exist, the simpler additive model has been suggested to be a good approximation (80). As more experimental data become available with the application of high-throughput methods for characterizing TF-binding sites, it remains to be seen if significant interdependence between positions of DNA-binding sites is prevalent in TF–DNA interactions, and whether the more complex sequence models provide a considerable enhancement in modeling accuracy over the additive PWMs in many cases.

Since the structural contexts of both DNA and protein can contribute towards specific DNA binding by TFs (82), some studies have investigated the use of structural information of protein–DNA recognition in building predictive models for DNA-binding sites (82–88). The structure-based approaches are promising as they can predict binding sites for TFs where no previous sites have been characterized before (85), and can provide improved predictive power over the simple sequence profile models (83). However, they have been limited in their general use so far, since derivation of the quantitative structural parameters is dependent on the small number of solved protein–DNA complex structures. A more thorough discussion of the structural aspects of protein–DNA recognition is beyond the scope of this review and the readers are referred to the articles above (and references therein) for further information on this topic.

With larger datasets, the issues with modeling complex DNA–protein interactions can be investigated more thoroughly than what has been done so far, and the building of models (with or without structural parameters) with higher accuracy is likely to become more feasible.

DISCOVERY OF DNA MOTIFS

Discovery of motifs in sequence data was an early problem to be addressed in computational biology. The DNA motif discovery algorithms that have been developed can be divided into two categories, namely pattern driven methods (those that identify DNA patterns) and sequence driven or alignment driven methods (those that identify profile models) (16,20). The concepts behind these algorithms, their advantages and limitations, and a few example methods are discussed below.

In the pattern driven methods, given a set of DNA sequences and a length L of pattern that we wish to find, the challenge is to identify the most significant patterns of that length. The solution to the problem can be obtained by generating all possible patterns of length L, searching for the number of occurrences of each pattern, and then reporting the ones with highest frequency as being the most significant in the given data (52–58,89–91). This enumerative approach is exact and guaranteed to find optimal solutions in the restricted search space. Approximate sequence patterns, or patterns that contain degeneracy at one or more positions, can also be identified from the sequence. The similarity between any two patterns may be given by the Hamming distance (the number of positions in which they differ) or the Levinstein distance (the number of substitutions, insertions or deletions needed to transform one string into another) (20). When multiple patterns are close in terms of their distance, they can be merged into one approximate pattern (89,92). Although the exact enumeration is an advantage of these methods, one limitation is that searching for long patterns is computationally expensive, and an exhaustive search through the sequence space of 4^L words often becomes impractical for $L > 10$ (93). Two general strategies have been taken to address this limitation: (i) the use of efficient methods [e.g. pattern graphs (93) or projections (94)] for preprocessing the data so that the search space is reduced and actual pattern search becomes less expensive, and (ii) combining the shorter overlapping patterns found from the data to yield longer or more complex patterns (89,92,95,96). In addition to the exact enumerative methods, efficient data structures like the suffix trees (97) have also been applied to the DNA pattern discovery problem (98–100). Although not exact algorithms, the advantage of the suffix trees is that they allow one to search for patterns of longer lengths since the search time is not exponential in the length of the patterns, but exponential in the number of mutations to be tolerated in the sites (98,100).

In the sequence driven methods the challenge is to find the location of the sites and the representative PWM using only the sequence data, without any assumptions on the statistical distributions of patterns in the sequences. If the locations of sites are known, building a DNA profile for them is trivial. However in the *ab initio* motif discovery problems, this information is not known ('missing information') and has to be learned from the input data. Such problems with missing information can be solved by employing machine learning algorithms. Several machine-learning approaches have been applied to the problem of motif finding (59,62–64,76,101–107). Unlike some pattern driven methods where the most significant motifs can be identified by exact enumeration, obtaining the globally optimal results cannot be guaranteed in these methods. But motifs of arbitrary lengths can be searched, since the search time does not depend significantly on the length of sites.

The first amongst the sequence driven methods was the greedy algorithm (59,61,62). Given a set of n sequences, and a motif length L to be searched, this algorithm progressively builds matrices by including the sites which maximize the IC (Equation 1). The algorithm first builds a set of significant matrices by comparing all pairs of sequences. In subsequent iterations, sites that increase the IC of the alignment are

added to the matrix. Another method that has been commonly used is the expectation–maximization (EM) algorithm (63,102,103). The EM algorithm simplifies the analysis of problems with missing information by iteratively substituting the locations of sites by expected locations. The algorithm starts with a guess PWM, which can be random or based on some prior knowledge about the binding sites [see e.g. (52,94)]. Using the PWM, the probability of each subsequence being a binding site is estimated, and the PWM is updated based on those probabilities. This cyclic process is iterated until a convergence criterion is reached. A stochastic variant of the EM algorithm that is now very widely used in sequence motif recognition is the Gibbs sampling method (64,106). Gibbs sampling, which is a type of Markov chain Monte Carlo (MCMC) algorithm, tends to provide a more robust optimization of the PWMs as the probabilistic sampling of sites helps the avoidance of local minima. Variations of the Gibbs sampling technique have been implemented in many motif finding software that are used in the bioinformatics community (76,104,105,108,109). For stochastic algorithms like the Gibbs sampling, multiple searches have to be performed with the input dataset in order to confirm that the same motifs are discovered starting from different random points in the sequence space.

Controlling for background sequence is an important issue in DNA motif discovery that has been effectively addressed in recent years. Variation in local base composition can adversely affect sequence alignment and discovery of relevant motifs. Because such variations can be complex, and since binding motifs are often AT- or GC-rich, these adverse effects can be difficult to control using existing masking algorithms. In addition, some low complexity sequence motifs [like ploy(A) or poly(GC)] are widespread in genomes of certain organisms. As a consequence, often the strongest motifs that are discovered from any input data are these common motifs that are prevalent in the genome but do not represent the specific regulatory elements that are being sought. It is therefore important to detect those motifs that are significantly more frequent in the positive sequence set (the input set in which we want to discover motifs) relative to the background (91,92). The available programs which identify such motifs (often called discriminative motifs) do so by modeling the background with a Markov chain (105,109,110), or by accounting for the motifs that are frequent in a given background set (26,92,108) through sampling or enumeration.

Recently, Tompa *et al*. (111) reported the most comprehensive comparative study yet performed for different *ab initio* motif-finding algorithms. Assessment was done for 13 commonly used DNA motif discovery tools that do not use any auxiliary information, such as comparative sequence analysis, mRNA expression levels or chromatin immunoprecipitation (ChIP–chip) data. Predictions were compared with known binding sites, using various statistics to assess their accuracy. One important, but not unexpected, observation from this work was that a few different tools tend to complement each other's performance. For example, MotifSampler's (109) predictions complement well the predictions of MEME (101,102), oligo/dyad-analysis (56,90), ANN-Spec (108) and YMF (110). The authors rightly suggest that biologists would be well advised to use a few complementary tools in combination rather than relying on a single one, and to pursue the top few predicted motifs of each rather than the single most significant motif (111).

There are currently some examples where *ab initio* motif finding algorithms have been employed to identify novel regulatory elements that have been validated by follow-up experimental studies. A few of these are described below with the objective of illustrating the types of input data that can be used and the approaches that were taken in the studies. Using the program AlignACE on the upstream promoter region of 248 distinct groups of genes in yeast *S.cerevisiae*, Hughes *et al*. (104), generated a list of 3311 putative regulatory motifs. By applying stringent thresholds and selecting motifs which had highest specificity for certain groups of genes, this large set was reduced to a small set of 54 motifs that were then grouped to 25 motif clusters based on the similarity of multiple motifs. Of the 25 motif clusters, 16 were previously known motifs representing the binding sites for the yeast TFs. One of the previously unidentified motifs, representing the binding sites for the TF Rpn4, was verified using mRNA expression analysis of the Rpn4 protein knockout and overexpressing strains. GuhaThakurta *et al*. (112) have described the identification of one novel regulatory element in the nematode *C.elegans* heat-shock response starting from a set of microarray experiments in which genes were robustly upregulated on heat-shock treatments at both early and late time points. Using a set of 28 heat-shock upregulated genes they used the *ab initio* motif discovery algorithms, Ann-Spec (108) and Consensus (59,108), to identify two strong motifs that are over-represented in the promoters of these genes. One of those motifs was the previously characterized heat-shock element that is broadly conserved in eukaryotes and known to bind to the TF called heat-shock factor. The second motif, which was novel, was shown to be biologically functional *in vivo* through transgenic and mutational studies. In fact, the two elements were shown to function in a cooperative manner; the contribution to heat-shock mediated expression by either one was weak, but when placed together in close proximity they strongly regulated expression. Studies similar to the one described above was done to identify several new muscle regulatory elements in *C.elegans,* which were shown to contribute to muscle expression in a cooperative fashion (113). The identified elements were also highly predictive of additional muscle-specific genes in that organism (114). These studies show that *ab initio* prediction methods can be valuable in elucidating unknown regulatory elements not only in unicellular organisms, but also the more complex metazoan genomes.

Although much progress has been made in the methods for *ab initio* detection of DNA regulatory elements, the problem still remains a difficult one, especially where the input sequences are long and motifs are weak. Therefore, the incorporation of auxiliary information into such methods can be of significant benefit. Since the availability of many complete genomes, the application of comparative genomics in the identification of DNA regulatory elements has become an important area of study and it is covered in detail in the next section. Here, as an example, we discuss a different approach which leverages mRNA expression data in DNA motif discovery. Bussemaker *et al*. (24) described recently a method to discover regulatory elements that uses correlation of DNA patterns with the expression levels of genes in a

single profiling experiment. They use a simple linear regression model to fit the logarithm of the expression of each gene to the sum of contributions from a set of DNA patterns in the upstream promoter region. All the genes are simultaneously fit, and statistically significant patterns that best fit the expression data are selected. Using yeast expression datasets, they reconfirm most of the motifs originally found by clustering of expression data and then running motif finding algorithms on clustered gene sets (115,116). Since there are frequently cooperative (non-additive) interactions between TFs regulating a gene, the modeling is simplistic. However, the advantage of this method is the fact that limited expression data are required for analysis and discovery of the DNA motifs. A similar approach has also been used recently by Conlon *et al*. (117). In addition to the above methods which integrate motif discovery with expression data, several tools and studies have been described that identify DNA-binding site motifs for TFs from a set of sequences defined by mRNA profiling (112,118,119) or ChIP–chip data (120,121).

COMPARATIVE GENOMICS IN THE SEARCH FOR REGULATORY ELEMENTS

In multicellular organisms the sequence space in which regulatory elements can be present in the genome is vast. In addition to the core promoter region, auxiliary transcription regulatory elements like enhancers, silencers and insulators can be present in the distant 5′ upstream region, 3′ downstream region and the introns (122). In *Drosophila* such DNA elements can be spread over a region of 10 kb around the genes whereas the average transcribed DNA is 2–3 kb, and in mammals these elements can be scattered over distances of hundreds of kb (123). Approaches that help to limit this search space are hence of significant value to the analysis of regulatory elements. Also, it is intriguing to study those regulatory mechanisms and elements that are likely to be of fundamental importance in maintaining certain cellular functions and therefore conserved in evolution. Consequently, phylogenetic footprinting (124,125), which is based on the premise that functional elements are likely to be under selection pressure and thereby evolve at a rate that is slower than the surrounding non-functional sequence, has been widely applied in recent years to the identification of regulatory sequences (21,123,126–129). Purely for the purpose of illustration, a simple example of the application of phylogenetic footprinting for the identification of putatively conserved TF-binding sites is given in Figure 3.

Enrichment of regulatory elements has been clearly demonstrated in non-coding DNA in the human–mouse conserved regions (21,127,130). For example, Wasserman *et al*. (21) found that 98% of the known muscle regulatory elements are located within 19% of the sequence that is most conserved between human and mouse. Non-coding sequences that can be aligned across two or more organisms have been shown to have functional regulatory roles (123,129). Analysis of motif frequencies and correspondence in conserved regions across multiple species has been used to identify known and novel regulatory elements in organisms ranging from yeast (25,131–133) to the mammals (134,135). Based on human–rodent sequence alignments, and known sets of regulatory sequences and ancient repeats in those genomes, methods have been developed to distinguish conserved regulatory regions from neutral sequences (136,137). In addition to its application in eukaryotic genomes, comparative genomic approaches have helped in elucidation of regulatory elements in prokaryotes and archaea (22,138–141). Therefore, significant developments have been made over the past several years in utilizing the sequences from multiple species to identify functional regulatory elements in all phyla.

There are now numerous methods that are available for alignment of genomic sequences from two or more species (142–151). Some of these have been integrated with motif finding and visualization programs to provide practical tools for analysis of regulatory elements within cross-species conserved regions (152–154). Multiple sequence alignment methods can be advantageous in comparative genomics since they utilize more information relative to pairwise sequence alignments; they can also be more powered to identify regulatory motifs (135,140,155). Prakash and Tompa (135) recently compared the performance of six global and local multiple alignment tools with respect to their potential of identifying highly conserved short (10mers) DNA patterns from the immediate upstream promoter sequences of orthologous genes from multiple vertebrate species (human, chimp, mouse, rat, chicken). Two of the methods tested, namely MLAGAN (147) and TBA (156), appeared to perform better than several others for this purpose (135). More advanced methods have now been developed that directly take into consideration the phylogenetic distances between the organisms that are being aligned in order to identify conserved DNA motifs (107,126,157–159).

In addition to phylogenetic footprinting, comparative genomics is now being utilized in other sophisticated and intriguing ways to identify regulatory elements of potentially conserved function. Some methods have not only utilized sequence conservation but also gene network conservation (based on the hypothesis that multiple sets orthogonal genes may be regulated by common TFs) to identify sets of regulatory motifs (133,160). Although developed in yeast, one of these studies show that such methods can be sufficiently powered to detect regulatory elements in much longer sequences in the multicellular organisms, including mammalian genomes (133).

There are now many examples where comparative genome sequence analysis has been used successfully to elucidate novel regulatory elements which would have been difficult to identify without that information; three are illustrated below. McCue *et al*. (140) used an extended Gibbs sampling algorithm to identify probable transcription regulatory sites upstream of *Escherichia coli* genes by cross-species comparison. A set of 184 genes with orthologs from two or more other gamma proteobacterial organisms were analyzed. Of their predictions 81% corresponded with the documented sites known to regulate these genes, whereas 67% corresponded when data from only one other species were available, suggesting that addition of one more species aids in sensitivity of the binding site detection. One of the novel predictions, a DNA site bound by the TF YijC, was verified by experiments. In an elegant study, Loots *et al*. (123) demonstrated the utility of phylogenetic footprinting by identifying a regulatory region conserved between human and

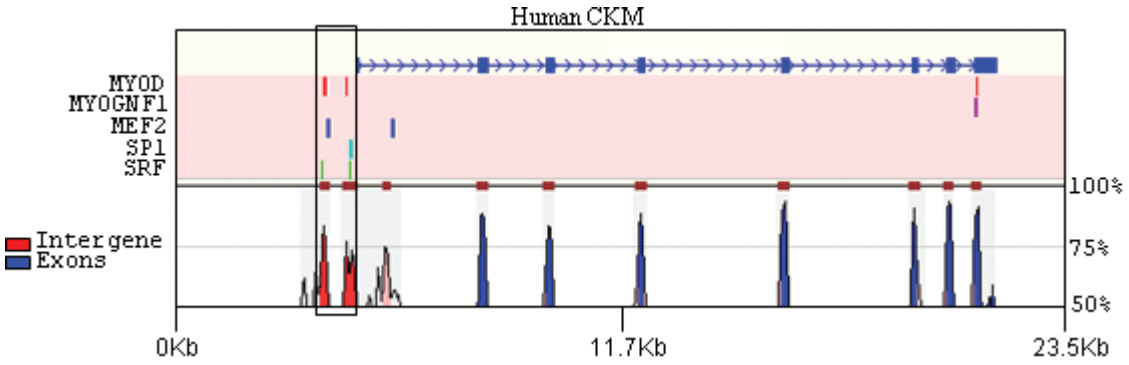

Figure 3. Predicted TF-binding sites in human–mouse conserved regions around the CKM (creatine kinase, muscle) gene. The genomic regions, along with 5 kb upstream and 2 kb downstream, of the CKM gene were extracted from the human and mouse genomes and aligned using the BLASTZ software (151). The BLASTZ alignments were then fed into the rVISTA program (153) through the website http://rvista.dcode.org. Binding sites for several TFs that are known to regulate gene expression in the muscle tissue were then predicted on the human sequence using the PWM models available from the TRANSFAC database (12). The predicted sites can be dynamically viewed and clustered through the above website. For the purpose of this current figure, we required that at least two binding sites belonging to different TFs be present within a window of 100 nt. A cluster of sites was observed in the immediate 5′ upstream region of this gene (boxed). Percent conservation between the two sequences is shown; regions with ⩾75% conservation are colored. The human gene structure is shown at the top in blue.

mouse for *IL-5* that is located ∼120 kb away from the gene itself. In another study comparing genomic sequences from multiple primate species, Boffelli *et al.* (129), identified regulatory elements for APOA, a recently evolved primate gene. A region of high conservation adjacent to the transcription start site was shown to interact with one or more DNA-binding proteins using electrophoretic mobility shift assays with nuclear extracts from liver cells. Identification of such primate-specific functional elements would be unattainable through the comparison of species that are evolutionarily more distant.

As evident from the discussions above, approaches that utilize conservation across species are very useful in elucidating functional DNA regulatory elements. Significant advances have been made in this area over the past 6 years. Despite its utility and widespread use, there are limitations in phylogenetic footprinting approaches with respect to their use in the identification of regulatory elements. The short regulatory elements may be missed if the genomic sequences that are being aligned come from distant organisms (112,161,162). If the organisms are too close, however, the alignments are extensive and therefore unable to distinguish the conserved functional elements from the non-functional ones (161,162). It is unclear at this time whether cross-species alignments would be equally useful in finding regulatory elements in all phylogenetic clades (163,164). In a recent comparison of two *Drosophila* species, the known regulatory elements were found to be only modestly enriched in the conserved regions (164), although the amount of conservation in the non-coding regions of these *Drosophila* species was roughly the same as in human–mouse. In another study, individual binding sites appear to be conserved across two nematode species, but they were not located in sequences that were aligned by the software for phylogenetic footprinting (112). Therefore, whether cross-species sequence alignment is likely to be an effective approach in all organisms in elucidating most of their functional regulatory elements, as well as the issues such as optimal phylogenetic distances, and the number of organisms needed to detect these elements, are still matters of debate and investigation (163).

COMPOSITE MOTIFS AND *CIS*-REGULATORY MODULES

In eukaryotes, TFs rarely act alone in regulating the expression of a given gene. In most cases multiple factors bind DNA, often in close proximity with each other, forming regulatory modules (13,33,165,166). By utilizing combinatorial interactions between multiple factors these *cis*-regulatory modules (CRMs) confer specific spatial and temporal patterns of transcription. Therefore identification of composite modules and higher order regulatory structures is currently an active area of research in computational analysis of regulatory sequences. The problem is difficult however since the combinatorial interactions between the regulating factors can be very complex (e.g. see the array of regulatory interactions in the immediate upstream region of *Endo*16 in sea urchin (13)]. There have been some developments in the past 6 years in addressing this problem in DNA sequence analysis. The current approaches can be classified as follows:

(i) Methods that identify modules given a set of DNA motifs representing sites for TFs that are known to act together in regulating transcription (27,29–31,74,167–171).
(ii) Methods for *ab initio* identification of multiple DNA motifs representing the sites for a CRM (26,93,98, 105,172–174).

Softwares that identify CRMs given a set of known DNA motifs fall into two categories, those that use hidden Markov models to represent the CRMs (27,169), and those that rely on observation of frequent joint occurrence of sites within a certain window, modeling the frequency of multiple sites with an appropriate statistical (e.g. Poisson) distribution (30,31,74,167,169,170,175,176). Because the DNA-binding site models for the majority of TFs are currently unknown and the information on TFs that bind DNA together in CRMs is very limited (166), several *ab initio* methods have been developed to identify composite DNA elements given just a set of input sequences (26,28,98,105,172–174). A few of these methods use efficient suffix-tree structures to identify multiple or dyad patterns (28,98), and others employ Gibbs

sampling or Monte Carlo strategy to identify multiple PWMs that may represent binding sites in regulatory modules (26,105,172–174). By combining information from multiple sites, these methods have the potential to identify motifs that are too weak (i.e. information poor) to be identified individually (26).

Two examples of the application of computational methods for identification of novel CRMs are discussed. The first describes the utility of an *ab initio* motif finder in identifying a novel composite motif regulating cell-cycle genes in yeast, and the second describes how the information on TFs that are known to bind DNA jointly can be leveraged to make genome-wide predictions of regulatory modules involving sites for those factors. With the application of an *ab initio* composite motif discovery program, CoBind (26), Pramila *et al*. (177) identified a CRM involving the binding sites for Yox1 and Mcm1 from the promoters of a set of 28 genes upregulated in the yeast Yox1 knockout strain. Yox1, which represses the expression of genes in the M/G_1 interval of yeast cell-cycle, binds DNA in conjunction with the generic MADS family TF, Mcm1. The binding sites for both TFs were jointly identified using the software and subsequently validated through mutational and gel mobility shift studies. The second example is of the transcriptional program in *Drosophila* embryo. By using known DNA binding specificity data for five TFs, Berman *et al*. (170) identified genomic regions containing unusually high concentrations of predicted binding sites for these factors. A significant fraction of these binding site clusters overlap known CRMs. In addition, many of the remaining clusters were adjacent to genes that were expressed in a pattern characteristic of those regulated by these factors. The authors tested one of the newly identified clusters, mapping upstream of the gap gene *giant* (*gt*), and showed that it acted as an enhancer that recapitulates the posterior expression pattern of *gt*.

Although the above approaches show potential, the developments in computational identification of CRMs are recent and there is significant room for further investigations and improvement, since the arrangements of sites in the modules are complex. The number of datasets containing collections of known composite regulatory elements that exist today is very limited [a few examples are (74,164,167)], which poses an obstacle in the training and testing of general computational methods in this realm.

CONCLUDING REMARKS

Understanding the mechanisms of transcriptional regulation has been an object of extended and difficult quest in biological disciplines. Clearly, significant advances have been made over the past two and half decades not only in the representation and modeling of the DNA regulatory elements, but also methods for their identification in genomic DNA. However, our knowledge of the transcriptional regulatory elements in the genome and their contribution to gene expression in different spatial and temporal contexts is still limited. Given the complex pattern of regulatory interactions, the challenges involved in the complete elucidation of these elements in the genome are substantial.

One of the major challenges is to associate the computationally identified regulatory elements with their cognate TFs. Genome-wide analyses often identify a host of putative regulatory elements (25,132–135). In order to get an understanding of the regulatory processes it is essential to associate the regulatory proteins with these elements. There have been some investigations into this problem (119,178,179) in prokaryotes and yeast, but further developments are required.

An important utility of characterizing the *cis*-regulatory elements is their use in the computational reconstruction of gene regulatory networks. Several studies have applied the information on *cis*-regulatory elements, either in isolation or in combination with other orthogonal sources of information (e.g. microarray expression data), to construct regulatory networks (180–183). These studies demonstrate how the information on transcriptional regulatory elements can be integrated to create gene networks. However, there are significant opportunities for investigations in this area.

The *ab initio* motif discovery tools and comparative genomics approaches have made it possible to detect regulatory elements in many genomes. In addition, information that can guide the search for regulatory elements to the most relevant regions of the genome is becoming available, e.g. the accurate location of transcriptional start sites (184), DNA-ase hypersensitive sequences within nuclear chromatin that represent regulatory regions (including promoters, enhancers, silencers, locus-control regions) (43), and TF binding locations from the ChIP–chip experiments (38, 40,41). Individually, there are methodological or practical limitations in the computational and experimental approaches in elucidating the location and function of the full set of regulatory elements. For example, the computational DNA motif finding algorithms have limitations in terms of the sensitivity and specificity of signals they can detect, whereas the high-throughput TF-binding site location technologies (e.g. ChIP–chip) are currently limited in the extent of intergenic sequences they can explore (185,186). Therefore it appears that the most efficient path to elucidating the novel regulatory elements and mechanisms will lie in the judicious integration of the various methods and data (182).

Despite inherent challenges in the field, rapid progress has been made over the past few years in the computational identification of regulatory elements. Successes of the computational methods have been demonstrated through experimental validations, and efficient methods for comparative genomics and analysis of CRMs are being fruitfully utilized to elucidate complex regulatory elements in organisms ranging from the unicellular bacteria to the mammals.

ACKNOWLEDGEMENTS

We are indebted to Gary Stormo for many discussions over the years. We wish to thank Yanqing Chen, Amit Kulkarni, Haoyuan Zhu, and especially David Haynor, Tao Xie and Jeffrey Sachs for comments on the manuscript. One anonymous reviewer is thanked for helpful comments and suggestions. Funding to pay the Open Access publication charges for this article was provided by Merck & Co., Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Chiaromonte,F., Weber,R.J., Roskin,K.M., Diekhans,M., Kent,W.J. and Haussler,D. (2003) The share of human genomic DNA under selection estimated from human–mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol*., **68**, 245–254.
4. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
5. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. *C.elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C.elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
7. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
8. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol*., **193**, 723–750.
9. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol*., **188**, 415–431.
10. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res*., **13**, 2381–2390.
11. Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci*., **23**, 109–113.
12. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
13. Yuh,C.H., Bolouri,H. and Davidson,E.H. (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.
14. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol*., **5**, 201.
15. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res*., **8**, 1202–1215.
16. Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol*., **5**, 279–305.
17. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet*., **5**, 276–287.
18. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
19. Pavesi,G., Mauri,G. and Pesole,G. (2004) *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform*., **5**, 217–236.
20. Pavesi,G., Mauri,G. and Pesole,G. (2001) Methods for pattern discovery in unaligned biological sequences. *Brief Bioinform*., **2**, 417–430.
21. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet*., **26**, 225–228.
22. Gelfand,M.S., Koonin,E.V. and Mironov,A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res*., **28**, 695–705.
23. Cooper,G.M. and Sidow,A. (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev*., **13**, 604–610.
24. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet*., **27**, 167–171.
25. Kellis,M., Patterson,N., Birren,B., Berger,B. and Lander,E.S. (2004) Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol*., **11**, 319–355.
26. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
27. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
28. Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18**, S354–S363.
29. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*., **31**, 3666–3668.
30. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics*, **19**, II5–II14.
31. Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19**, i169–176.
32. Jegga,A.G., Gupta,A., Gowrisankar,S., Deshmukh,M.A., Connolly,S., Finley,K. *et al*. (2005) CisMols Analyzer: identification of compositionally similar *cis*-element clusters in ortholog conserved regions of coordinately expressed genes. *Nucleic Acids Res*., **33**, W408–W411.
33. Davidson,E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego, CA.
34. Gold,L., Brown,D., He,Y., Shtatland,T., Singer,B.S. and Wu,Y. (1997) From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. *Proc. Natl Acad. Sci. USA*, **94**, 59–64.
35. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nat. Biotechnol*., **17**, 573–577.
36. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
37. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*., **29**, 2471–2478.
38. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev*., **12**, 130–136.
39. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
40. Hanlon,S.E. and Lieb,J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev*., **14**, 697–705.
41. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
42. Rodriguez,B.A. and Huang,T.H. (2005) Tilling the chromatin landscape: emerging methods for the discovery and profiling of protein–DNA interactions. *Biochem. Cell Biol*., **83**, 525–534.
43. Crawford,G.E., Holt,I.E., Whittle,J., Webb,B.D., Tai,D., Davis,S., Margulies,E.H., Chen,Y., Bernat,J.A., Ginsburg,D. *et al*. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*., **16**, 123–131.
44. CBN, U.-I.C.o.B.N. (1970) Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Eur. J. Biochem*., **15**, 203–208.
45. Stormo,G.D. (1998) Information content and free energy in DNA–protein interactions. *J. Theor. Biol*., **195**, 135–137.
46. Fields,D.S., He,Y., Al-Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol*., **271**, 178–194.
47. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
48. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic

transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
49. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res*., **32**, D75–D77.
50. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
51. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.
52. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol*., **338**, 207–215.
53. Reinert,G., Scabath,S. and Waterman,M.S. (2000) Probabilistic and statistical properties of words. *J. Comput. Biol*., **71**, 1–48.
54. Galas,D.J., Eggert,M. and Waterman,M.S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli. *J. Mol. Biol*., **186**, 117–128.
55. Pesole,G., Prunella,N., Liuni,S., Attimonelli,M. and Saccone,C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res*., **20**, 2871–2875.
56. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol*., **281**, 827–842.
57. Tompa,M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **1999**, 262–271.
58. Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*., **30**, 5549–5560.
59. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
60. Nagarajan,N., Jones,N. and Keich,U. (2005) Computing the *P*-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21**, i311–i318.
61. Stormo,G.D. and Hartzell,G.W.,III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
62. Hertz,G.Z., Hartzell,G.W.,III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci*., **6**, 81–92.
63. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
64. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
65. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
66. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.
67. Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci*., **5**, 89–96.
68. Ashraf,S.I. and Ip,Y.T. (1998) Transcriptional control: repression by local chromatin modification. *Curr. Biol*., **8**, R683–R686.
69. Razin,A. (1998) CpG methylation, chromatin structure and gene silencing—a three-way connection. *EMBO J*., **17**, 4905–4908.
70. Farkas,G., Leibovitch,B.A. and Elgin,S.C. (2000) Chromatin organization and transcriptional control of gene expression in *Drosophila*. *Gene*, **253**, 117–136.
71. Mellor,J. (2005) The dynamics of chromatin remodeling at promoters. *Mol. Cell*, **19**, 147–157.
72. Benitez-Bellon,E., Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA. *Genome Biol*., **3**, RESEARCH0013.
73. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J. Mol. Biol*., **284**, 241–254.
74. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol*., **278**, 167–181.
75. Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*., **30**, 1255–1261.
76. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
77. King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*., **31**, e116.
78. Gershenzon,N.I., Stormo,G.D. and Ioshikhes,I.P. (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res*., **33**, 2290–2301.
79. O'Flanagan,R.A., Paillard,G., Lavery,R. and Sengupta,A.M. (2005) Non-additivity in protein–DNA binding. *Bioinformatics*, **21**, 2254–2263.
80. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res*., **30**, 4442–4451.
81. Barash,Y., Kaplan,T., Friedman,N. and Elidan,G. (2003) Modeling dependencies in protein–DNA binding sites. In *Proceedings of the Seventh International Conference in Research in Computational Molecular Biology (RECOMB)*, April, 10–13, Berlin, Germany, 28–37.
82. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein–DNA binding. *Bioinformatics*, **18**, S22–S30.
83. Liu,R., Blackwell,T.W. and States,D.J. (2001) Conformational model for binding site recognition by the *E.coli* MetJ transcription factor. *Bioinformatics*, **17**, 622–633.
84. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res*., **33**, 5781–5798.
85. Kaplan,T., Friedman,N. and Margalit,H. (2005) *Ab initio* prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol*., **1**, e1.
86. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein–DNA binding sites. *Nucleic Acids Res*., **26**, 2306–2312.
87. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
88. Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein–DNA interactions. *J. Mol. Biol*., **344**, 59–70.
89. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
90. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*., **28**, 1808–1818.
91. Sinha,S. (2003) Discriminative motifs. *J. Comput. Biol*., **10**, 599–615.
92. Wang,G., Yu,T. and Zhang,W. (2005) WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Res*., **33**, W412–W416.
93. Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 269–278.
94. Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. *J. Comput. Biol*., **9**, 225–242.
95. Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.

96. Wang,G. and Zhang,W. (2005) An iterative learning algorithm for deciphering stegoscripts: a grammatical approach for motif discovery. *Technical Report No. 12*, Department of Computer Science and Engineering, Washington University, St Louis, MO.
97. Sagot,M.F. (1998) Spelling approximate repeated or common motifs using a suffix tree. *Lecture Notes Comput. Sci.*, **1380**, 111–127.
98. Marsan,L. and Sagot,M.F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.*, **7**, 345–362.
99. Apostolico,A., Bock,M.E., Lonardi,S. and Xu,X. (2000) Efficient detection of unusual words. *J. Comput. Biol.*, **7**, 71–94.
100. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**, S207–S214.
101. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
102. Bailey,T.L. and Elkan,C.P. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51–80.
103. Cardon,L.R. and Stormo,G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.*, **223**, 159–170.
104. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
105. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.
106. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
107. Siddharthan,R., Siggia,E. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
108. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.
109. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
110. Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.
111. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
112. GuhaThakurta,D., Palomar,L., Stormo,G.D., Tedesco,P., Johnson,T.E., Walker,D.W., Lithgow,G., Kim,S. and Link,C.D. (2002) Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res.*, **12**, 701–712.
113. GuhaThakurta,D., Schriefer,L.A., Waterston,R.H. and Stormo,G.D. (2004) Novel transcription regulatory elements in *Caenorhabditis elegans* muscle genes. *Genome Res.*, **14**, 2457–2468.
114. Guhathakurta,D., Schriefer,L.A., Hresko,M.C., Waterston,R.H. and Stormo,G.D. (2002) Identifying muscle regulatory elements and genes in the nematode *Caenorhabditis elegans*. *Pac. Symp. Biocomput.*, 425–436.
115. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
116. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
117. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
118. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
119. Zhu,Z., Pilpel,Y. and Church,G.M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.*, **318**, 71–81.
120. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
121. Hong,P., Liu,X.S., Zhou,Q., Lu,X., Liu,J.S. and Wong,W.H. (2005) A boosting approach for motif modeling using ChIP–chip data. *Bioinformatics*, **21**, 2636–2643.
122. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
123. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
124. Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L. and Jones,R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
125. Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.
126. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
127. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
128. Corcoran,D.L., Feingold,E., Dominick,J., Wright,M., Harnaha,J., Trucco,M., Giannoukakis,N. and Benos,P.V. (2005) Footer: a quantitative comparative genomics method for efficient recognition of *cis*-regulatory elements. *Genome Res.*, **15**, 840–847.
129. Boffelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K.D., Ovcharenko,I., Pachter,L. and Rubin,E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
130. Levy,S., Hennenhalli,S. and Workman,C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, **17**, 871–877.
131. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
132. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
133. Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of *cis*-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.
134. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
135. Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, **23**, 1249–1256.
136. Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R. and Chiaromonte,F. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.*, **14**, 700–707.
137. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W. and Chiaromonte,F. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.

138. Gelfand,M.S., Novichkov,P.S., Novichkova,E.S. and Mironov,A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform.*, **1**, 357–371.
139. McCue,L.A., Thompson,W., Carmack,C.S. and Lawrence,C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, **12**, 1523–1532.
140. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
141. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.
142. Zhu,J., Liu,J.S. and Lawrence,C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.
143. Frazer,K.A., Elnitski,L., Church,D.M., Dubchak,I. and Hardison,R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.*, **13**, 1–12.
144. Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.
145. Schwartz,S., Elnitski,L., Li,M., Weirauch,M., Riemer,C., Smit,A., Green,E.D., Hardison,R.C. and Miller,W. (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, **31**, 3518–3524.
146. Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.
147. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
148. Bray,N. and Pachter,L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.*, **14**, 693–699.
149. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
150. Shah,N., Couronne,O., Pennacchio,L.A., Brudno,M., Batzoglou,S., Bethel,E.W., Rubin,E.M., Hamann,B. and Dubchak,I. (2004) Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics*, **20**, 636–643.
151. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
152. Berezikov,E., Guryev,V. and Cuppen,E. (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res.*, **33**, W447–W450.
153. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
154. Aerts,S., Van Loo,P., Thijs,G., Mayer,H., de Martin,R., Moreau,Y. and De Moor,B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.*, **33**, W393–W396.
155. Margulies,E.H., Blanchette,M., Haussler,D. and Green,E.D. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res.*, **13**, 2507–2518.
156. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
157. Blanchette,M., Schwikowski,B. and 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.
158. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.
159. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
160. Pritsker,M., Liu,Y.C., Beer,M.A. and Tavazoie,S. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.*, **14**, 99–108.
161. Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterston,R.H. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.
162. Tompa,M. (2001) Identifying functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1143–1144.
163. Siggia,E.D. (2005) Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.*, **15**, 214–221.
164. Emberly,E., Rajewsky,N. and Siggia,E.D. (2003) Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics*, **4**, 57.
165. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.
166. Kel-Margoulis,O.V., Kel,A.E., Reuter,I., Deineko,I.V. and Wingender,E. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.
167. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
168. Liu,R., McEachin,R.C. and States,D.J. (2003) Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. *Genome Res.*, **13**, 654–661.
169. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19**, II16–II25.
170. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (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**, 757–762.
171. Alkema,W.B., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195–W198.
172. Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.
173. Gupta,M. and Liu,J.S. (2005) *De novo cis*-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.
174. Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.
175. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
176. Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
177. Pramila,T., Miles,S., GuhaThakurta,D., Jemiolo,D. and Breeden,L.L. (2002) Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes Dev.*, **16**, 3034–3045.
178. Birnbaum,K., Benfey,P.N. and Shasha,D.E. (2001) *cis* element/transcription factor analysis (cis/TF): a method for discovering transcription factor/*cis* element relationships. *Genome Res.*, **11**, 1567–1573.
179. Tan,K., McCue,L.A. and Stormo,G.D. (2005) Making connections between novel transcription factors and their DNA motifs. *Genome Res.*, **15**, 312–320.
180. Bolouri,H. and Davidson,E.H. (2002) Modeling DNA sequence-based *cis*-regulatory gene networks. *Dev. Biol.*, **246**, 2–13.
181. Haverty,P.M., Hansen,U. and Weng,Z. (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res.*, **32**, 179–188.
182. Blais,A. and Dynlacht,B.D. (2005) Constructing transcriptional regulatory networks. *Genes Dev.*, **19**, 1499–1511.
183. Shannon,M.F. and Rao,S. (2002) Transcription. Of chips and ChIPs. *Science*, **296**, 666–669.
184. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005)

The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
185. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K. *et al*. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.
186. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al*. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
187. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.