BIOINFORMATICS REVIEW Vol. 24 no. 11 2008, pages 1325–1331
doi:10.1093/bioinformatics/btn198

Gene expression

Eukaryotic transcription factor binding sites—modeling and integrative search methods

Sridhar Hannenhalli*

Penn Center for Bioinformatics and Department of Genetics, University of Pennsylvania, Philadelphia, USA

Received on February 11, 2008; revised and accepted on April 18, 2008

Advance Access publication April 21, 2008

Associate Editor: Jonathan Wren

ABSTRACT

A comprehensive knowledge of transcription factor binding sites (TFBS) is important for a mechanistic understanding of transcriptional regulation as well as for inferring gene regulatory networks. Because the DNA motif recognized by a transcription factor is typically short and degenerate, computational approaches for identifying binding sites based only on the sequence motif inevitably suffer from high error rates. Current state-of-the-art techniques for improving computational identification of binding sites can be broadly categorized into two classes: (1) approaches that aim to improve binding motif models by extracting maximal sequence information from experimentally determined binding sites and (2) approaches that supplement binding motif models with additional genomic or other attributes (such as evolutionary conservation). In this review we will discuss recent attempts to improve computational identification of TFBS through these two types of approaches and conclude with thoughts on future development.

Contact: sridharh@pcbi.upenn.edu

1 INTRODUCTION

A substantial portion of a cell's morphological and functional attributes is determined at the level of gene transcription. Thus, a comprehensive mechanistic understanding of transcriptional regulation is an important long-term goal. Eukaryotic protein coding genes are transcribed by RNA polymerase II, however the basal transcription is tightly regulated by complex processes involving chromatin modifying proteins, transcription factors (TF), co-factors and RNA polymerase (Wasserman and Sandelin, 2004). A critical component of transcription control relies on sequence-specific binding of multiple TF to short (~13 bps on average) DNA sites in the relative vicinity of the target gene (Kadonaga, 2004). Mutations in the transcription factor binding sites (TFBS) are known to underlie several human diseases and are also likely to underlie a substantial component of the phenotypic variability within and across species (Wray, 2007). A comprehensive knowledge of TFBS is thus critical for understanding the mechanism of transcriptional regulation, disease etiology and phenotypic variability.

*To whom correspondence should be addressed.

Genome-scale identification of TFBS involves three main steps: (1) experimentally identifying binding sites, (2) constructing a model or a motif to represent the set of binding sites for a TF and (3) searching for novel instances of binding sites using the model. Additionally, binding sites for an unknown TF can be identified computationally through *de novo* motif discovery.

1.1 Experimental identification of binding sites

A variety of experimental techniques have been used to identify specific genomic regions bound by a TF. We refer the reader to (Elnitski *et al.*, 2006) for a detailed review of these techniques. Below we provide a brief summary of the techniques. Genomic regions that are hypersensitive to the DNase I enzyme (DNase I HS regions) represent open chromatin regions likely to harbor functional TFBS. A number of experimental techniques exist for determining DNase I HS regions with varying resolution ranging from a few hundred bases to a single nucleotide. Given a DNase I HS region, several follow-up experiments can be done to define the precise boundaries of TFBS, such as DNase I protection or footprinting assays and deletion/mutation experiments—the so-called 'promoter bashing'. Although useful for the discovery of binding sites, these experiments do not identify the associated TFs. To this end, several other techniques have been extensively utilized. *In vitro* techniques include the *Electro-Mobility Shift Assay (EMSA), Systematic Evolution of Ligands by EXponential enrichment* (*SELEX*) and protein-binding DNA microarrays. The most common high-throughput technique for *in vivo* identification of binding sites for a specific TF is chromatin immunoprecipitation of bound DNA followed either by hybridization (*ChIP-chip*) or sequencing (*ChIP-seq*). A detailed review and the related references for the various experimental techniques are provided in (Elnitski *et al.*, 2006). Experimentally determined binding sites are compiled in databases such as TRANSFAC (Matys *et al.*, 2006) and JASPAR (Sandelin *et al.*, 2004). TRANSFAC is a licensed database which currently includes ~900 positional weight matrices (PWMs) constructed from published, experimentally determined binding sites for individual TFs. The individual binding sites are assigned a quality score corresponding to the strength of experimental evidence. JASPAR is a freely accessible resource which currently includes 138 non-redundant PWMs, also constructed from literature data, however based on a more stringent set of criteria. Despite the

difference in the sizes of TRANSFAC and JASPAR, various TF structural families are comparably represented in these two databases.

1.2 Binding motif representation

Given a collection of experimentally determined binding sites for a TF, a concise representation, or a model, of sites is referred to as a *motif*. For instance, consider a TF that binds to DNA sites *ACAAGATAA*, *ACAAGATGA* and *TCAAGATCA*. This collection of sites can be represented as the *consensus* motif *ACAAGATNA*, where '*N*' indicates any base. The consensus model belongs to a class of string-match-based models where instances of the motif are defined by the degree of base pair match from a seed set of strings. In contrast, now widely used, PWM model provides a probabilistic representation of binding sites. Unlike the consensus representation, a PWM captures relative preference for all four bases at each position. A detailed description and theoretical foundations of PWMs are provided in (Stormo, 2000). The PWM is an attractive model due to its simplicity and theoretical foundation. However, the PWM representation assumes independence among positions within a binding site which may not always be biologically reasonable (Bulyk *et al*., 2002; Man and Stormo, 2001).

1.3 Searching for binding sites in the genome

Given the PWM M for a TF, the next task is to search a genomic region of interest for sites that 'match' the PWM M and thus represent potential binding sites for the TF. The degree to which a DNA site S matches M can be quantified by a raw score R, that aggregates the PWM entry for the observed base at each position. There are a variety of ways in which R is used to decide if S is a potential binding site. For instance, software MatInspector, first transforms R into a percentile score $P = (R - min)/(max - min)$, where min and max are the minimum and the maximum scores that M can achieve, and then applies a user-defined threshold on P to define a match (Quandt *et al*., 1995). An alternative is to transform R into a P-value, which estimates the random expectation of observing a raw score of R or greater. The P-value can either be estimated theoretically based on an extreme value distribution, as in PATSER (ural.wustl.edu/software.html), or empirically based on genome-wide distribution of scores, as in (Levy and Hannenhalli, 2002). Regardless, the inherent problem with binding site search is that the binding motifs are short and degenerate, which leads to a high-error rate in a genome-wide scan. The term 'error' is meant to encapsulate sensitivity (or recall) as well as specificity (or precision). A variety of strategies have been used—most notably, evolutionary conservation—to reduce the false-positive rate. These strategies are the main focus of this review.

1.4 Binding motif discovery

In contrast to the term 'search', we use the term 'discovery' to refer to the situation in which a motif is unknown. For instance, given the promoter sequences of a set of co-regulated genes, the task of uncovering the binding sites corresponding to an unknown motif in these sequences is termed 'discovery'. Although outside the scope of this review, we mention the 'discovery' area of research for the sake of completeness. Numerous methods and software tools for *de novo* motif discovery have been published and evaluated (Sandve *et al*., 2007; Tompa *et al*., 2005).

Here, we will focus on approaches to improve the accuracy of binding site search. These approaches fall into two broad categories: (1) use of alternative (other than PWMs) binding motif representations; referred to as *modeling* approaches and (2) use of additional genomic and other biological information; referred to as *integrative* approaches. Figure 1 provides the outline for the review. We will conclude by discussing future directions to improve binding site identification using computational methods.

2 IMPROVED MOTIF MODELS FOR TFBS SEARCH

Here we discuss prior attempts to find an alternative to the PWM as a model for representing a binding motif. These include mixture models which capture multiple subclasses of binding sites for a single TF and other models that attempt to capture the dependence between positions within the binding site.

2.1 Motif subtypes

Human TF ELK1, as a monomer, can bind only to high-affinity binding sites because its *B box* domain interacts with its DNA binding *ETS* domain thereby interfering with its DNA binding activity. However, in an ELK1-SRF heterodimer, the *B box* domain interacts with SRF thus freeing up the *ETS* domain to bind to low-affinity sites (Treisman *et al*., 1992). In a different study, based on an evolutionary analysis of *REB1* sites in yeast, Tanay *et al*. (2004) identified two classes of Reb1 motifs (Fig. 2), a high-affinity motif occurring upstream of REB1 gene and a low-affinity motif occurring downstream of the REB1 transcription start site, forming a positive and a negative auto-regulatory loop.

ELK1 and REB1 provide examples of motif subtypes for a TF. In a genome-wide study, it was shown that clustering the binding sites for a TF and representing the binding motif by a mixture of PWMs, as opposed to a single PWM, improved binding site prediction (Hannenhalli and Wang, 2005). Based on the observation that most of the variability between subtypes is likely to be limited to a few positions within the binding site, a position-dependent mixture modeling approach was proposed yielding further improvement (Georgi and Schliep, 2006). In all known examples of motif subtypes for a TF, the two subtypes differ mainly in their binding affinity. It is not clear whether a TF can bind to qualitatively distinct binding sites, and if so, what are the genomic and cellular attributes that determine the usage of a specific motif subtype?

2.2 Inter-position dependence within a binding motif

A potential shortcoming of the PWM model is that it presumes independence among binding site positions, even though there are examples that point to the contrary—for Mnt repressor (Man and Stormo, 2001) and for EGR1 (Bulyk *et al*., 2002).

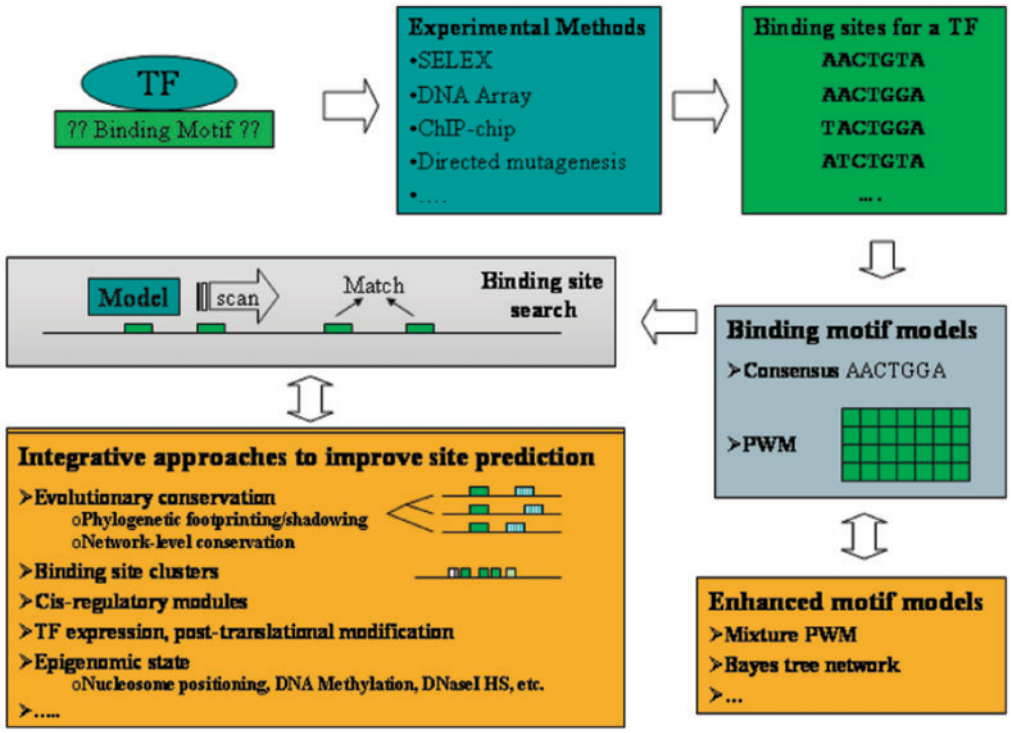

Fig. 1. Outline of the review. The overall goal is to identify transcription factor binding sites on a genome-wide scale. Starting with a few experimentally determined sites, a model of the binding site is constructed which is then used in a genome-wide scan to search for additional instances of the binding site. Besides enhanced motif models, additional, evolutionary, genomic, epigenomic, transcriptomic and proteomic data can be used in an integrative fashion to improve the accuracy of binding site search.

Fig. 2. Adopted from (Tanay *et al.*, 2004). REB1 gene promoter harbors a high-affinity binding site for Reb1 protein and mediates a positive feedback loop while a low affinity site downstream of transcription start site, active at high concentrations of Reb1, mediates a negative feedback loop. The two binding site differ in their last position.

However, in these two cases the independence assumption provides a reasonable approximation (Benos *et al.*, 2002).

2.2.1 Evolutionary evidence of inter-position dependence Based on multiple alignments of five mammalian species and a genome-wide set of high-confidence predicted binding sites in human proximal promoters, Evans *et al.* (2006) analyzed the evolutionary patterns at each binding site position and found evidence of correlated evolution. In particular, they found that the presence of a specific evolutionary pattern at a certain position affected the likelihood of the evolutionary pattern at another position. Moreover, the overall tendency for dependence between two positions decreases with increasing genomic distance between the two positions, adjacent positions being the most inter-dependent.

2.2.2 Modeling inter-position dependence Large-scale mutation experiments for determining inter-position dependence are currently not feasible. However, if computational models which explicitly incorporate inter-position dependence improve binding site prediction accuracy, then this may be perceived as indirect evidence of dependence among positions within a binding site. Based on a large set of *Escherichia coli* TFs and their binding sites, Osada *et al.* (2004) showed that incorporating dependencies among adjacent bases improves binding site prediction. Barash *et al.* (2003) proposed several models based on a *Bayesian tree network* that captures general patterns of dependencies between positions within binding sites. One of these models—a mixture of tree networks—generalizes the motif mixture model proposed in (Hannenhalli and Wang, 2005).

2.2.3 TF family motif In addition to the modeling approaches, to address redundant binding site prediction due to similarity in binding motifs among structurally related TFs, Sandelin and Wasserman (2004) have proposed a small set of family-wise PWMs. However, the initial family-wise analysis should be followed by a more refined investigation of individual motifs.

3 INTEGRATIVE APPROACHES TO IMPROVE TFBS SEARCH

We will now summarize previous attempts and future directions for improving binding site prediction accuracy by using other genomic, transcriptomic, proteomic and epigenomic markers of functional binding sites.

3.1 Evolutionary conservation—Phylogenetic footprinting and phylogenetic shadowing

Non-coding regions of the genome that are conserved across distantly related species are likely to be under purifying selection and are thus expected to harbor functional elements. *Phylogenetic footprints* are islands of highly conserved regions embedded within a background of neutrally evolving sequences (Tagle *et al*., 1988). Due to the recent availability of numerous alignable genome sequences, phylogenetic footprinting has been widely used to identify binding sites (Levy and Hannenhalli, 2002; Wasserman and Fickett, 1998; Xie *et al*., 2005). Footprinting of distantly related species, such as human and mouse, is likely to identify only ancient regulatory elements. To identify more recently created regulatory elements, a related technique—*phylogenetic shadowing*—has been proposed, which utilizes the conservation patterns in multiple closely related species (Boffelli *et al*., 2003). For a more detailed review of phylogenetic footprinting we refer the reader to (Wasserman and Sandelin, 2004). Although reliance on evolutionary conservation is an effective means to reduce the false-positive rate in binding site prediction, conservation is neither a sufficient (Nobrega *et al*., 2004), nor a necessary condition for biological functionality. Dermitzakis and Clark (2002) show that there is an extensive divergence between human and mouse within the TFBS. Similar findings have been reported in *Drosophila* (Emberly *et al*., 2003). A limitation of alignment-based approaches such as phylogenetic footprinting is that it does not allow for 'binding site turnover', the process by which a functional binding site may be lost due to the creation of an equivalent binding site nearby (Dermitzakis and Clark, 2002; Doniger and Fay, 2007; Moses *et al*., 2006). To address this limitation, a '*network level*' conservation approach has been proposed where a binding site is considered conserved as long as it appears somewhere in the orthologous promoter region (Elemento and Tavazoie, 2005; Pritsker *et al*., 2004; Kheradpour *et al*., 2007). Conserved regions may serve a variety of functional roles. To specifically characterize the conserved regions that are likely to harbor TFBS, a measure called '*regulatory potential*' has been proposed (Elnitski *et al*., 2003). Genome-wide pre-computed values of regulatory potential are provided on the UCSC genome browser (http://genome.ucsc.edu).

3.2 Clustering of binding sites

Part of the complexity of eukaryotic gene regulatory programs is achieved through combinatorial interactions among TFs. For instance, five TF proteins—Bcd, Cad, Hb, Kr and Kni—combinatorially regulate the expression of *Drosophila* genes involved in anterior-posterior axis formation (Niessing *et al*., 1997). The requirement for TF interactions is reflected in the genomic clustering of binding sites. By searching *Drosophila* promoter sequences for dense clusters of putative binding sites for Bcd, Cad, Hb, Kr and Kni, Berman *et al*. (2002) identified novel developmental genes. Since then, several approaches and tools have been proposed to detect significant binding site clusters in the genome (Rebeiz *et al*., 2002; Sinha *et al*., 2003). Although searching for clusters of binding sites for a given collection of TFs is useful, it can result in high error rates without the knowledge of functionally interacting TFs. Thus, independent approaches are needed to determine groups of potentially interacting TFs. A first step is to establish pairs of potentially interacting TFs. Kel-Margoulis *et al*. (2002) have compiled experimentally determined TF–TF interactions and their cognate motifs. In addition, several computational approaches have been proposed to predict pairs of interacting TFs. For instance, a greater than expected frequency of co-localized binding sites for two distinct TFs is indicative of their interaction (Hannenhalli and Levy, 2002; Pilpel *et al*., 2001). Also, if the genes that are putative targets of a pair of TFs have similar expression patterns across multiple spatio-temporal contexts (e.g. cell types, developmental times or environmental perturbations), then the two TFs are likely to interact functionally (Banerjee and Zhang, 2003; Pilpel *et al*., 2001). Numerous other general computational approaches for inferring interactions between proteins have been proposed which in principle, can be applied to TFs (Ramani and Marcotte, 2003; Wong *et al*., 2004).

3.3 Modeling interaction-dependent TF–DNA binding

There is evidence to suggest that the binding of a TF may depend on the presence/absence of additional binding sites (Hochschild and Ptashne, 1986; Lomvardas and Thanos, 2001). Thus the approach to identify binding sites for a TF of interest can be informed by the presence/absence of binding sites of interacting TFs. Binding models have been proposed to exploit such sequence contexts (Das *et al*., 2004; Wang *et al*., 2005).

3.4 *Cis* regulatory modules

Cooperatively interacting TFs, or transcriptional modules, often bind to a genomic region containing binding sites for the TFs in the module. A cluster of such functionally interacting sites, typically with multiple instances in the genome, is referred to as a *cis-regulatory module* (CRM) (Bolouri and Davidson, 2002; Ludwig *et al*., 1998). Similar to binding site clusters, knowledge of CRMs can aid in accurate identification of individual binding sites. For instance, to identify functional binding sites of the TF GLI, Hallikas *et al*. (2006) first identified CRMs in the mouse genome and then restricted their experimental validation to the CRMs that contained two or more GLI binding sites. Numerous computational approaches have been proposed to identify CRMs. These approaches have been discussed in a few recent review articles (Fickett and Wasserman, 2000; Hannenhalli, 2007; Wasserman and Sandelin, 2004). Below we very briefly summarize these approaches and refer the readers to previous reviews for further details.

Given a collection of interacting TFs, several methods have been proposed for genome-wide identification of the CRMs

(Berman *et al*., 2002; Sinha *et al*., 2003; Wasserman and Fickett, 1998). There are other methods that attempt to discover CRMs without a prior knowledge of the TFs. These methods can be broadly categorized as graph-theoretic, combinatorial, statistical or other heuristic approaches. The graph-theoretic approaches attempt to find tightly interconnected sub-graphs in a TF-Gene network; the sub-graph represents a potential CRM (Everett *et al*., 2006; Hannenhalli and Levy, 2003). Statistical approaches to CRM discovery include *Expectation Maximization* (EM) (Segal *et al*., 2003b) and *Gibbs sampling* (Gupta and Liu, 2005; Thompson *et al*., 2004; Zhou and Wong, 2004). Given the promoter sequence and genome-wide expression data under various conditions, Segal *et al*. (2003b) use EM to simultaneously optimize three probability models so as to best explain the observed data—gene expression model, motif model and regulation model. The CRM is implicit in the 'regulation model'. As an example of a Gibbs sampling approach, given a set of promoter sequences, Thompson *et al*. (2004) estimate three parameters using a sampling approach—PWMs for a number of unknown factors, number of sites in each promoter and the neighborhood relationship between a pair of PWMs. Numerous other approaches have been proposed (Aerts *et al*., 2003; Beer and Tavazoie, 2004; Roven and Bussemaker, 2003; Singh *et al*., 2007), and we refer the reader to other recent reviews on this topic (Fickett and Wasserman, 2000; Wasserman and Sandelin, 2004). Two recent papers provide large scale evaluations of several CRM discovery methods. Ivan *et al*. (2008) use experimentally identified CRMs in *Drosophila* to evaluated methods that do not rely on a prior knowledge of the motifs, while Klepper *et al*. (2008) use composite motifs in TRANSFAC to evaluated several methods that require a collection of motifs, however the robustness of these methods were tested by introducing additional 'noisy' motifs as the input. No single method performed consistently better.

3.5 Integrating multiple evidence to improve binding site prediction

Prediction of binding sites based on sequence data alone does not capture the spatio-temporal context of binding site usage. One relevant attribute of cellular context is the concentration of active TF. However, it is difficult to directly quantify the active form of a TF largely due to the poor characterization of active forms of TF proteins and a lack of effective assays to quantify them. Therefore, the TF gene expression level determined from microarray experiments has been used as a proxy for TF protein concentration (Chen *et al*., 2007; Roven and Bussemaker, 2003; Segal *et al*., 2003a). To improve binding site prediction, Holloway *et al*. (2005) use support vector machines (SVM) to integrate multiple evidence—binding motif, evolutionary conservation, binding site clusters, expression correlation between TF gene and target gene, functional similarity among TFs whose sites occur within a cluster, etc. Similarly, Jiang *et al*. (2007) have used SVMs to capture location and co-occurrence preferences of binding sites in order to improve binding site prediction. There are additional attributes that have a bearing on the selection of binding sites in a condition-specific fashion. Incorporating these additional attributes to identify binding sites remains an important challenge.

4 CONCLUSION AND OUTLOOK

Advances made over the last several years in computational methods and experimental techniques have significantly improved our ability to identify functional TFBS. However, on a genome scale, the accuracy of binding site prediction remains far from satisfactory. Previous approaches to improve binding site predictions have either attempted to develop enhanced, more informative motif representations or models (modeling approaches), or tried to exploit additional genomic or transcriptomic attributes (integrative approaches).

Given several experimentally determined binding sites for a TF, an ideal representation is one that strikes an optimal balance between sensitivity and specificity by extracting maximal information. While PWM representation assumes independence among positions within a binding site, a full dependence model, on the other extreme, requires estimating an exponentially large joint distribution based on a small number of exemplars. Mixture models (Barash *et al*., 2003; Hannenhalli and Wang, 2005) represent a reasonable trade-off. However, the functional relevance of multiple motif subtypes is not always clear. The optimal choice among these possibilities may vary among TFs and a detailed evaluation of these choices needs to be done. Moreover, a significant portion of known binding sites have been determined using *in vitro* approaches, such as SELEX, or DNA arrays, which may be different *in vivo* because of additional factors such as chromatin structure, epigenetic state, and the availability of other TFs. An unbiased and comprehensive evaluation of the differences between binding sites recognized *in vivo* and *in vitro* needs to be done.

While previous integrative approaches have taken additional attributes into account, except for the inclusion of mRNA levels of TF genes, these approaches remain largely sequence based. Purely sequence-based approaches for binding site identification do not capture the cellular state and thus do not reflect the dynamic nature of transcriptional regulation. A highly relevant attribute of the cellular state is the chromatin structure and epigenetic state of the genome. Recently, computational models have been proposed to predict DNase I HS regions (Noble *et al*., 2005), nucleosome positioning (Segal *et al*., 2006) and unmethylated CpG islands (Fang *et al*., 2006). Incorporating these attributes should enhance binding site prediction. For instance, using predicted nucleosome positioning as a prior results in a significant improvement in *de novo* motif discovery (Narlikar *et al*., 2007). However, the computational predictions are still based on genomic sequence and do not capture the dynamic cellular state. Histone modification state can additionally help identify the condition-specific chromatin structure. Various high-throughput technologies to assess epigenetic state are being applied to the ENCODE region and genome-wide application of these technologies in the future will significantly enhance our ability to predict binding sites (Birney *et al*., 2007). Post-translational modification states of TF proteins can alter, directly or indirectly, the TF–DNA interaction (Neumann and Naumann, 2007). However the

high-throughput technologies to identify post-translational modifications are limited to certain types of modifications and our understanding of how these modifications affect TF–DNA interaction is not sufficiently detailed. Ultimately, improvement in the prediction of TFBS will come from independent progress in these various fronts and in the development of tools to integrate the varied information, in a way that effectively incorporates heterogeneous data.

ACKNOWLEDGEMENTS

The author would like to thank Praveen Sethupathy, Jonathan Schug, Larry Singh, Laura Elnitski and the anonymous reviewers for providing constructive comments on the manuscript.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2003) Computational detection of cis-regulatory modules. *Bioinformatics*, **19** (Suppl. 2), 14.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Barash,Y. *et al.* (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular biology, Berlin, Germany*. ACM Press, New York, NY, USA.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Benos,P.V. *et al.* (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.

Berman,B.P. *et al.* (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.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Boffelli,D. *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Bolouri,H. and Davidson,E.H. (2002) Modeling DNA sequence-based cis-regulatory gene networks. *Dev. Biol.*, **246**, 2–13.

Bulyk,M.L. *et al.* (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.

Doniger,S.W. and Fay,J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.*, **3**, e99.

Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, **6**, R18.

Elnitski,L. *et al.* (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.

Elnitski,L. *et al.* (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.

Emberly,E. *et al.* (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics*, **4**, 57.

Evans,P. *et al.* (2006) Conservation patterns in cis-elements reveal compensatory mutations. *Lecture Notes Comp. Sci.*, **4205**, 186–199.

Everett,L. *et al.* (2006) Dense subgraph computation via stochastic search: application to detect transcriptional modules. *Bioinformatics*, **22**, e117–e123.

Fang,F. *et al.* (2006) Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, **22**, 2204–2209.

Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.

Georgi,B. and Schliep,A. (2006) Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, **22**, e166–e73.

Gupta,M. and Liu,J.S. (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.

Hallikas,O. *et al.* (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.

Hannenhalli,S. (2007) Eukaryotic transcriptional regulation: signals, interactions and modules. In Stojanovic,N. (ed.) *Computational Genomics*. Horizon Bioscience, Norfolk, UK, pp. 55–82.

Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.

Hannenhalli,S. and Levy,S. (2003) Transcriptional regulation of protein complexes and biological pathways. *Mamm. Genome*, **14**, 611–619.

Hannenhalli,S. and Wang,L.S. (2005) Enhanced position weight matrices using mixture models. *Bioinformatics*, **21** (Suppl. 1), i204–i212.

Hochschild,A. and Ptashne,M. (1986) Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. *Cell*, **44**, 681–687.

Holloway,D.T. *et al.* (2005) Integrating genomic data to predict transcription factor binding. *Genome Inform.*, **16**, 83–94.

Ivan,A. *et al.* (2008) Computational discovery of cis-regulatory modules in Drosophila, without prior knowledge of motifs. *Genome Biol.*, **9**, R22.

Jiang,B. *et al.* (2007) OSCAR: one-class SVM for accurate recognition of cis-elements. *Bioinformatics*, **23**, 2823–2828.

Kadonaga,J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.

Kel-Margoulis,O.V. *et al.* (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.

Kheradpour,P. *et al.* (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.*, **17**, 1919–1931.

Klepper,K. *et al.* (2008) Assessment of composite motif discovery methods. *BMC Bioinformatics*, **9**, 123.

Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.

Lomvardas,S. and Thanos,D (2001) Nucleosome sliding via TBP DNA binding in vivo. *Cell*, **106**, 685–696.

Ludwig,M.Z. *et al.* (1998) Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, **125**, 949–958.

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.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Moses,A.M. *et al.* (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, **2**, e130.

Narlikar,L. *et al.* (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.

Neumann,M. and Naumann,M. (2007) Beyond IkappaBs: alternative regulation of NF-kappaB activity. *Faseb J.*, **21**, 2642–2654.

Niessing,D. *et al.* (1997) A cascade of transcriptional control leading to axis determination in Drosophila. *J. Cell Physiol.*, **173**, 162–167.

Noble,W.S. *et al.* (2005) Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics*, **21** (Suppl. 1), i338–i343.

Nobrega,M.A. *et al.* (2004) Megabase deletions of gene deserts result in viable mice. *Nature*, **431**, 988–993.

Osada,R. *et al.* (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, **20**, 3516–3525.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Pritsker,M. *et al.* (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.*, **14**, 99–108.

Quandt,K. *et al.* (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Ramani,A.K. and Marcotte,E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.

Rebeiz,M. *et al.* (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci. USA*, **99**, 9888–9893.

Roven,C. and Bussemaker,H.J. (2003) REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res.*, **31**, 3487–3490.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

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.

Sandve,G.K. *et al.* (2007) Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, **8**, 193.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Segal,E. *et al.* (2003a) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Segal,E. *et al.* (2003b) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19** (Suppl. 1), i273–i82.

Singh,L.N. *et al.* (2007) TREMOR–a tool for retrieving transcriptional modules by incorporating motif covariance. *Nucleic Acids Res.*, **35**, 7360–7371.

Sinha,S. *et al.* (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19** (Suppl. 1), I292–I301.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tagle,D.A. *et al.* (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.

Tanay,A. *et al.* (2004) A global view of the selection forces in the evolution of yeast cis-regulation. *Genome Res.*, **14**, 829–834.

Thompson,W. *et al.* (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–74.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Treisman,R. *et al.* (1992) Spatial flexibility in ternary complexes between SRF and its accessory proteins. *Embo J.*, **11**, 4631–4640.

Wang,L.S. *et al.* (2005) An interaction-dependent model for transcription factor binding. *Lecture Notes Comp. Sci.*, **4023**, 225–234.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Wong,S.L. *et al.* (2004) Combining biological networks to predict genetic interactions. *Proc. Natl Acad. Sci USA*, **101**, 15682–15687.

Wray,G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, **8**, 206–216.

Xie,X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

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.