Review

Computational prediction of transcription-factor binding site locations

Martha L Bulyk

Address: Division of Genetics, Departments of Medicine, Pathology and Harvard/MIT Division of Health Sciences and Technology, Brigham and Women's Hospital and Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.
E-mail: mlbulyk@rascal.med.harvard.edu

Published: 23 December 2003

Genome ***Biology*** 2003, **5:**201

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/5/1/201

Abstract

Identifying genomic locations of transcription-factor binding sites, particularly in higher eukaryotic genomes, has been an enormous challenge. Various experimental and computational approaches have been used to detect these sites; methods involving computational comparisons of related genomes have been particularly successful.

The publication of a nearly complete draft sequence of the human genome is an enormous achievement, but characterizing the entire set of functional elements encoded in the human and other genomes remains an immense challenge [1]. Francis Collins, Director of the National Human Genome Research Institute (USA), has proposed that "the next phase of genomics is to catalog, characterize and comprehend the entire set of functional elements [including those that do not encode protein] encoded in the human and other genomes" [1]. Two of the most important functional elements in any genome are transcription factors (TFs) and the sites within the DNA to which they bind. These interactions between protein and DNA control many important processes, such as critical steps in development and responses to environmental stresses, and defects in them can contribute to the progression of various diseases. Much progress has been made recently in the accumulation and analysis of mRNA transcript profiles of a variety of cell and tissue types, including those associated with various human diseases [2]; much remains to be understood, however, about the transcriptional regulatory networks that govern these expression profiles. A more complete understanding of transcription factors, their DNA binding sites, and their interactions, will permit a more comprehensive and quantitative mapping of the regulatory pathways within cells, as well as a deeper understanding of the potential functions of individual genes regulated by newly identified DNA-binding sites.

The binding specificities of only a small number of TFs are well characterized. Transcription-factor binding sites (TFBSs) are usually short (around 5-15 base-pairs (bp)) and they are frequently degenerate sequence motifs (Figure 1a); potential binding sites thus can occur very frequently in larger genomes such as the human genome. The sequence degeneracy of TFBSs has been selected through evolution and is beneficial, because it confers different levels of activity upon different promoters, thus causing some genes to be transcribed at higher levels than others, as may be required by the cell [3]. The function of TFBSs is often independent of their orientation. In yeast, their position within a promoter can vary, and in higher eukaryotes they can occur upstream, downstream, or in the introns of the genes that they regulate; in addition, they can be close to or far away from regulated gene(s). Moreover, the human genome is about 200 times larger than yeast genome, and approximately 95-99% of it does not encode proteins. For all these reasons, it can be very difficult to find TFBSs in noncoding sequences using relatively simple sequence-searching tools like BLASTN or CLUSTALW [4].

Experimental methods for identifying transcription-factor binding sites

Much of the information on TF binding specificity has been determined using traditional methodologies such as

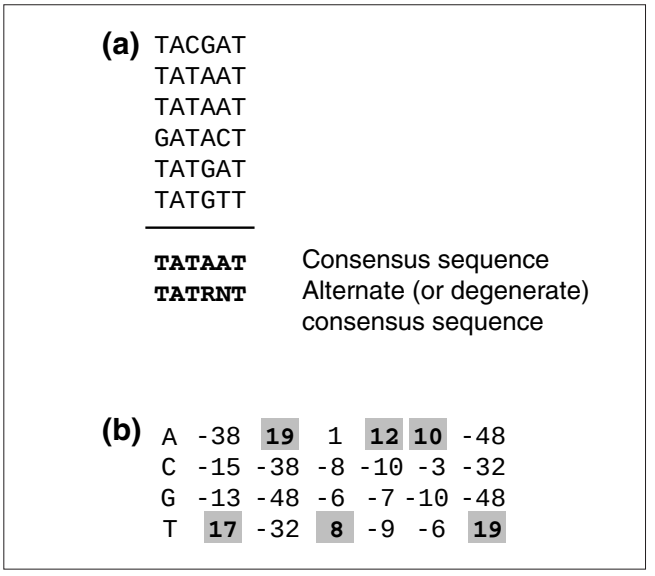

Figure 1

Representation of transcription-factor binding sites. **(a)** An example of six sequences and the consensus sequence that can be derived from them. The consensus simply gives the nucleotide that is found most often in each position; the alternate (or degenerate) consensus sequence gives the possible nucleotides in each position; R represents A or G; N represents any nucleotide. **(b)** A position weight matrix for the -10 region of *E. coli* promoters, as an example of a well-studied regulatory element. The boxed elements correspond to the consensus sequence (TATAAT). The score for each nucleotide at each position is derived from the observed frequency of that nucleotide at the corresponding position in the input set of promoters. The score for any particular site is the sum of the individual matrix values for that site's sequence; for example, the score for TATAAT is 85. Note that the matrix values in (b) do not come from the example shown in (a) but rather are derived from a much larger collection of -10 promoter regions. Adapted, with permission, from [3].

footprinting methods that identify the region of DNA protected by a bound protein, nitrocellulose binding assays, gel-shift analysis that monitors the change in mobility when DNA and protein bind, Southwestern blotting of both DNA and protein, or reporter constructs. These methods are generally quite time-consuming and not readily scaled up to whole genomes, however. In recent years, therefore, a number of high-throughput technologies have been developed, for identifying TFBSs both *in vitro* and *in vivo*. One high-throughput method for finding high-affinity binding sequences *in vitro* is the selection (frequently referred to as SELEX (systematic evolution of ligands by exponential evolution)) from randomized double-stranded DNAs those that bind with high affinity to a protein of interest [5]. This method has been further modified into genomic SELEX, which uses a genomic library as the starting material for the selections [6]. More recently, the sequence specificities of DNA-binding proteins have been determined by direct binding of proteins to double-stranded DNA microarrays [7,8].

Similarly, high-throughput methods have also been developed for measuring the interactions between DNA and TFs *in vivo*. Microarray-based readout of chromatin immunoprecipitation assays ('ChIP-chip'), also referred to as genome-wide location analysis [9], is currently the most widely used method for identifying genomic TFBSs *in vivo* and in a high-throughput manner (see [10] for a review). This approach has been used to characterize a number of TFs in the yeast *Saccharomyces cerevisiae* [9,11-15] and, more recently, to identify genomic targets in mammalian cells [16-18]. Another recently developed method that takes advantage of DNA microarrays for the identification of TFBSs *in vivo* uses TFs tethered to DNA adenine methyltransferase (Dam) [19,20], resulting in DNA methylation near sites bound by the TF-Dam fusion protein [19,20]. This approach has been used to identify binding sites *in vivo* in *Drosophila* [20,21] and *Arabidopsis* [22].

Identifying candidate TFBSs *in silico*

Once a potential regulatory sequence motif has been identified, the next goal is frequently to identify candidate target genes that may be regulated through it, potentially by a TF that may bind to it. Although degenerate consensus sequences (Figure 1a) are still frequently used to depict the binding specificities of TFs, they do not contain precise information about the relative likelihood of observing the alternate nucleotides at the various positions of a TFBS. Thus, a common way of representing the degenerate sequence preferences of a DNA-binding protein is by a position weight matrix (PWM; Figure 1b) [3]. It is important to note that predicted TFBSs may not serve a direct regulatory function, or even be bound, *in vivo*. A number of collections of experimentally defined TFBSs have therefore been assembled. Genes can also be classified according to whether they are likely to be regulated through a particular motif or combination of motifs, such as by using Hidden Markov Models [33] to statistically model the number and relative locations of TFBSs within a sequence [34].

The prediction and experimental identification of regulatory regions in higher eukaryotes is more difficult than in model organisms with smaller genomes. Not only are the genomes larger, but also a greater proportion of the genomes are noncoding. In addition, regulatory elements can be found far from the transcription start site of the genes they regulate, making the search for them difficult. One method to enrich for shared sets of candidate regulatory elements is to focus on the noncoding sequence surrounding genes with very similar mRNA expression patterns. A number of studies have been successful in extracting sequence motifs from expression data or groups of functionally related genes in yeast [35-39]. This is much more difficult for higher eukaryotes, however, because the much greater amount of input sequence that must go into the motif-search algorithms increases the background noise levels in the motif search. For these reasons, it has been suggested that comparisons between genomes be incorporated when searching higher

eukaryotic expression clusters for regulatory motifs [40]. Further details of PWMs and collections of experimentally defined TFBSs are available with the complete version of this article, online.

Phylogenetic footprinting

A major method for enriching for candidate regulatory elements is to identify regions of sequence conservation between genomes, as it is these conserved regions that are likely to contain important regulatory sites. This method of performing phylogenetic comparisons to reveal conserved *cis* elements in the noncoding regions of homologous genes is referred to as 'phylogenetic footprinting' [41]. It has been described as searching for "islands of conserved sequences in seas of less conserved noncoding sequence" [40].

An important first step in phylogenetic footprinting is to identify orthologs, genes in different species that are derived from the same gene in the last common ancestral species and thus usually have similar functions in the genomes being compared. In contrast, paralogs are duplicate gene pairs within a genome that have diverged and typically have different functions. Orthologs need to be distinguished from paralogs, because it can be expected that as the functions of paralog has diverged, their transcriptional regulators may also have diverged. At relatively close evolutionary distances - divergence around 40-80 million years ago (Mya) - it can be difficult to distinguish between undiscovered coding sequences and functional noncoding sequences, so comparison with distantly related species can improve the ability to distinguish these classes of conserved sequences [42]. Frazer and colleagues [42,43] have reviewed methods for cross-species sequence comparisons.

In the initial sequencing and comparative analysis of the mouse genome, Waterston and colleagues [44] found that a much higher fraction of short segments in the mammalian genome are under selection than can be explained by protein-coding sequences alone. In a comparison of 1 megabase (Mb) of orthologous human and mouse sequences surrounding three interleukin genes [45], 90 conserved noncoding sequences were found, of which one has so far been shown experimentally to regulate the genes. Comparisons have also been made between the pufferfish and human genomes [49]; it is important to remember that comparisons between such distantly related organisms will miss sequences specific to one lineage.

The first step in finding upstream TFBSs is to identify the transcriptional start site, so that searches can be focused on sequence upstream of 5′ untranslated regions; intronic and transcriptionally downstream sequences are also searched for TFBSs. Next, either global or local sequence alignments are performed to identify regions of sequence conservation. It is important to note that the level of sequence conservation varies considerably across genomes, so fixed percentage-identity cutoffs for alignments may not be suitable. TFs associated with expression specific to skeletal muscle have been studied extensively, probably as a result of good cell-culture models for differentiation. Wasserman and Fickett [66] have done a literature search for experimentally defined TFBSs for five TFs associated with skeletal-muscle-specific expression and found that high-scoring sites occurred more frequently in sequences linked to muscle-specific expression. In a comparison of 28 orthologous human-mouse gene pairs that are specifically upregulated in skeletal muscle, Wasserman's group [68] found that 98% of experimentally defined sequence-specific TFBSs specific to skeletal muscle are confined to the 19% of human noncoding sequences that are most conserved in the orthologous rodent sequences [68]. Further details of the factors that need to be considered in phylogenetic footprinting studies and the results from these analyses are available with the complete version of this article, online.

Clustering of transcription-factor binding sites

In higher eukaryotes, TFs frequently bind DNA within segments of sequence, typically hundreds of base-pairs long, termed *cis*-regulatory modules or enhancers. A given gene can have multiple such modules in its surrounding noncoding sequence; they typically direct expression in either a cell-type-specific or temporal-specific manner [69]. Typically four to eight different TFs bind within an enhancer, and each factor can bind to multiple sites within it [53,70] (for reviews on transcriptional regulation in metazoans, see [69,70]). Because pairs of sites may correspond to TFs that coregulate expression of the nearby gene(s) [71], a number of approaches have been developed to identify pairs of binding sites [72-78]. For example, one study focusing on the MEF2 and MyoD families of TF found that where the two bind in the same regulatory region, their binding sites occur at precise distances relative to the helical turn of DNA, and thus probably allow cooperative protein-protein interactions [79]. Although some TFs may require specific distances between their binding sites for cooperative binding, it has been thought that in many cases the exact spacing and order of TFBSs is not important for enhancer function [80].

More recently, approaches have been developed to identify higher-order site clusterings [81-93]. Such clusters can be homotypic, containing multiple sites for one particular TF, or heterotypic, containing one or more binding sites for multiple TFs [89]. A search of vertebrate genomic sequence revealed that sites bound by the liver regulatory TF hepatocyte nuclear factor 1 (HNF1) occurred more frequently in hepatic genes than expected by chance, that HNF1-binding sites in liver genes are more often associated in clusters with sites for other TFs than expected by chance, and that the enrichment is more pronounced in promoter regions [94]. In a search for matches to TRANSFAC PWMs within conserved

noncoding sequences surrounding a set of human and mouse genes, conserved segments in upstream regions contained TFBS pairs colocalized in a manner consistent with experimentally known pairwise co-occurrences of TFs [95].

In a recently published study, Wasserman and colleagues [96] performed human-mouse sequence comparisons of 14 well-studied genes and searched for matches to TFBS PWMs within the conserved noncoding regions, using a range of PWM score thresholds. The choice of PWM score cutoffs is a critical issue in all predictions of sites from PWMs, as the requirement for a more stringent match (a higher cutoff) is likely to result in fewer false-positive predictions but can potentially result in more sites being missed (false negatives). The same kind of problem occurs when conserved regions are used: the assumption is that fewer of the motif 'hits' will be false positives than when searching the whole genome, but a greater number of functional sites may be missed because they occur outside conserved regions. Considering regions with 70% sequence identity and a 75% relative matrix score threshold, Wasserman and colleagues found that 66% of previously verified TFBSs were detected with phylogenetic footprinting, compared with 73% when just single sequences were scanned. At a 60% matrix score threshold, looking just within the conserved regions, they were able to detect 83% of TFBSs [96] (although one has to keep in mind that decreasing the PWM score threshold will increase the number of likely false-positive hits).

Full-genome comparisons of yeast noncoding sequences

The yeasts are good organisms for phylogenetic footprinting because of the compact noncoding portions of yeast genomes, the available complete *S. cerevisiae* sequence, the well-characterized phylogeny, and the ease of experimental validation in *S. cerevisiae*. Yeast strains closely related to *S. cerevisiae* can be divided into three sub-groups: *Saccharomyces sensu stricto*, *Saccharomyces sensu lato* and petite-negative. A survey by Mark Johnston's group [4] of orthologous genomic loci in seven yeast strains from these sub-groups showed conservation of TFBSs and their spacing in *sensu stricto* species but not *sensu lato* species. Subsequently, the same group [97] sequenced the genomes of three *sensu stricto* strains (*S. mikatae, S. kudriavzevii,* and *S. bayanus*) and two more distantly related strains (*S. castellii* and *S. kluyveri*), and performed genome sequence alignments. They identified most characterized motifs that met their stringent criteria, and also 79 unique unknown conserved elements of length 6-30 bp with no gaps, with some evidence for functionality.

In a similar study using slightly different criteria, Lander and colleagues [99] compared four *sensu stricto* species - *S. cerevisiae, S. paradoxus*, *S. mikatae*, and *S. bayanus* - and focused on Gal4-binding sites as a test case (Figure 2). They found 72 motifs, both known and novel, and could assign a tentative biological function to many novel motifs. Most were usually upstream of genes, although some were preferentially found downstream of genes (note that many studies that aim to find regulatory DNA elements in yeast have searched only upstream of the target gene(s)). Even in these high-resolution genome sequence comparisons, not all known motifs were found. Further discussion of these yeast studies is available with the complete version of this article, online.

Phylogenetic footprinting in other organisms

Similar phylogenetic footprinting approaches have been taken to try to identify regulatory elements in the noncoding portions of other genomes. A comparison of the *Escherichia coli* and *Haemophilus influenzae* genomes led to the identification of a novel motif that had not been found previously in any of the individual genomes, and to the discovery of new members of known regulons [100]. In a search within alignments of a set of orthologous intergenic regions from the *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes (which are 23-40 Mya apart), an uneven distribution of short conserved sequence blocks was found across the genomes, again suggesting the potential co-occurrence of TFBSs within transcriptional enhancers [101]. In an analysis of conservation over four *Drosophila* species spanning a range of divergence times, it was also found that conserved noncoding sequences tend to cluster spatially, with conserved spacing between them, and that there is a strong tendency for known *cis*-regulatory elements to overlap clusters of conserved noncoding sequences [102]. Such clusters may correspond to functional interactions among transcriptional enhancers.

In a landmark paper examining enhancer function in *Drosophila*, Ludwig and co-workers [103] found that in a comparison of 13 species, none of 16 surveyed *D. melanogaster* TFBSs was completely conserved. They also observed differences in the spacing between TFBSs. Despite these differences between species, each enhancer drove reporter-gene expression at identical times and locations in the early *D. melanogaster* embryo. Chimeric enhancers did not recapitulate the wild-type expression pattern, however. The authors proposed that stabilizing selection has maintained phenotypic constancy, but has allowed mutation within the enhancer, and that substitutions within TFBSs and changes in the lengths of spacer regions between TFBSs would result in weak changes, with many functionally compensatory mutations. One of their significant conclusions was that this "may make it difficult to identify homologous elements in different species groups by sequence comparison alone" [103]. This is an important observation to keep in mind in the development and application of algorithms for discovery *in silico* of transcriptional enhancers and TFBSs conserved across genomes, because conserved TFBSs may not necessarily occur within longer stretches of conserved sequence.

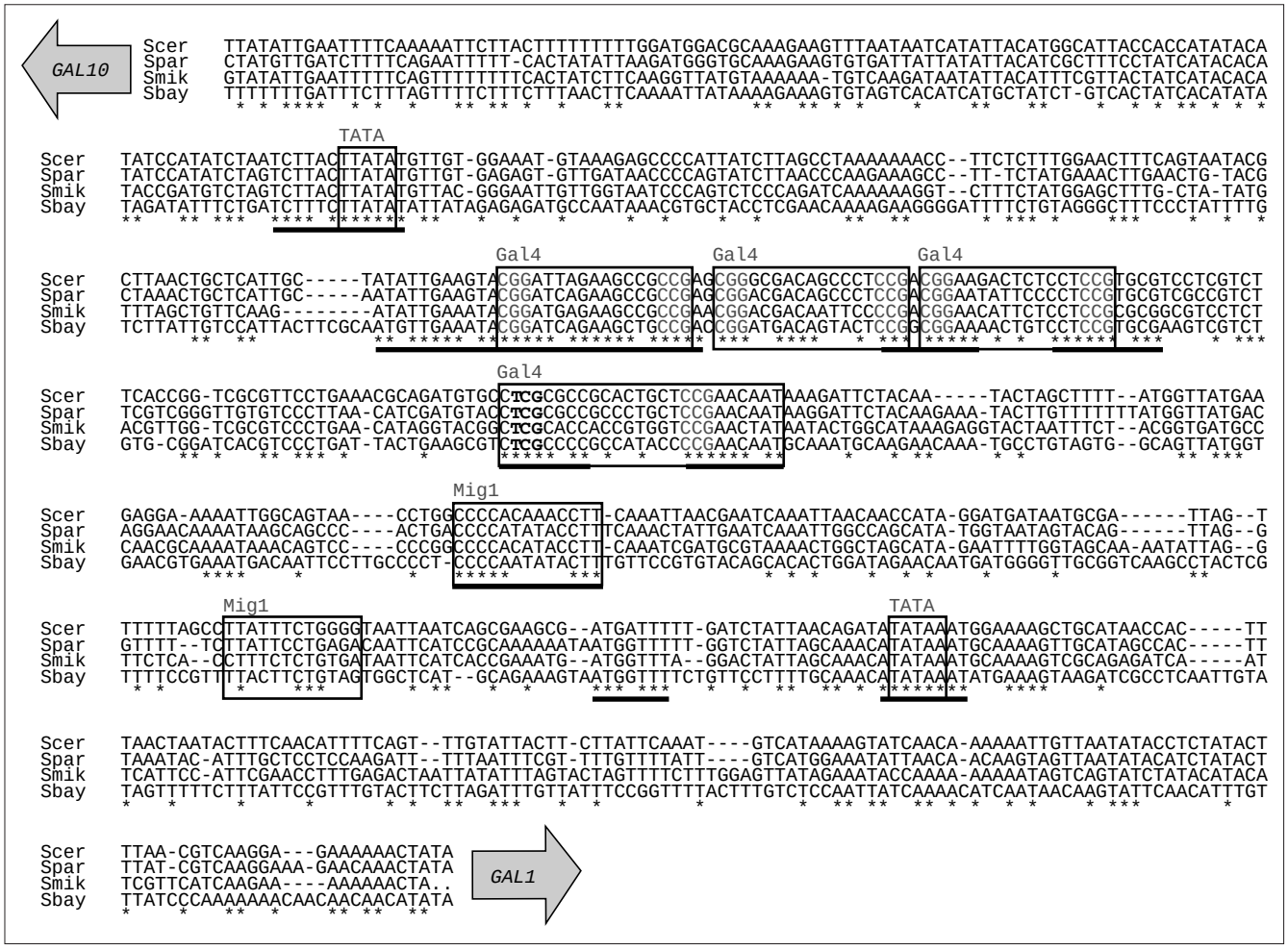

Figure 2

Sequence comparison of the *GAL1-GAL10* intergenic region across four yeast species. Scer, *S. cerevisiae*; Spar, *S. paradoxus*; Smik, *S. mikatae*; Sbay, *S. bayanus*. Arrows indicate the start and transcriptional orientation of the *GAL1* and *GAL10* open reading frames; dashes in the alignment indicate gaps; nucleotide positions conserved across all four species are denoted by asterisks. Stretches of conserved nucleotides are underlined, and experimentally validated transcription-factor binding-site footprints are boxed and labeled with the name of the footprinted transcription factor. Underlined regions that are not boxed correspond to potential, previously unknown, transcription-factor binding sites. Note that not all nucleotide positions of a footprinted binding site are necessarily conserved across all four species in this comparison (note the Mig1 sites, for example). The nucleotides matching the published Gal4 binding-site motif are in gray; for the fourth Gal4 site, non-standard consensus motif nucleotides are shown in boldface. Reproduced with permission from [99].

In an important recent study, Boffelli and colleagues [104] sequenced four different regions from over a dozen primate species, including Old World and New World monkeys and hominoids. The premise of their approach was that the human-mouse comparisons can fail to align meaningfully, and thus can fail to identify functional elements, and that the additive collective divergence of higher primates as a group is comparable to that of humans and mice [104]. An additional consideration is that in comparing just human and mouse sequences there is the potential problem that some regions of the genome are highly conserved [105]. In this 'phylogenetic shadowing' approach, they took into account the phylogenetic relationships of the analyzed species. The authors noted that the most informative subset of four to seven species can capture most of the discriminative power of the approach using the full set of species. Using gel-shift assays and luciferase reporter assays, they found that conserved regions were bound by protein more frequently, and thus were presumably more likely to be functional, than nonconserved regions [104].

In a similar study, Thomas and colleagues [106] compared sequences from 12 evolutionarily diverse vertebrate species, for sequences orthologous to a human chromosomal region containing 10 genes, including the gene mutated in cystic fibrosis (CFTR). The authors noted that the 'multi-species conserved regions' that they detected overlapped with 63% of the functionally validated regulatory elements in the CFTR

genomic region, and that many of the remaining missed known regulatory elements may have been missed either because they are shorter than their approach could detect (< 25 bp), or because they are primate-specific. Interestingly, their results suggest that the power to detect multi-species conserved regions seems to depend mainly on the total divergence of the subset of species rather than on the particular distribution of the species among lineages, and thus that combined phylogenetic branch length may be a useful metric for guiding the selection of additional genomes to sequence.

Future directions in the discovery of transcription-factor binding sites

Francis Collins has said [1] that further multi-species comparisons, especially those occupying distinct evolutionary positions, will lead to significant refinements in our understanding of the functional importance of conserved sequences and are thus crucial to the functional characterization of the human genome. Sidow [107] noted that identification of the majority of functional elements relevant to human biology requires placental genomes beyond those of human, mouse, and rat. Sidow commented that "Building a parts list is important, but multiple sequence alignments by themselves do not quantify conservation and allow only limited inference as to which conserved functional element is more constrained than another" [107].

In recent years, a number of efforts have been focused on attempting to predict TFBSs using structural information on the TF protein itself or related protein-DNA complexes. Some of these studies have attempted to determine what 'recognition rules' or 'recognition code' may exist that stipulate which DNA base-pairs are likely to be bound by which amino acids, in the context of a particular structural class of DNA-binding protein. There is no obvious, simple code like the genetic code, and any recognition rules that might exist are likely to be quite degenerate and highly dependent upon the docking arrangement of the protein with its DNA binding site [118,119]. An important challenge will be to characterize the binding specificities of the approximately 1,850 TFs in the human genome [120]. The high-throughput technologies described earlier will help with these studies. Further details of these future developments are available with the complete version of this article, online. Finally, there is a need for the development of high-throughput transgenic bioassays for validating predicted enhancers, so that we can be sure that *in silico* predictions translate to *in vivo* understanding.

→ .reviews

The complete version of this article, available online at http://genomebiology.com/2003/5/1/201, includes the following additional information:

Further details of position weight matrices and collections of experimentally defined TFBSs.

Further details of the factors that need to be considered in phylogenetic footprinting studies, and the results from these analyses.

Further discussion of full-genome comparisons of yeast noncoding sequences.

Further details of likely future developments and challenges in finding TFBSs.

Additional references.

Acknowledgements

I thank Mike Berger, Anthony Philippakis, and Pete Estep for helpful comments on the manuscript. M.L.B. was supported in part by an Informatics Research Starter Grant from the PhRMA Foundation, a Taplin Award from the John F. and Virginia B. Taplin Foundation, and a Harvard Medical School William F. Milton Fund Award.

References

1. Collins F, Green E, Guttmacher A, Guyer M, US National Human Genome Institute: **A vision for the future of genomics research.** *Nature* 2003, **422:**835-847.
2. Lockhart D, Winzeler E: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405:**827-836.
3. Stormo G: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
4. Cliften P, Hillier L, Fulton L, Graves T, Miner T, Gish W, Waterston R, Johnston M: **Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1175-1186.
5. Oliphant A, Brandl C, Struhl K: **Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein.** *Mol Cell Biol* 1989, **9:**2944-2949.
6. Gold L, Brown D, He Y-Y, Shtatland T, Singer B, Wu Y: **From oligonucleotide shapes to genomic SELEX: Novel biological regulatory loops.** *Proc Natl Acad Sci USA* 1997, **94:**59-64.
7. Bulyk ML, Huang X, Choo Y, Church GM: **Exploring the DNA-binding specificities of zinc fingers with DNA microarrays.** *Proc Natl Acad Sci USA* 2001, **98:**7158-7163.
8. Bulyk ML, Gentalen E, Lockhart DJ, Church GM: **Quantifying DNA-protein interactions by double-stranded DNA arrays.** *Nat Biotechnol* 1999, **17:**573-577.
9. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
10. Wyrick J, Young R: **Deciphering gene expression regulatory networks.** *Curr Opin Genet Dev* 2002, **12:**130-136.
11. Reid JL, Iyer VR, Brown PO, Struhl K: **Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase.** *Mol Cell* 2000, **6:**1297-1307.
12. Lieb JD, Liu X, Botstein D, Brown PO: **Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association.** *Nat Genet* 2001, **28:**327-334.
13. Lee T, Rinaldi N, Robert R, Odom D, Bar-Joseph Z, Gerber G, Hannett N, Harbison C, Thompson C, Simon I, *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.

14. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
15. Simon I, Barnett J, Hannett N, Harbison C, Rinaldi N, Volkert T, Wyrick J, Zeitlinger J, Gifford D, Jaakkola T, *et al.*: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106:**697-708.
16. Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M: **GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIP-chip analysis.** *Proc Natl Acad Sci USA* 2002, **99:**2924-2929.
17. Weinmann A, Yan P, Oberley M, Huang T, Farnham P: **Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis.** *Genes Dev* 2002, **16:**235-244.
18. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young R, Dynlacht B: **E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints.** *Genes Dev* 2002, **16:**245-256.
19. van Steensel B, Henikoff S: **Identification of *in vivo* DNA targets of chromatin proteins using tethered dam methyltransferase.** *Nat Biotechnol* 2000, **18:**424-428.
20. van Steensel B, Delrow J, Henikoff S: **Chromatin profiling using targeted DNA adenine methyltransferase.** *Nat Genet* 2001, **27:**304-308.
21. van Steensel B, Delrow J, Bussemaker H: **Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding.** *Proc Natl Acad Sci USA* 2003, **100:**2580-2585.
22. Tompa R, McCallum C, Delrow J, Henikoff J, van Steensel B, Henikoff S: **Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3.** *Curr Biol* 2002, **12:**65-68.
33. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press; 1998.
34. Pavlidis P, Furey T, Liberto M, Haussler D, Grundy W: **Promoter region-based classification of genes.** *Pac Symp Biocomput* 2001:151-163.
35. Tavazoie S, Hughes J, Campbell M, Cho R, Church G: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
36. Hughes J, Estep P, Tavazoie S, Church G: **Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214.
37. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16:**939-945.
38. Bussemaker H, Li H, Siggia E: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
39. Chiang D, Brown P, Eisen M: **Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles.** *Bioinformatics* 2001, **17 Suppl 1:**S49-S55.
40. Pennacchio L, Rubin E: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2:**100-109.
41. Tagle D, Koop B, Goodman M, Slightom J, Hess D, Jones R: **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* 1988, **203:**439-455.
42. Frazer K, Elnitski L, Church D, Dubchak I, Hardison R: **Cross-species sequence comparisons: a review of methods and available resources.** *Genome Res* 2003, **13:**1-12.
43. Dubchak I, Frazer K: **Multi-species sequence comparison: the next frontier in genome annotation.** *Genome Biol* 2003, **4:**122.
44. Waterston R, Lindblad-Toh K, Birney E, Rogers J, Abril J, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
45. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
49. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S: **Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*.** *Proc Natl Acad Sci USA* 1995, **92:**1684-1688.
53. Davidson E: *Genomic Regulatory Systems: Development and Evolution*. San Diego: Academic Press; 2001.
66. Wasserman W, Fickett J: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-181.
68. Wasserman W, Palumbo M, Thompson W, Fickett J, Lawrence C: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.
69. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424:**147-151.
70. Arnone M, Davidson E: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
71. Pilpel Y, Sudarsanam P, Church G: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
72. GuhaThakurta D, Stormo G: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17:**608-621.
73. Gelfand M, Koonin E, Mironov A: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28:**695-705.
74. Li H, Rhodius V, Gross C, Siggia E: **Identification of the binding sites of regulatory proteins in bacterial genomes.** *Proc Natl Acad Sci USA* 2002, **99:**11772-11777.
75. van Helden J, Rios A, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28:**1808-1818.
76. Eskin E, Pevzner P: **Finding composite regulatory patterns in DNA sequences.** *Bioinformatics* 2002, **18 Suppl 1:**S354-S363.
77. Quandt K, Grote K, Werner T: **GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences.** *Genomics* 1996, **33:**301-304.
78. Bulyk ML, McGuire AM, Masuda N, Church GM: **A motif co-occurrence approach for genome-wide prediction of transcription factor binding sites in *E. coli*.** *Genome Res* 2004, in press.
79. Fickett J: **Coordinate positioning of MEF2 and myogenin binding sites.** *Gene* 1996, **172:**GC19-GC32.
80. Wagner A: **Distribution of transcription factor binding sites in the yeast genome suggests abundance of coordinately regulated genes.** *Genomics* 1998, **50:**293-295.
81. Markstein M, Markstein P, Markstein V, Levine M: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo.** *Proc Natl Acad Sci USA* 2002, **99:**763-768.
82. Halfon M, Grad Y, Church G, Michelson A: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-1028.
83. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762.
84. Rajewsky N, Vergassola M, Gaul U, Siggia E: **Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo.** *BMC Bioinformatics* 2002, **3:**30.
85. Krivan W, Wasserman W: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:**1559-1566.
86. Frith M, Hansen U, Weng Z: **Detection of *cis*-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17:**878-889.
87. Frith M, Spouge J, Hansen U, Weng Z: **Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences.** *Nucleic Acids Res* 2002, **30:**3214-3224.
88. Frith M, Li M, Weng Z: **Cluster-Buster: finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3666-3668.
89. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15:**776-784.
90. Markstein M, Levine M: **Decoding *cis*-regulatory DNAs in the *Drosophila* genome.** *Curr Opin Genet Dev* 2002, **12:**601-606.

91. Pickert L, Reuter I, Klawonn F, Wingender E: **Transcription regulatory region analysis using signal detection and fuzzy clustering.** *Bioinformatics* 1998, **14:**244-251.
92. Frech K, Danescu-Mayer J, Werner T: **A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter.** *J Mol Biol* 1997, **270:**674-687.
93. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15:**180-186.
94. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266:**231-245.
95. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17:**871-877.
96. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman W: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**13.
97. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen B, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301:**71-76.
99. Kellis M, Patterson N, Endrizzi M, Birren B, Lander E: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
100. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo G: **A comparative genomics approach to prediction of new members of regulons.** *Genome Res* 2001, **11:**566-584.
101. Webb C, Shabalina S, Ogurtsov A, Kondrashov A: **Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*.** *Nucleic Acids Res* 2002, **30:**1233-1239.
102. Bergman C, Pfeiffer B, Rincon-Limas D, Hoskins R, Gnirke A, Mungall C, Wang A, Kronmiller B, Pacleb J, Park S, *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0086.1-0086.20.
103. Ludwig M, Bergman C, Patel N, Kreitman M: **Evidence for stabilizing selection in a eukaryotic enhancer element.** *Nature* 2000, **403:**564-567.
104. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-1394.
105. Koop B, Hood L: **Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA.** *Nat Genet* 1994, **7:**48-53.
106. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature* 2003, **424:**788-793.
107. Sidow A: **Sequence first. Ask questions later.** *Cell* 2002, **111:**13-16.
118. Pabo C, Nekludova L: **Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?** *J Mol Biol* 2000, **301:**597-624.
119. Benos P, Lapedes A, Stormo G: **Is there a code for protein-DNA recognition? Probab(ilistical)ly...** *Bioessays* 2002, **24:**466-475.
120. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.