Molecular Systems Biology (2006) doi:10.1038/msb4100062
© 2006 EMBO and Nature Publishing Group All rights reserved 1744-4292/06
www.molecularsystemsbiology.com

molecular systems biology

NEWS AND VIEWS

Biological code breaking in the 21st century

Alan M Michelson[1] and Martha L Bulyk[1,2,3]

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [2] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA and [3] Harvard/MIT Division of Health Sciences and Technology (HST), Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

Molecular Systems Biology 2 May 2006; doi:10.1038/msb4100062

Just as early molecular biologists made the conceptual leap from the unique structure of DNA—with its precise sequence of nucleotides and complementary strands—to a theoretical and empirical solution to the genetic coding problem, four decades later a new generation of biologists is tackling a similar scientific challenge, this time of even greater complexity. How do DNA sequences specify the coordinated temporal and spatial expression patterns of functionally related genes throughout the life cycle of an organism? Whereas recognition of the fundamental relationships among DNA, RNA and protein provided the impetus for unraveling the genetic code, comparative sequence analysis, gene expression profiles, transcription factor binding site specificities, chromatin immunoprecipitation and powerful computational tools to analyze and integrate these diverse data sets are priming the way for deciphering the *cis*-regulatory codes that direct specific gene expression patterns. A paper by Ukkonen, Taipale and co-workers (Hallikas *et al*, 2006) in a recent issue of *Cell* is among the latest important contributions to this quest.

Detailed studies of transcriptional *cis*-regulatory elements from different species and biological contexts have yielded the general view that promoters and enhancers are modular; that is, they consist of closely clustered binding sites for one or more transcription factors. This structure facilitates combinatorial interactions among transcription factors, a mechanism that plays an essential role in generating gene expression specificity and diversity in metazoans (Levine and Tjian, 2003). Modularity also enables *cis*-regulatory elements to integrate convergent inputs from intrinsic factors—acquired early in the development of a cell—with later-acting extrinsic signals (Carroll *et al*, 2005). Superimposed on this complexity is the need for a cell to simultaneously express genes encoding proteins with related functions. Perhaps the most economical solution to the latter problem is for coexpressed genes to contain similar *cis*-regulatory elements, an assumption that is at the heart of current efforts to dissect transcriptional codes on a genomic scale.

A variety of computational methods have been developed to search whole genomes for related clusters of transcription factor binding sites that might constitute functional enhancers in coexpressed genes (reviewed by Bulyk, 2003). These studies have been most successful when transcription factors with known coregulatory functions are used for genome-wide scans, and when the binding sites of the individual factors have been very well characterized (reviewed by Michelson, 2002). At present, however, both of these prerequisites are met in only a limited number of systems. Additional barriers to enhancer prediction strategies include the relatively low information content of many individual transcription factor binding sites, and the large amount of noncoding sequence, particularly in vertebrate genomes.

Hallikas *et al* present a new computational approach that significantly advances current enhancer prediction strategies. They first adapted a method previously developed to study zinc-finger DNA-binding domains (Choo and Klug, 1994) in order to determine the binding affinities of the transcription factors GLI1-3, Ci, Tcf4 and cETS1 for specific DNA sequences. In their assay, Hallikas *et al* derive DNA-binding site data from a quantitative competition binding assay in which every single-base mismatch substitution of the consensus sequence is systematically examined. For a transcription factor of interest, this technique yields a position weight matrix of binding sequences based on their relative affinities. Although yielding useful information, a major limitation of this approach is that it requires knowledge of the likely consensus site to be used as the starting sequence, a requirement that is not inherent in other technologies for determining DNA-binding specificities (Oliphant *et al*, 1989; Mukherjee *et al*, 2004; Liu *et al*, 2005; Warren *et al*, 2006). Another drawback of this binding assay is that it considers only single-base substitutions, and so the resulting affinity-derived position weight matrix does not take into account potential effects of neighboring nucleotides. Using the present method, a much larger set of competing oligonucleotides would be required to address this potential variable, a goal that is more readily achieved using microarray-based approaches (Mukherjee *et al*, 2004; Liu *et al*, 2005; Warren *et al*, 2006).

Second, and perhaps more importantly, Hallikas *et al.* developed a novel computational algorithm called enhancer element locator (EEL), which searches for conserved patterns of transcription factor-binding sites in orthologous genes from two different species. In doing so, EEL does not require that the specific sequences of the binding sites be conserved. Rather, for the transcription factors of interest, it first uses position weight matrices—either affinity-based, as derived from the above method, or based on nucleotide frequencies present in high-quality sites taken from the literature and from the JASPAR2 database (Sandelin *et al*, 2004)—to find all possible binding site sequence matches beyond a predetermined motif match threshold. In the case of literature or JASPAR2 motifs, position weight matrix scores serve as proxies for affinity. EEL then uses a novel alignment strategy in which pairs of sites with conserved spacing are identified in orthologous genomic regions. The alignment scoring function considers the distances between paired binding sites such that a difference in this parameter between orthologous sequences is penalized. The penalty takes into account the distance separating the sites and the energy required to twist DNA such that the two aligned helices contain transcription factors bound at the same helical spacing. This penalty decreases as the site separation increases because a longer stretch of DNA is more flexible, thus more readily facilitating interactions between bound proteins and compensating for loss of factor proximity. On the other hand, the score is increased by the presence of high-affinity binding sites in both species.

A candidate enhancer identified by EEL is therefore characterized by the conservation of type, spacing and order (i.e., a pattern) of high-affinity transcription factor binding sites within a cluster. In this regard, EEL differs from other existing enhancer search algorithms that score actual binding site sequences—with or without evolutionary conservation considered—but which do not require a rigid spacing or order of sites (reviewed by Bulyk, 2003). Increased specificity is theoretically achieved by EEL's preference for conservation of spacing between orthologous sites as this feature rewards the expected strong selective pressure of preserving phase-dependent cooperative interactions among transcription factors. However, a major limitation of such a rigid enhancer model is that it does not take into account the evolutionary shuffling of binding sites that has been observed in some well-studied cases, a situation that is tolerated due to coevolving, compensatory changes occurring elsewhere in the enhancers (Ludwig *et al*, 2000, 2005). The overall density of known sites in a particular *cis*-regulatory element and the extent of their mutational turnover will determine whether or not EEL is sensitive to the latter effect. Indeed, one enhancer with a very high density of sites, only some of which have changed during evolution, was identified by this algorithm (Hallikas *et al*, 2006). Thus, the total number of sites available to be analyzed, the evolutionary distance between the species being compared and the number of comparison genomes used are likely to be critical factors in the successful application of one enhancer prediction algorithm or another in any given context (Moses *et al*, 2004).

A valuable feature of EEL is that it is readily scalable to large numbers of transcription factors and amounts of input genomic sequences. As presented, pairwise alignments of ~20 000 human and mouse gene pairs—including entire exonic, intronic and 100 kb of upstream and downstream sequences—were independently undertaken for each of 107 transcription factors for which high-quality binding site data are available. These alignments were then stored in a relational database that can be queried at the authors' website (http://www.cs.helsinki.fi/u/kpalin/EEL) for cooccurrences of the same or different types of sites (Hallikas *et al*, 2006). In this way, it is possible to identify candidate enhancers comprising transcription factor combinations that are inferred from independent lines of biological evidence.

Gene expression data provide another useful source of information that can be considered in the prediction of *cis*-regulatory elements (Beer and Tavazoie, 2004). Hallikas *et al* (2006) exploit this concept by seeking correlations between subsets of genes expressed under certain conditions, and the presence of specific transcription factor binding sites in those genes. As an example, they searched for overrepresentation of single and paired occurrences of all 107 transcription factor binding sites in sequences associated with genes that are upregulated with loss of the APC tumor suppressor. This analysis yielded a statistically significant overrepresentation of paired Tcf4 sites among APC target genes relative to the rest of the genome, a finding that is consistent with the documented activation of Tcf4 as a mediator of Wnt signaling in this system. Similarly, all pairwise combinations of the 107 transcription factors were examined. In a separate analysis, a genome-wide scan identified conserved cooccurrences of Tcf4 and GLI sites in genes known to be regulated by the two corresponding growth factors, Wnt and Hedgehog, respectively. Although the hypothesis that the genomic sequences associated with these sites represent a single enhancer capable of integrating Wnt and Hedgehog inputs remains to be empirically tested, this example highlights the potential power of the present strategy for identifying novel *cis*-regulatory elements.

Significant progress has clearly been made in the genome-wide analysis of *cis*-regulatory elements, but what does the future hold for this field? Emphasis must be placed on expanding and integrating five types of data or data analysis methods in next generation studies. First, binding specificities and affinities must be determined for all transcription factors in several model organisms, including yeast, *Drosophila* and mouse, which span a range of genome sizes and modes of transcriptional regulation. High-throughput methods now exist to make such studies feasible (Mukherjee *et al*, 2004; Liu *et al*, 2005; Warren *et al*, 2006).

Second, more refined gene expression studies must be undertaken to identify genes that are coexpressed at the resolution of single cells in whole organisms. These studies require multiple time points to provide a dynamic description of gene expression, and should evaluate responses to experimental manipulations such as appropriate environmental stimuli or genetic perturbations (Estrada *et al*, 2006). Searching for overrepresentation of particular binding sites associated with such gene sets should identify candidate coregulatory DNA elements (Spellman *et al*, 1998; Beer and Tavazoie, 2004; Hallikas *et al*, 2006; Philippakis *et al*, 2006). Development of even more powerful computational algorithms to search whole genomes for multiple combinations of transcription factor binding sites would also advance such studies by permitting *in silico* testing of how well novel regulatory models fit specific sets of coexpressed genes (Philippakis *et al*, 2006).

Third, *in vivo* transcription factor binding site localization data provide useful complementary information on binding site occupancy in intact cells under specific conditions (Wyrick and Young, 2002). More data of this type must be obtained and incorporated into enhancer prediction strategies to refine their output. Fourth, ways of simultaneously considering sequence alignments for multiple species are required to augment evolutionary conservation as a valuable factor in predicting enhancers (Moses *et al*, 2004).

Fifth, considerable effort must be placed on *in vivo* testing of predicted regulatory elements, including both enhancers and their constituent motifs. Once large numbers of both true-positive enhancers and false-positive candidates are available for any given model, these examples could be used as training sets to uncover details of *cis*-regulatory rules other than just the identities of the motifs involved. Features of binding site organization such as their orientation, spacing, number and position relative to transcription initiation are likely to profoundly influence enhancer activity (Senger *et al*, 2004).

Sleuths of the genetic code only needed to account for how a four-nucleotide alphabet is translated into a 20-amino-acid protein language. In contrast, 21st century code breakers are faced with a more daunting problem of how to functionally link hundreds (*Drosophila* and nematode) or thousands (mammals) of transcription factors—acting in as yet undefined combinations—to innumerable gene expression patterns in a single organism. Given the progress that has been made to date, along with the tools and concepts that are now available, we are optimistic that this field is engaged in a challenging and complex problem that, with further experimental data and more advanced computational analyses, will yield significant new advances in the not too distant future.

References

Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* **117:** 185–198

Bulyk M (2003) Computational prediction of transcription factor binding site locations. *Genome Biol* **5:** 201

Carroll SB, Grenier JK, Weatherbee SD (2005) *From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design*. Malden, MA: Blackwell Publishing

Choo Y, Klug A (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci USA* **91:** 11168–11172

Estrada B, Choe SE, Gisselbrecht SS, Michaud S, Raj L, Busser BW, Halfon MS, Church GM, Michelson AM (2006) An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genet* **2:** e16

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124:** 47–59

Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* **424:** 147–151

Liu X, Noll DM, Lieb JD, Clarke ND (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res* **15:** 421–427

Ludwig M, Bergman C, Patel N, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M (2005) Functional evolution of a *cis*-regulatory module. *PLoS Biol* **3:** e93

Michelson AM (2002) Deciphering genetic regulatory codes: a challenge for functional genomics. *Proc Natl Acad Sci USA* **99:** 546–548

Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* **5:** R98

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **36:** 1331–1339

Oliphant AR, Brandl CJ, Struhl K (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol* **9:** 2944–2949

Philippakis AA, Busser BW, Gisselbrecht SS, He FS, Estrada B, Michelson AM, Bulyk ML (2006) Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Comp Biol* (in press)

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32:** D91–D94

Senger K, Armstrong GW, Rowell WJ, Kwan JM, Markstein M, Levine M (2004) Immunity regulatory DNAs share common organizational features in *Drosophila*. *Mol Cell* **13:** 19–32

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9:** 3273–3297

Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips Jr GN, Ansari AZ (2006) Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci USA* **103:** 867–872

Wyrick JJ, Young RA (2002) Deciphering gene expression regulatory networks. *Curr Opin Genet Dev* **12:** 130–136