Commentary

Deciphering genetic regulatory codes: A challenge for functional genomics

Alan M. Michelson*

Howard Hughes Medical Institute, Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 20 Shattuck Street, Boston, MA 02115

In the past two decades, considerable effort has been devoted to elucidating the mechanisms of transcriptional regulation in metazoans. A number of fundamental principles have been established concerning the functions of many transcription factors (TFs) and the cis-acting sequences to which they bind (1). One hypothesis that has emerged from these studies is that genes with similar temporal and spatial expression patterns are subject to a common regulatory logic. That is, unique ''transcriptional codes'' govern the activation and repression of genes in particular developmental contexts (2, 3). However, because of the laborious nature of cis-regulatory sequence dissection, few comprehensive examples exist to support this concept. A more efficient approach to the identification of coexpressed genes and their associated regulatory elements would accelerate this field greatly. The availability of complete human and model organism genome sequences represents a tremendous windfall for those interested in this problem. Two papers appearing in this issue of PNAS (4, 5) exemplify how a marriage between computational and experimental biology can yield a powerful approach for exploiting genomic information to predict and validate new genes, the expression of which are subject to similar transcriptional codes.

The complex spatial and temporal patterns of gene expression occurring in development are orchestrated by cis-acting regulatory modules (CRMs) or enhancers (2, 3). CRMs comprise sets of short oligonucleotide motifs, each of which has an affinity for one or more sequence-specific DNA-binding proteins or TFs. TFs in turn interact with each other and the basal transcriptional machinery to either activate or repress the expression of coding regions associated with the particular CRM. A recurring theme in the organization of CRMs is their ability to integrate multiple convergent inputs through the binding of TFs belonging to different classes, often in a cooperative manner (6–10). The clusters of binding sites found within a CRM can include multiple copies of the same or different motifs. This combinatorial nature of CRMs contributes to the response specificity of proteins that individually would not discriminate among different targets effectively and at the same time broadens the diversity of potential outputs generated by a limited set of factors. For example, unique combinations of tissue-specific selector proteins and signal-activated TFs can induce target gene expression in exquisitely precise domains, and a given TF can activate distinct genes in different developmental contexts (6, 8–10).

A major goal of developmental biologists is to understand in fine detail how the genomic control apparatus is organized and functions, that is, how sets of related genes are coexpressed. The traditional approach to this problem is to use *in vitro* and *in vivo* methods to analyze individual CRMs, a slow and painstaking process for the characterization of genetic networks. For larger scale discovery of candidate regulatory regions, computational algorithms have been developed for genome-wide scans (reviewed in refs. 11–13). However, with a purely computational approach, uncertainty remains as to whether a predicted CRM actually possesses the expected function. The work of Markstein *et al.* (4) and Berman *et al.* (5) represents an important step forward by addressing this issue directly.

With a purely computational approach, uncertainty remains as to whether a predicted CRM actually possesses the expected function.

These two groups used related but distinct computational strategies for the prediction of coexpressed genes and their associated CRMs in the *Drosophila melanogaster* genome (4, 5). Each identified sets of similar CRMs based on the dense clustering of individual TF binding sites. However, whereas Markstein *et al.* used a single class of TF, Berman *et al.* employed five different TFs with known concerted functions during *Drosophila* embryogenesis. Most importantly, both efforts involved not only computational predictions but also experimental evaluation of the candidates obtained from their respective genome-wide searches. This is a critical aspect of this approach that distinguishes it from prior studies where experimental validation of novel CRMs derived from a search was not undertaken (14–18).

The work of Berman *et al.* (5) builds on earlier efforts that defined the hierarchy of TFs controlling anteroposterior patterning of the *Drosophila* embryo (3). Five such TFs with well defined DNA binding specificities were selected: Bicoid (Bcd), Caudal (Cad), Hunchback (Hb), Krüppel (Kr), and Knirps (Kni). A position weight matrix (19), which reflects the frequency with which a given nucleotide appears in each position of a binding site, was constructed for each set of available TF recognition sequences. The position weight matrixes then were used to search the *Drosophila* genome for the locations of potential TF sites. A further parameter was added to the search algorithm to eliminate those sites with a theoretical low affinity for each factor. An initial test run of 1 megabase of sequence surrounding *even skipped* (*eve*), a known target gene of these regulators, identified the majority of experimentally documented binding sites as well as many others. To distinguish between random and functionally relevant occurrences of these sites, an additional consideration was introduced: the latter should be densely clustered, as defined arbitrarily by the colocalization of at least 13 sites in a 700-bp window. The validity of this *in silico* analysis was established with the identification of three previously characterized *eve* stripe enhancers.

Extending their search strategy to the entire *Drosophila* genome but increasing the requisite density of TFs imposed by their program, Berman *et al.* (5) found an addi-

See companion articles on pages 757 and 763.

*E-mail: michelson@rascal.med.harvard.edu.

 www.pnas.org/cgi/doi/10.1073/pnas.032685999

tional 28 clusters that defined 49 candidate target genes (some clusters fell within introns, whereas others were located in intergenic regions, the latter defining two potential targets). Of these, ≈40% were found to be expressed in early embryos in patterns consistent with regulation by the TFs used to model the search. As a further empirical test of this approach, one of the high-density clusters that is found just upstream of the gap gene, *giant* (*gt*), was evaluated for enhancer activity in transgenic embryos. Strikingly, this genomic sequence directed reporter gene transcription in a pattern that faithfully recapitulated the posterior domain of endogenous *gt* expression. Thus, a previously unknown enhancer was identified by a purely computational approach.

In designing their computational search, Markstein *et al.* focused on only one TF, Dorsal (Dl; ref. 4). Dl is involved in patterning the early *Drosophila* embryo by activating or repressing particular genes in discrete regions along the dorsoventral axis (3). A Dl-responsive silencer from one gene, *zerknullt* (*zen*), was used to develop a specific model for a genome-wide computational scan to identify related CRMs. In this case, clusters of at least three high-affinity Dl binding sites were sought in a 400-bp window. This search yielded only 15 matches, an improbable occurrence by chance alone. Furthermore, three of these matches were associated with known Dl-responsive genes. One such candidate Dl-dependent CRM was found in an intron of *short gastrulation* (*sog*). When fused to a reporter transgene, this putative enhancer activated transcription in a lateral embryonic domain that corresponds to that of endogenous *sog*. Two additional candidate Dl target genes were found to be expressed in patterns that are consistent with direct regulation by this TF, although the functions of the associated CRMs were not assessed. These examples affirm the feasibility of computationally discovering new regulatory elements that adhere to a common code defined by Dl binding.

Markstein *et al.* built a successful search algorithm around a single TF binding site (4). However, even here combinatorics can be applied. Some characterized Dl-responsive enhancers are known to contain binding sites for other TFs such as the zinc finger protein Snail (20). At least one of the new CRMs identified also contained potential Snail sites, which can explain features of its function. Twist, a basic helix–loop–helix TF, also acts together with Dl in regulating some enhancers (21). Inclusion of these additional classes of sites in the search algorithm might reduce the rate of false-positive returns. One explanation offered by the authors for such false positives is the possibility that these genes are targets of other Dl family members with similar DNA binding specificities but with activities in other stages of development. If this hypothesis is correct, then it is unlikely that these CRMs would contain the same additional binding sites as true Dl-responsive enhancers. Rather, they should have their own combinations of interacting TFs that contribute to their specificities. A similar case can be made for false positives obtained in the screen for segmentation gene targets (5).

Although the two studies reviewed here represent a significant advance in the application of bioinformatics to understanding genetic regulatory networks, how generally applicable are the approaches? There are ≈700 TFs in the *Drosophila* genome (22), most of which are active at much later stages of development than those examined in the present papers. This is a potentially significant issue, because these studies involved several TFs that function as morphogens early in development when the fly embryo is a single syncytial cell; that is, such TFs generate unique threshold responses at different concentrations produced by free diffusion of the proteins within the embryonic syncytium (3, 23). This point is relevant to the clustering of cognate DNA binding sites, because the number and affinities of these sites provide a readout of the local TF concentrations. For example, a CRM with a large number of high-affinity sites would be activated in response to low levels of the corresponding TF (24). Although other mechanisms might generate cell-specific TF levels, the majority of *Drosophila* transcriptional regulators probably do not act as morphogens and thus may be less likely to be associated with high binding-site densities. Rather, combinatorial interactions of multiple TFs each binding to relatively few sites may be the rule in these cases. Any computational algorithm designed to identify such CRMs must take this point into account. Similar arguments apply to attempts at analyzing the ≈1,850 TFs estimated to be encoded by the human genome (25).

A previously unknown enhancer was identified by a purely computational approach.

Another limitation of computational efforts to predict CRMs relates to the sequence complexity of the binding sites included in the search algorithm. A single TF was successfully used by Markstein *et al.* (4) due not only to the dense clustering of its sites in the genome but also to its relatively high binding specificity. It is less likely that this approach would be effective for TFs such as Hox proteins, the binding sites of which have a much lower information content (26). Although protein cofactors can modify Hox binding specificity, in these cases a combinatorial paradigm also would be useful for CRM predictions. Indeed, this was a critical design feature contributing to the success of Berman *et al.* where homeodomain proteins and other TFs with relatively degenerate binding sequences were included in the combinatorial model (5). A similar approach should be applicable to other TFs that have modest binding site specificities and are known to act in biologically relevant combinations (6, 8–10).

Future attempts to represent the combinatorial logic underlying CRM architecture in a specific biological context will require not only consideration of binding-site type and number but also their spacing, orientations, affinities, and order. These parameters will be difficult if not impossible to predict accurately *a priori*. Instead, the derivation of a directed computational approach to identity members of a particular regulatory network will benefit from the availability of at least one well defined representative of that network to serve as a starting paradigm. Additional candidates then might be predicted computationally from this initial example. Empirical testing of these candidates should enable further refinement of the model, which in turn can be applied in another round of computational screening. Again, this expectation underscores the importance of combining informatics with wet laboratory methodologies, as exemplified by the precedent established in the present papers.

Although both of these studies successfully predicted CRMs belonging to defined regulatory networks, the extent to which the individual TF binding sites within these CRMs contribute to enhancer activity was not evaluated. Not all the newly identified sites are necessarily functional, because there is a reasonable probability of random occurrence even within a *bona fide* module. Testing the functions of individual binding sites is of obvious relevance to the aforementioned goal of refining a particular combinatorial model through an iterative screening and validation strategy. Although a time-consuming effort, such tests are valuable, because their results will lead to revised models that more closely resemble authentic CRMs, thereby increasing the sensitivity and specificity of the derived computational algorithms.

In addition to searching an entire genome for known TF binding sites, it is possible to screen for novel motifs shared by a given set of sequences (27–29). Al-

though neither of the present papers took advantage of such an approach, this could be an informative way of identifying as yet unrecognized cis-regulatory sequences. This strategy also can be used to increase the combinatorial complexity of a specific transcriptional code that could be applied to a sequential screening strategy for CRM characterization. In addition, empirical efforts to identify more TF binding specificities will provide a larger database on which to draw in formulating such paradigms.

Several other experimental approaches can contribute valuable information to increase the precision of computational screens designed to identify coregulated genes. One is the use of comparative sequence data from a related species, so-called phylogenetic footprinting (13). This method relies on the evolutionary conservation of orthologous noncoding sequences caused by selection for regulatory functions. Such an approach should be possible soon for the fruit fly, because an effort is under way to sequence the genome of another *Drosophila* species. Large-scale comparisons between mouse and human genomic sequences should be useful also in this regard (30, 31).

Another set of data that can be incorporated into CRM search algorithms is expression data as derived from genome-wide expression profiling or high throughput *in situ* hybridization screens of cDNA collections. This approach can be used as a filter for identifying the most likely coregulated genes from among the computational candidates or as a means of grouping biologically related TFs in an initial combinatorial model. Chromatin immunoprecipitation studies offer yet another source of information that can contribute to the assignment of genes to a common regulatory network (32, 33). Finally, the construction and functional assessment of synthetic enhancers can be used to test a particular combinatorial model to ensure that the incorporated features reproduce the intended effect (6, 7).

There is clearly much work to be done before we have a comprehensive understanding of transcriptional codes on a genomic scale. However, the promising strategies and associated findings reported by Berman *et al.* and Markstein *et al.* suggest that this is a tractable problem. It seems reasonable to anticipate, therefore, that in the not-too-distant future we should gain considerable fresh insights into the organization of genetic regulatory networks in both invertebrate and vertebrate systems.

I thank Martha Bulyk, Yonatan Grad, and Marc Halfon for helpful discussions and thoughtful comments on the manuscript. A.M.M. is an Associate Investigator of the Howard Hughes Medical Institute.

1. Blackwood, E. M. & Kadonaga, J. T. (1998) *Science* **281,** 60–63.
2. Davidson, E. H. (2001) *Genomic Regulatory Systems* (Academic, San Diego).
3. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. (2001) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell Science, Malden, MA).
4. Markstein, M., Markstein, P., Markstein, V. & Levine, M. S. (2002) *Proc. Natl. Acad. Sci. USA*. **99,** 763–768.
5. Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. (2002) *Proc. Natl. Acad. Sci. USA*. **99,** 757–762.
6. Guss, K. A., Nelson, C. E., Hudson, A., Kraus, M. E. & Carroll, S. B. (2001) *Science* **292,** 1164–1167.
7. Yuh, C.-H., Bolouri, H. & Davidson, E. H. (1998) *Science* **279,** 1896–1902.
8. Xu, C., Kauffmann, R. C., Zhang, J., Kladny, S. & Carthew, R. W. (2000) *Cell* **103,** 87–97.
9. Flores, G. V., Duan, H., Yan, H., Nagaraj, R., Fu, W., Zou, Y., Noll, M. & Banerjee, U. (2000) *Cell* **103,** 75–85.
10. Halfon, M. S., Carmena, A., Gisselbrecht, S., Sackerson, C. M., Jiménez, F., Baylies, M. K. & Michelson, A. M. (2000) *Cell* **103,** 63–74.
11. Pennacchio, L. A. & Rubin, E. M. (2001) *Nat. Rev. Genet.* **2,** 100–109.
12. Ohler, U. & Niemann, H. (2001) *Trends Genet.* **17,** 56–60.
13. Fickett, J. W. & Wasserman, W. W. (2000) *Curr. Opin. Biotechnol.* **11,** 19–24.
14. Gailus-Durner, V., Scherf, M. & Werner, T. (2001) *Mamm. Genome* **12,** 67–72.
15. Frech, K., Danescu-Mayer, J. & Werner, T. (1997) *J. Mol. Biol.* **270,** 674–687.
16. Wasserman, W. W. & Fickett, J. W. (1998) *J. Mol. Biol.* **278,** 167–181.
17. Krivan, W. & Wasserman, W. W. (2001) *Genome Res.* **11,** 1559–1566.
18. Frith, M. C., Hansen, U. & Weng, Z. (2001) *Bioinformatics* **17,** 878–889.
19. Stormo, G. D. (2000) *Bioinformatics* **16,** 16–23.
20. Ip, Y. T., Park, R. E., Kosman, D., Bier, E. & Levine, M. (1992) *Genes Dev.* **6,** 1728–1739.
21. Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K. & Levine, M. (1992) *Genes Dev.* **6,** 1518–1530.
22. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287,** 2185–2195.
23. Gurdon, J. B. & Bourillot, P.-Y. (2001) *Nature (London)* **413,** 797–803.
24. Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H. & Jackle, H. (1995) *Nature (London)* **376,** 253–256.
25. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) *Science* **291,** 1304–1351.
26. Mann, R. S. & Morata, G. (2000) *Annu. Rev. Cell Dev. Biol.* **16,** 243–271.
27. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
28. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000) *J. Mol. Biol.* **296,** 1205–1214.
29. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998) *Nat. Biotechnol.* **16,** 939–945.
30. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000) *Nat. Genet.* **26,** 225–228.
31. Hardison, R. C. (2000) *Trends Genet.* **16,** 369–372.
32. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000) *Science* **290,** 2306–2309.
33. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. (2001) *Nature (London)* **409,** 533–538.