Decoding cis-regulatory DNAs in the Drosophila genome
Michele Markstein* and Michael Levine†

Cis-regulatory DNAs control the timing and sites of gene expression during metazoan development. Changes in gene expression are responsible for the morphological diversification of metazoan body plans. However, traditional methods for the identification and characterization of cis-regulatory DNAs are tedious. During the past year, computational methods have been used to identify novel cis-DNAs within the entire Drosophila genome. These methods change the way that cis-DNAs will be analyzed in future studies and offer the promise of unraveling complex gene networks.

Addresses
* Department of Molecular Cell Biology, Division of Genetics & Development, 401 Barker Hall, University of California, Berkeley, California 94720, USA
† Committee on Developmental Biology, University of Chicago, Cummings Life Sciences Center, 755, 920 East 58th Street, Chicago, Illinois 60637, USA

0959-437X/02/$ — see front matter © 2002 Elsevier Science Ltd. All rights reserved.

Abbreviations
ChIP chromosome immunoprecipitation
eve even-skipped
hbr heartbroken
sog short-gastrulation
Su(H) Suppressor of Hairless
zen zerknüllt

Introduction
A clear revelation of the post-genome era is that organismal complexity does not correlate with gene number. The fruitfly, Drosophila melanogaster, contains <14,000 genes [1], whereas the nematode worm Caenorhabditis elegans—a considerably less complex animal—contains ~20,000 genes [2]. Complexity is more likely to be a manifestation of the total number of gene expression patterns that are produced during development. Perhaps Drosophila contains a greater number of cis-regulatory DNAs than C. elegans, and thereby exhibits more sophisticated morphologies and behaviors. The average fly gene might be regulated by three or four different cis-DNAs, and consequently, the fly genome could produce something like 50,000 distinct patterns of gene expression. If worms contain an average of just one or two cis-DNAs per gene, then the worm genome might produce half the total number of gene-expression patterns than flies, even though it contains 50% more genes. Here, we discuss recent efforts to employ computational methods to identify cis-regulatory DNAs within complex metazoan genomes.

Pre-genomics analysis of cis-regulatory DNAs
There are several classes of cis-regulatory DNAs, including enhancers [3], silencers [4], and insulators [5]. Enhancers represent the most thoroughly analyzed type of cis-regulatory DNA. The characterization of enhancers in transgenic worms, flies, sea urchins, ascidians, fish, frogs, chicks, and mice, suggests that a typical enhancer mediating cell-type specific expression is 300 bp to 1 kb in length and contains clustered binding sites for both transcriptional activators and repressors [6]. In flies, a typical enhancer such as the even-skipped (eve) stripe 2 enhancer [7] contains a total of 10 binding sites for three different sequence-specific transcription factors; two of the factors function synergistically to activate gene expression, whereas the third represses transcription in inappropriate cell types.

Emergence of computational methods: from in silico to in vivo
The detailed characterization of cell-specific enhancers in transgenic embryos is a laborious process. Consequently, comparatively few enhancers—something like 100 in all animals combined—have been examined [6]. Although several computational approaches have been developed to identify cis-regulatory motifs and regions in silico, until recently, very few of these predictions have been tested in multicellular animals [8,9]. In the past year, bioinformatics methods have been used to identify authentic enhancers within the Drosophila genome [10–13]. Although these are early days for the computational identification of enhancers, the methods we briefly summarize in this review have permanently changed the way such cis-regulatory DNAs will be characterized in the future. Several of these methods are available as web-based tools at the URLs listed in Box I.

Computing Dorsal target enhancers
We begin with the dorsal–ventral patterning of the early Drosophila embryo. Localized patterns of gene expression depend on a sequence-specific transcription factor called Dorsal [14]. The Dorsal protein is distributed in a broad nuclear gradient, with peak levels in ventral regions and progressively lower levels in lateral and dorsal regions (Figure 1a). The Dorsal gradient differentially regulates as many as 25 different target genes in a concentration-dependent manner. Some of the target genes are activated only by high concentrations of the gradient. As a result, these genes exhibit localized expression in ventral regions that form the mesoderm. By contrast, other Dorsal target genes are regulated by low levels of the gradient, and are activated, or repressed, in lateral regions that form the neurogenic ectoderm.

A total of seven different Dorsal target cis-regulatory DNAs were characterized through the traditional method of attaching random DNA fragments from the 5′ flanking regions of the target genes to a lacZ reporter gene. These lacZ fusion genes were individually integrated into the Drosophila genome using P-element gene transfer.
Embryos were collected from transgenic strains, and stained for lacZ gene expression to determine whether any of the DNA fragments were sufficient to recapitulate authentic aspects of the endogenous gene expression patterns. In this way, minimal tissue-specific enhancers were identified for two genes expressed in the mesoderm, twist and snail [15–17], and two genes expressed in the neurogenic ectoderm, single-minded and rhomboid [18,19]. Another three cis-regulatory DNAs, silencers, were likewise identified for the tolloid, zerknöllt (zen), and decapentaplegic genes [20–22]. These silencers keep expression off in ventral and lateral regions in response to high and low levels of the Dorsal gradient.

The characterization of these seven cis-regulatory DNAs required several years of effort from several laboratories. Markstein et al. [10*] recently developed a computational method for identifying clusters of Dorsal-binding sites in the Drosophila genome. This method led to the rapid identification of a new Dorsal target enhancer. The Dorsal protein binds DNA as an obligate dimer, and recognizes a spectrum of sites with dyad symmetry [23]. Dorsal-binding sites have been defined by SELEX assays as well as DNase I and chemical footprinting assays. These studies provided concise consensus sequences representing 106 optimal, high-affinity binding sites [23]: GGGWWWWWCCM and GGGWDWWWCCM (W = A or T, M = C or A, D = A or T or G). Four optimal Dorsal sites conforming to these consensus matrices have been observed within a 400 bp region of the 600 bp zen silencer sequence [23]. These high-affinity sites mediate the repression of zen even in dorso-lateral regions where there are vanishingly low levels of the Dorsal gradient. A survey of the entire Drosophila genome for clusters containing at least three optimal Dorsal-binding sites within 400 bp identified 15 novel clusters in addition to the zen cluster. As only four Dorsal clusters would be expected by chance alone — using the binomial distribution as an approximation — the occurrence of 16 clusters is consistent with the notion that positive evolutionary selection has maintained the functional integrity of the clusters.

One of the novel clusters is located within the first intron of the short-gastrulation (sog) gene, which was identified as a potential target of the Dorsal gradient in genetic studies [24]. A 393 bp genomic DNA fragment that encompasses this cluster was placed 5′ of a lacZ reporter gene containing a minimal, 42-bp ‘naïve’ promoter (Figure 1a). The sog-lacZ fusion gene exhibits broad lateral stripes of expression in transgenic embryos that are similar to those observed for the endogenous gene. Another four of the novel clusters are associated with genes that exhibit early, localized expression across the dorsal–ventral axis. One of these clusters is located ~10.5 kb 5′ of the brinker gene, which is a known genetic target of the Dorsal gradient [25]. These new clusters are currently being tested for enhancer activities in transgenic embryos. It is possible that 5 of the 15 novel Dorsal binding clusters in the fly genome, one-third, correspond to authentic enhancers.

Computing Suppressor of Hairless target enhancers

A similar approach was used by Rebeiz et al. [11*] to examine the clustering of Suppressor of Hairless (Su[H]) binding sites in the Drosophila genome. Su(H) is a sequence-specific transcription factor that mediates Notch signaling in a variety of developmental processes in Drosophila [26]. A major function of Notch signaling and Su(H) is to inhibit neurogenesis. This is accomplished, at least in part, by the activation of a set of sequence-specific transcriptional repressors encoded by the Enhancer of Split gene complex. These Enhancer of split proteins bind to specific sites within the regulatory regions of achaete–scute genes, which are essential for the development of neurons within the central and peripheral nervous systems [26]. The Su(H) protein is localized within the nucleus, where it appears to function as a transcriptional repressor in the absence of Notch signaling. Upon signaling, the Notch receptor is proteolytically processed. The Notch intracytoplasmic domain (NotchIC) is released from the cell surface and interacts with Su(H) within the nucleus [27]. The Su(H)–NotchIC complex functions as a transcriptional activator and induces the expression of Enhancer of split genes, which were formerly repressed by Su(H) in the absence of Notch signaling [26].

Several Notch/Su(H) target genes have been characterized, and all of the regulatory regions that were identified contain clusters of Su(H)-binding sites conforming to the consensus sequence YGTGTDGAA [11*]. A search of the Drosophila genome for statistically significant clustering of high-affinity Su(H) sites — defined by YGTGREGAA and CGTGDGAA — identified 46 novel clusters, ranging in size from 300 bp to 5 kb. On the basis of expression patterns of the associated genes, and the analysis of one of the clusters in transgenic flies, it would appear that at least
12 of the 46 clusters, ~25% correspond to Notch target enhancers. This 'hit rate' is similar to that described for Dorsal-binding clusters.

### Computing segmentation enhancers

The preceding analyses document the efficacy of exploiting binding site matrices of single transcription factors to identify authentic cis-regulatory DNAs. Approximately one-fourth to one-third of the clusters identified by Dorsal or Su(H)-binding sites might correspond to either enhancers or silencers. However, as we discussed earlier, a typical enhancer is regulated by multiple factors [6]. It is reasonable to anticipate that the analysis of multiple factors should lead to a higher 'hit rate'. That is, perhaps most of the clusters for multiple factors engaged in a common process correspond to authentic cis-regulatory DNAs. A seemingly ideal test case is provided by the segmentation process in the Drosophila embryo.

Opposing gradients of two maternal homeodomain proteins, Bicoid and Caudal, lead to the localized expression of several gap genes, which encode zinc finger repressors, including Hunchback, Krüppel, and Knirps (e.g. see [28]). Segmentation stripes of gene expression, such as the localized expression of *eve stripe 2*, are established through the interplay of the maternal Bicoid and Caudal activators, and the gap repressors (Figure 1b). The *eve stripe 2* enhancer contains five binding sites for the Bicoid activator, as well as six binding sites for gap repressors [7]. Berman et al. [12•] created position weight matrices for the binding sites of Bicoid, Caudal, Hunchback, Krüppel, and Knirps using a broad spectrum of binding sites compiled from 19 segmentation matrices were used to identify clusters of Bicoid, Caudal, Hunchback, Krüppel, and Knirps using a broad spectrum of binding sites compiled from 19 segmentation enhancers previously described in the literature. These matrices were used to identify clusters of Bicoid, Caudal, and gap-binding sites throughout the non-coding portion of the fly genome. A search for clusters containing 15 sites within 700 bp identified half of the 19 enhancers upon which the search was based and an additional 22 novel clusters. These clusters were pooled with 6 other novel clusters identified by an equally stringent screen for clusters containing 4 rather than 5 of the transcription factors. Of these pooled 28 novel clusters ~10 are associated with genes showing localized expression along the anterior–posterior axis, such as *odd* and *pdm2* (Figure 1b). One of the 10 clusters was shown to mediate the posterior expression pattern of the segmentation gene, *giant* [12•,29,30]. So, something like one-third of the novel clusters might correspond to authentic segmentation enhancers. This value is similar to the hit-rate observed for individual Dorsal- and Su(H)-binding site clusters. Bicoid, Caudal, and the gap proteins
bind DNA as monomers and recognize degenerate sequences with low-binding affinity (e.g. [31]). Perhaps the clustering of multiple monomeric factors provides no more specificity than the clustering of a single dimeric protein such as Dorsal. One possible way to improve studies based on multiple sequences with low specificity (or information content) may be to use Boolean operators (i.e. ‘and’, ‘or’, ‘not’) to require specific combinations of sites, thereby increasing stringency.

Computing heart enhancers

An effort to use specific combinations of multiple binding sites (using the Boolean operator ‘and’) was recently reported by Halfon et al. [13•]. These authors focused their analysis on the well-characterized heart enhancer from the S′ regulatory region of the eve gene [32–34]. This enhancer is regulated by a variety of transcription factors, including dIcf, Mad, and Pointed, which are mediators of Wingless, Transforming growth factor-β (TGF-β), and Sevenless (Sev) receptor tyrosine kinase signaling pathways, respectively [11•]. The eve heart enhancer also contains binding sites for the mesoderm determinant, Twist, as well as Tinman, a homeodomain transcription factor essential for heart differentiation in flies and mice [35]. The Drosophila genome was examined for clusters containing at least two instances of each of the binding sites — Mad, Pointed, Twist, and Tinman — and one instance of the Icf site, in an effort to identify additional heart-specific enhancers. A total of 33 novel regions were identified. One is located within the first intron of the heartbroken (hbr) gene, which is specifically expressed in the heart [36]. A DNA fragment containing the clustered sites was shown to direct heart-restricted expression in transgenic embryos. Two of the remaining clusters appear to be associated with heart-specific genes. However, when examined by transgenic analysis, these clusters did not recapitulate the endogenous profiles of the associated genes. This lower than expected hit rate may be due to the limited information content of some of the binding site matrices used in the study (e.g. Mad-binding sites). This suggests that the use of Boolean operators may not be able to overcome the noise created by poorly defined sites and highlights the need for well-defined, information-rich binding site matrices.

More motifs, higher hit rate?

In all of the cases we have discussed, dorsal–ventral patterning, Notch signaling, segmentation, and heart morphogenesis, the bioinformatics methods led to both false negatives and false positives. For example, many of the genes that are known to be regulated by the Dorsal gradient were not identified on the basis of optimal binding clusters because they are regulated by low-affinity recognition sequences that possess extensive degeneracy [10•]. A major goal of future efforts will be to increase the hit-rate and eliminate false negatives. Toward this end we anticipate the need to develop computational methods to identify shared sequence motifs among coordinately regulated enhancers. The use of such motifs has the potential to refine subsequent computational searches for new enhancers that mediate related patterns of gene expression.

Whereas several wet-lab methods are available for identifying coordinately expressed genes and enhancers in (e.g. cDNA microarray assays, chromatin immunoprecipitation [ChIP],
automated large-scale in situ hybridizations), relatively few attempts have been made to identify and test novel motifs shared among co-expressed genes in multicellular animals [37••,38]. A recent study comparing Drosophila segmentation enhancers for shared motifs accurately identified both known and novel motifs. The known motifs correspond to binding sites for maternal and gap transcription factors [39]. The novel shared motifs may therefore be relevant to enhancer function but functional tests for these motifs have not yet been reported. The use of such motifs has the potential to refine subsequent computational searches for new enhancers that mediate related patterns of gene expression. Some of the computational methods that can identify shared motifs are available as web-based tools at the URLs listed in Box 1.

An initial attempt to test computational predictions of novel binding sites was reported by Halfron et al. [13•]. The eve enhancer was compared with the novel clusters identified as potential heart enhancers in the Drosophila genome. A conserved sequence motif was identified that is located in both the eve and Hbr heart enhancers. This motif is related to the binding site for the Oct-1 transcription factor. Mutations in binding sites was reported by Halfron [13•]. A recent study comparing even-skipped stripe 2 in the Drosophila embryo, EMBO J 1992, 11:4047-4057.

Conclusions

The preceding studies demonstrate that bioinformatics methods can be used to identify novel enhancers. This will forever change the way that cis-regulatory DNAs are characterized in complex metazoan genomes. However, these studies have not yet revealed a ‘cis-regulatory code’, whereby gene-expression patterns can be inferred from simple sequence analysis. Better predictions may be achieved through the identification of additional conserved features of cis-regulatory DNAs, beyond the simple clustering of binding sites. For example, the intergenic ‘enhancosome’ contains binding sites that are separated by fixed distances, which facilitate protein–protein interactions [40]. The sog intronic enhancer contains four evenly spaced Dorsal binding sites that are separated by distances of ~80 bp — about one turn of the nucleosome (Figure 2). Helical phasing, nucleosome phasing, and the stereochemistry of binding, provide a ‘grammar’ that should increase the hit-rate in future computational searches for cis-regulatory DNAs.

Acknowledgements

We thank Yutaka Nibu for help with the figures and Kate Senger for helpful discussions. This work was funded by a grant from the NIH (GM46638).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


10. Review of classic wet-lab and genomic approaches (including phylogenetic footprinting, position weight matrices and inter-species comparisons) to identify cis-regulatory DNAs in mammals.


13. Clusters of multiple maternal and gap proteins are shown to be associated with segmentations genes in the Drosophila genome. One such cluster corresponds to an authentic enhancer in the giant 5’ regulatory region. This study used the programme ‘cis-analyst’ to identify and plot the clusters, and it is available on the web at http://www.fruitfly.org/cis-analyst.


15. The eve heart enhancer is used as a model to identify an intronic enhancer in the Hbr gene. Comparison of the eve and Hbr enhancers identifies a new conserved motif that mediates repression within the eve heart enhancer. To find new motifs the study used the programme ‘coccur_scan’ available at http://arep.med.harvard.edu/Halfon_Grad_etal/supplemental.html.


