Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*

Yonatan H. Grad¹, Frederick P. Roth², Marc S. Halfon³, and George M. Church*¹

¹The Lipper Center for Computational Genetics and Department of Genetics, 77 Avenue Louis Pasteur, Harvard Medical School, Boston, Massachusetts, 02115, USA

²Department of Biological Chemistry and Molecular Pharmacology, 250 Longwood Avenue, Harvard Medical School, Boston, Massachusetts, 02115, USA

³Department of Biochemistry and Center of Excellence in Bioinformatics, SUNY at Buffalo, Buffalo, New York, 14214, USA

*To whom correspondence should be addressed (see http://arep.med.harvard.edu)
Abstract:

Motivation: To date, computational searches for cis-regulatory modules (CRMs) have relied on two methods. The first, phylogenetic footprinting, has been used to find CRMs in non-coding sequence, but does not directly link DNA sequence with spatio-temporal patterns of expression. The second, based on searches for combinations of transcription factor (TF) binding motifs, has been employed in genome-wide discovery of similarly acting enhancers, but requires prior knowledge of the set of TFs acting at the CRM and the TFs’ binding motifs.

Results: We propose a method for CRM discovery that combines aspects of both approaches in an effort to overcome their individual limitations. By treating phylogenetically footprinted non-coding regions (PFRs) as proxies for CRMs, we endeavor to find PFRs near co-regulated genes that are composed of similar short, conserved sequences. Using Markov chains as a convenient formulation to assess similarity, we develop a sampling algorithm to search a large group of PFRs for the most similar subset. When starting with a set of genes involved in Drosophila early blastoderm development and using phylogenetic comparisons of D. melanogaster and D. pseudoobscura genomes, we show here that our algorithm successfully detects known CRMs. Further, we use our similarity metric, based on Markov chain discrimination, in a genome-wide search, and uncover additional known and many candidate early blastoderm CRMs.

Availability: Software is available via http://arep.med.harvard.edu/enhancers

Contact: see http://arep.med.harvard.edu/email.html

Supplementary Information: Can be accessed at http://arep.med.harvard.edu/enhancers
**Introduction:**

The complex regulation of metazoan gene expression is substantially controlled through the interaction of transcription factors (TFs) and *cis*-regulatory DNA sequences. These *cis*-regulatory sequences are organized into modules, where each module integrates input from a specific set of transcription factors to direct a corresponding spatiotemporal expression pattern. These key regulatory sequences, termed *cis*-regulatory modules (CRMs) and often referred to as “enhancers”, share a number of important features. They are usually around 500 to 1000 bp in length, are located in genomic sequence near the genes they regulate, and contain one or more binding sites for each of a set of transcription factors (Carroll et al. 2001; Davidson 2001). CRMs can thus be thought of as sitting at the nexus of gene regulatory networks; they are DNA sequences which assist in translating a combinatorial code of transcription factor inputs into a specific gene expression output.

While little is understood about the evolutionary processes affecting CRMs, studies have observed that they undergo stabilizing selection, with maintenance of the overall set of TF inputs and resulting expression pattern coupled with species-specific gain and loss of transcription factor binding sites, a process known as “turnover” (Ludwig et al. 1998; Ludwig et al. 2000; Dermitzakis and Clark 2002; Dermitzakis et al. 2003). The degree of turnover appears to increase with evolutionary distance. However, the CRMs are under functional constraint and appear to change much more slowly than non-functional sequence. They are expected to have a degree of sequence conservation that, given sufficient evolutionary distance, is significantly higher than background and that is
related to the rate of binding site turnover, change, and/or loss of CRM function (Fickett and Wasserman 2000; Levy et al. 2001; Dermitzakis and Clark 2002; Moses et al. 2003).

Existing sequence analysis-based tools for identification of CRMs take two main approaches. The first involves inter-species comparisons designed to take advantage of evolutionary conservation of regulatory sequences, an approach termed “phylogenetic footprinting” (Tagle et al. 1988). Here, non-coding sequences conserved between two or more related species are treated as likely candidates for regulatory regions. The observed conservation arises from TF binding sites that remain stable in relative location to one another and hence form “anchors” that seed long linear sequence alignments.

Phylogenetic footprinting has routinely served as a guide to the discovery of regulatory sequences (Blackman and Meselson 1986; Gumucio et al. 1993; Vuillaumier et al. 1997; Loots et al. 2000; McGuire et al. 2000; Bergman et al. 2002). This method is often used as a first step to define the boundaries of potential regulatory sequences, and is followed by experimental studies to elucidate the spatio-temporal pattern of expression, if any, that those potential regulatory sequences might direct. Since metazoan genes generally possess complex expression patterns directed by multiple CRMs, a search for a module responsible for a particular pattern requires laborious experimental testing of all nearby candidate sequences.

Recently, several groups have designed a second set of approaches that predict CRMs by identifying clusters of potential transcription factor binding sites. Starting with a set of coordinately acting transcription factors and their experimentally described binding
motifs, some of these algorithms search the genome for sequences of roughly 500-1000 bp with uncommonly high concentrations or co-occurrences of predicted binding sites (Crowley et al. 1997; Fickett and Wasserman 2000; Frith et al. 2001; Berman et al. 2002; Halfon et al. 2002; Markstein et al. 2002; Rajewsky et al. 2002; Rebeiz et al. 2002). The presumption underlying these approaches is that binding-motif rich sequences provide evidence of TF binding and hence may direct similar expression patterns. In several cases, a subset of predictions has been experimentally verified. However, the arduous task of testing all predictions generated by these algorithms and the absence of extensive and well-characterized CRM data sets have hindered careful evaluation of the false positive and false negative rates. It is also important to point out that all tests in D. melanogaster so far have concentrated on only a small number of regulatory networks (chiefly gene expression in the early blastoderm). The general applicability of TF-binding motif clustering methods to other pathways has not yet been examined, and we may yet learn that other pathways are regulated in a manner refractory to or requiring more sophistication than site clustering-type analyses. Two obstacles preventing broader evaluation of these algorithms are that (1) binding motifs have been confidently described for only a handful of transcription factors and (2) there are few well-characterized networks of coordinately acting transcription factors that could serve as starting points for additional binding site clustering/co-occurrence studies.

Just as the availability of full genome sequence prompted the development of computational tools to comprehensively search for CRMs, the availability of the genomes of pairs of related species presents the opportunity to combine the above approaches by
introducing systematic phylogenetic comparisons into algorithms designed to decipher the regulatory code (Cliften et al. 2003; Kellis et al. 2003; Moses et al. 2003). The sequencing of *D. pseudoobscura* and *D. melanogaster*, a pair of species separated by approximately 24 million years of evolution (Russo et al. 1995), provides the tools to pursue such aims in metazoans (Adams et al. 2000);

http://www.hgsc.bcm.tmc.edu/projects/drosophila/.

We propose here an approach to identify similarly acting *cis*-regulatory modules given genome sequences for *D. melanogaster* and *D. pseudoobscura* and a set of co-regulated genes. We hypothesize that similarly acting enhancers can be identified by conserved subsequence signatures in phylogenetic footprinted regions (PFRs). Drawing inspiration from Gibbs-sampling based motif-finding algorithms that successfully identify regulatory sequence motifs near co-regulated genes (Tavazoie et al. 1999; Hughes et al. 2000), we developed an approach to identify the most similar phylogenetic footprints near co-regulated genes in a program we call PFR-Sampler (see Figure 1a). As a first step, we assemble a list of PFRs. To do this, we use Avid/Vista (Dubchak et al. 2000; Mayor et al. 2000; Bray et al. 2003) to identify conserved non-coding regions across the genome, which we then collate into a genome wide set of PFRs. We treat each PFR as a potential CRM, and generate for each PFR a profile of the conserved sequences that anchor the PFR alignment. By keeping track of the conserved sequence profile with Markov chains, we can use Markov chain discrimination to assess PFR similarity (Durbin et al. 1998). We then use a sampling algorithm, PFR-Sampler, to identify a subset of maximally related PFRs among the full initial set of phylogenetic footprints near a set of previously
defined co-regulated genes. The conserved sequences which are discovered to anchor the footprinted regions are candidate TF binding sequences, while the PFRs in the subset of maximally related PFRs are candidate CRMs responsible for the observed co-regulation.

This method of CRM discovery effectively circumvents obstacles facing algorithms that aim to predict CRMs by searching for co-occurrence of TF binding motifs, in that it does not require prior knowledge of the constellation of TFs that might act in the pathway of interest, and further does not require any prior information regarding TF nucleotide binding specificities. The ability to use a combination of phylogenetic conservation and co-expression to compensate for lack of knowledge of TF constellation and binding sites has precedent in previous work (Wasserman et al. 2000), which demonstrates the feasibility of using human-rodent phylogenetic conservation together with co-expression data for genes active in skeletal muscle to identify binding sites for key transcription factors responsible for this expression.

To evaluate our algorithm, we focus on the regulatory network that coordinates blastoderm expression in early *Drosophila* embryo development. This system is very well characterized, and has been used in several other studies to evaluate the TF binding motif clustering algorithms (Fickett and Wasserman 2000; Berman et al. 2002; Rajewsky et al. 2002). Starting only with a list of co-regulated genes and genome sequences from *D. melanogaster* and *D. pseudoobscura*, we show that our approach successfully identifies PFRs that correspond to known blastoderm enhancers when blastodermally-
expressed genes are input, and we anticipate that this approach can be more broadly applied to other sets of co-regulated genes.

Extending these findings, we show that the output of the sampling algorithm can be used in genome-wide scans for similarly acting enhancers. The PFR-Sampler output set of similar PFRs can be used to train a Markov chain discrimination algorithm, which we call PFR-Searcher (see Figure 1b), and all PFRs can then be scored to evaluate the extent of similarity to the model. Again taking phylogenetic footprint regions as possible bona fide cis-regulatory modules, we show that this approach performs well in recognizing additional known blastodermal CRMs and predicting new CRMs throughout the genome.
Figure 1. Overview of PFR-Sampler and PFR-Searcher methods. 

*a.* PFR-Sampler overview. A set of co-regulated genes, together with the conserved sequence profiles (see Methods) for all genomic PFRs are input into PFR-Sampler, which identifies the set of PFRs surrounding the co-regulated genes, and identifies the subset with the most similar conserved word profiles that is also most distinct from background. 

*b.* PFR-Searcher overview. The training set of PFRs along with the PFRs to search are input into PFR-Searcher, which constructs a model from the training set, and reports a ranked list of
the scanned PFRs according to similarity to the model. Abbreviation: PFR – phylogenetic footprint region.
Algorithm:

Identification of putative regulatory regions by phylogenetic footprinting. Our algorithm for identifying putative regulatory regions starts with genome sequences for *Drosophila pseudoobscura* ([http://www.hgsc.bcm.tmc.edu/projects/drosophila/](http://www.hgsc.bcm.tmc.edu/projects/drosophila/)) and *D. melanogaster* and successively analyzes them for conserved noncoding sequences (CNS), conserved noncoding subsequences (CNSS), and phylogenetically footprinted regions (PFRs), which later are analyzed using Markov chain discrimination algorithm. The genome sequences were first aligned by Avid global alignment software (Bray et al. 2003) available online from the Berkeley Genome Pipeline (version: 8 July 2003; [http://pipeline.lbl.gov/pseudo/](http://pipeline.lbl.gov/pseudo/)). Non-coding sequence was extracted from these aligned sequences using Release 3.1 annotations ([http://www.fruitfly.org/annot/release3.html](http://www.fruitfly.org/annot/release3.html)). From the resulting non-coding sequence alignments, using Vista software (Dubchak et al. 2000; Mayor et al. 2000), we then identified CNSes as groups of aligned non-coding sequences with greater than 60% conservation over 100 nt, which are separated by less than 100 nt. These parameters were estimated by an assessment of conservation and distances between aligned non-coding sequences in a few experimentally defined enhancers. By examining the sequence alignments within a CRM, CNSSes were identified as maximal contiguous stretches of aligned, conserved, ungapped sequence which contained at most 2 mismatches within an 8 bp sliding window. This parameter was modeled on the observed changes of TF binding sites between *D. melanogaster* and *D. pseudoobscura* and the average nucleotide difference between two TF binding sites for the same TF in *melanogaster* (Ludwig et al. 1998; Ludwig et al. 2000; Dermitzakis et al. 2003). Note that CNSes are stretches of sequence alignment strings, each position of
which pairs either a base code A, C, G, T, or ‘-’ (a gap) from *D. melanogaster* against a base code or gap from *D. pseudoobscura*, and that CNSSes are constituent blocks of aligned nucleotides (no gaps) within a given CNS. Henceforth, we consider only the *D. melanogaster* sequence corresponding to these aligned sequence stretches.

CNSSes, which are ungapped, are analyzed in setting up our Markov chain algorithm: each CNSS string $s_1s_2...s_n$, is considered as a sequence of $n - 5$ 6-nucleotide (nt) windows $s_is_{i+1}s_{i+2}s_{i+3}s_{i+4}s_{i+5}$, and each such window is considered a “state transition” from the prefix 5-word $s_is_{i+1}s_{i+2}s_{i+3}s_{i+4}$ to the suffix 5-word $s_{i+1}s_{i+2}s_{i+3}s_{i+4}s_{i+5}$. Any CNS with over 300 state-transitions was identified as a PFR, a threshold we chose because it reflects the length of archetypal enhancers. Since a CNS may be composed of multiple CNSSes, separated by gaps or by stretches of sequences that fall below the threshold of 2 mismatches in an 8 bp window, we note that these 300 state-transitions need not be contiguous. A PFR, then, is a CNS that is composed of clustered CNSSes containing, in total, more than 300 state-transitions.

**PFR-Searcher: Markov chain discrimination algorithm.** A maximum likelihood (ML) Markov model is generated from the PFRs by computing a frequency table of transitions from all possible strings of five base codes (PFR model), and a ML model is generated by the same method for a set of randomly chosen PFRs (background model). As described above, we encode a first order Markov chain using the alphabet of all 5-mers (mathematically equivalent to a fifth order Markov chain). In an effort to improve signal to noise by diminishing the impact of sequences shared among PFRs throughout the
genome, we used all the PFRs identified on chromosome arm 2R as the background model. The Markov chain discrimination algorithm then computes a score, measured in bits, describing the similarity of a sequence to a model, as compared to the background model (Durbin et al. 1998). Ideally, ML estimators for a given set of state-transition probabilities should be constructed solely from observed frequencies, but this requires an extensive sampling of all transitions. However, PFR datasets generally yield only sparse estimates of transition counts. While pseudocounts are often used to overcome sparse estimates, we adjusted counts in a more biologically meaningful way: as binding sites for TFs generally are similar to a consensus sequence, and the 6mers considered in state-transitions represent possible TF binding sites, we adjusted frequency counts for any given transition with all sequence transitions within a Hamming distance of one nucleotide:

$$c_{st}^+ = c_{st} + \sum_{s_i, d(s-s_i) = 1} c_{s_i} \cdot w$$

Here, $c_{st}$ indicates the counts of state-transition from word $s$ to word $t$ is observed in the training set sequences, $w$ is the weight by which all the neighboring counts are multiplied, and $c_{st}^+$ represents the resulting consensus adjusted counts. By these means, many state-transitions that have an actual count of 0 in the PFR data may be assigned non-zero adjusted counts based on a biologically supported similarity with other transitions that actually did occur. Weights used in this work were 0.25. The ML estimators were then determined as follows:
\[ a_{st}^+ = \frac{c_{st}^+}{\sum_{s'} c_{s't}^+} \]

Where \( a_{st}^+ \) is the maximum likelihood estimator for the \( s \) to \( t \) transition in the training set.

Using this method, ML estimators were computed from state-transitions in PFRs to define the PFR model, and from those in the background to define the background model. The Markov chain discrimination score between the PFR and background models was then calculated for any sequence \( x \) as the log-odds ratio

\[
S(x) = \log \frac{P(x \mid \text{model}^+)}{P(x \mid \text{model}^-)} = \sum_{i=1}^{L} \log \frac{a_{s_i,s_{i+1}}^+}{a_{s_i,s_{i+1}}^-}
\]

for a query sequence with \( L \) transitions.

**PFR-Sampler: Sampling algorithm.** The aim of the sampling algorithm is to identify a subset of PFRs with the most closely related subsequence profiles. We begin with an input set of genes \( G = \{g_1, g_2, \ldots, g_n\} \). Each gene \( g_i \) is associated with a set of nearby PFRs (proximity is user-defined), denoted \( g_{i,1}, \ldots, g_{i,j} \). In these studies, all PFRs within 50,000 bp from gene boundaries were included. Where the 50 kb overlap two genes included in the analysis, PFRs in the overlapping region were assigned to the locus of the closest gene. The total set of all PFRs associated with genes in \( G \) is \( A \). The object of this algorithm is to find the subset of \( A \), termed \( S \), such that when the members of \( S \) are used as a training model for the Markov chain discrimination test (using the same formulation described above for the PFR-Searcher algorithm), the sum of their scores is maximal. A
maximum score indicates that the PFRs in set $S$ possess the most closely related subsequences profiles that are the most distinct from background of any possible subset of $A$.

Simulated annealing is used to avoid local minima and thus enable the sampling to search for the set of PFRs with maximum score (Press et al. 1992). The algorithm begins by initializing the set $S$ to a randomly selected assortment of PFRs from $A$. The number of initial elements is set by the user-defined parameter “-p”, and represents the number of similar PFRs the user expects to discover. In this study, the number of expected PFRs was set as equal to the number of input genes. Once the algorithm is underway, this number is allowed to fluctuate. Additionally, we include a requirement that each gene contribute no more than two PFRs to the set $S$ to ensure that the set is composed of contributions from multiple gene loci and that it is not overly biased by the sequence composition of a given locus. The program then cycles through each gene $g_i$, and evaluates the score of the current set $S$, and the scores under three additional models, including swapping PFRs, removal of a PFR, and inclusion of an additional PFR. The score of the highest scoring alternative model is compared to the score of the current model. If the best alternative model’s score is greater than that of the current model, $S$ is updated to the alternative. Otherwise, $S$ is updated to the lower-scoring alternative based on a probability determined from the scoring difference between the alternative and current models ($\Delta S$) and the simulated annealing temperature schedule, where the “temperature” is given below by $T$. 

14
\[ p = \exp\left( \frac{\Delta S}{T} \right) \]

The initial $T$ used in these studies is 20, with a schedule to decrease by 95% after each cycle through $G$. The algorithm halts when either no change is made to the set $S$ over the course of several cycles through $G$, or when either 50 updates or 30 cycles are completed. We note that simulated annealing is a stochastic optimization algorithm, and results may vary. Results are dependent on the “temperature” schedule and random number seed used. Varying schedules may alter the output by giving more or less opportunity to move towards the global optimum, and, as is standard for simulated annealing protocols, we encourage trials with varying parameters.

Software: The PFR-Sampler and PFR-Searcher programs were written in C, and use auxiliary modules written in C and PERL. All programs and instructions on their use are available at our website, [http://arep.med.harvard.edu/enhancers/](http://arep.med.harvard.edu/enhancers/).

Cross validation. We employed “leave-one-out” cross validation to assess the results of the sampling algorithm. In this procedure, a set of $n$ PFRs (members of set $S$ of the sampling algorithm) is used to generate $n$ sets of PFRs in which a different single PFR is left out. For each of the $n$ sets, the model used in the Markov chain discrimination algorithm is the set itself, and scores are determined for the left-out PFR and each of a test set of 1000 PFRs randomly selected from the genome. The rank of the PFR within the set of the 1000 randomly selected PFRs is an indication of its similarity to the training set.
Results:

Phylogenetic footprinting

A general and commonly mentioned architectural feature of cis-regulatory sequences is that they are often, though not always, separable and non-overlapping. This feature, and the evolutionary conservation of the sequences due to functional constraint, has led to the common use of conserved non-coding sequence in searching for CRMs. Unanswered questions remain about the appropriate evolutionary distance that maximizes the signal-to-noise ratio for picking out meaningfully conserved sequences and about the applicability of only a single species-pair comparison to identify all CRMs, since CRMs may evolve at different rates. Still, there are a constantly growing number of reported successes of CRM discovery using a range of species pairs, including mouse-human, e.g. (Loots et al. 2000) and D. pseudoobscura-D. melanogaster (Bergman et al. 2002), thus making genome scale analysis of CRM conservation in these species pairs at least a reasonable, if not rigorously established, venture. In addition, simulation studies are beginning to explore and construct solid foundations for identification of functionally constrained non-coding sequences (Pollard et al. 2004).

Here, we focus on the D. pseudoobscura-D. melanogaster pair. To assemble a complete set of genomic conserved non-coding sequences, we obtained global alignments of these two genomes, extracted non-coding aligned sequence, and identified conserved sequence and subsequence regions that we considered phylogenetically footprinted regions (PFRs) (see Methods). Using these procedures, we generated a genomic complement of 24,651
PFRs, with average length of 817 bp, average number of conserved nucleotides equal to 640 bp, and average of 506 state-transitions.

*PFR-Sampler uncovers PFRs correlating to cis-regulatory modules that regulate blastoderm expression*

One of the best studied developmental networks in metazoans is that of blastoderm development in *D. melanogaster*. With many gene loci closely studied and a number of characterized regulatory sequences known to direct blastodermal expression, it is the most attractive system with which to evaluate our hypothesis that similarly acting CRMs can be identified by conserved subsequence signatures in PFRs.

We examined the overlap of our PFRs with enhancers known for a set of blastodermal development genes (Figure 2, Supplemental Figure 1). The locations of CRMs are from the following references: Berman et al. 2002; Rajewsky et al. 2002; Lifanov et al. 2003. We observe that the vast majority of CRMs overlap coordinates of phylogenetic footprints (27/30; see Figure 2 and Supplemental Figure 1). Of the three that do not appear in the set of PFRs, the *oc* (also called *otd*) early enhancer and the *runt* stripe 5 enhancer both do not appear to be well conserved over an extended region; and the *runt* stripe 1+7 enhancer (which is adjacent to the *runt* stripe 5 enhancer, and represented together with the stripe 5 enhancer as one single regulatory region in the figures) appears to be below the conservation threshold used to generate the set of PFRs. The low conservation of *runt* enhancers has been observed in nearby regions of this locus, offering a precedent for the apparent low conservation seen here (Wolff et al. 1999). In cases of
multiple contiguous or overlapping CRMs, we observed two outcomes. First, the contiguous block of *hairy* stripe 3+4, stripe 7, and stripe 6+2 enhancers overlaps an area of high conservation which is split into two conserved non-coding regions by our phylogenetic footprinting method. The first PFR contains the stripe 3+4 enhancer and part of stripe 7 enhancer, and the second contains an additional part of the stripe 7 enhancer and stripe 6+2 enhancer. Similarly, while a conserved non-coding region appears within the boundaries of the *tll* PD enhancer, the abutting AD-PD enhancer bridges this conserved non-coding block and two others, making unclear the functional relationship between the conserved non-coding sequences and the enhancers. Second, the *eve* stripe 4+6 and 1+5 enhancers are situated in long stretches of highly conserved regions. Studies have shown that densely packed, often overlapping CRMs populate this sequence (Sackerson et al. 1999). In both cases, the PFRs containing the sequence of the two sets of stripe enhancers are far longer than the experimentally defined minimal enhancers themselves; the stripe 4+6 enhancer is an approximately 600 bp enhancer embedded within a phylogenetically footprinted sequence of length ~4kb, and stripe 1+5 is jointly about 1700 bp within a stretch of conserved sequence of roughly 2600 bp. As we see here and has been previously reported in Bergman et al. (2002), many enhancers have a convenient one-to-one correspondence with PFRs, but there are a few exceptions, including the absence of real enhancers from the PFR set, and the conjoining of multiple true and differing enhancers, particular in the case of long (greater than 2.5 kb) PFRs.
Figure 2: Examples of results from the phylogenetic footprinting procedure for loci surrounding 19 blastodermally expressed genes (Abbreviations: *abdominal-A* – *abd-A*; *fushi tarazu* – *ftz*; *hunchback* – *hb*; *Kruppel* – *Kr*; *ultrabithorax* – *Ubx*). The loci extend 50 kb flanking the start and stopping site of each gene, or to a midway point between two adjacent genes. Fuchsia arrows designate exons, blue triangles indicate known blastodermal enhancers, and green semicircles indicate PFRs. Figure generated using *gff2ps* (Abril and Guigo 2000). For full results, see Supplemental Figure 1.
We ran our CRM discovery algorithm, PFR-Sampler (see Methods), on loci surrounding genes with characterized expression in the early blastoderm. The algorithm takes the approach of sampling various subsets of PFRs to identify those out of a large set which bear most resemblance to each other, and uses leave-one-out cross validation to assess the outcome. In leave-one-out cross validation (see Methods for details), each PFR is serially removed; its score against the training set of remaining PFRs is calculated; and the rank of this score as compared to 1000 randomly selected PFRs is determined. The average rank of the left-out PFRs gives an indication of how similar the PFRs are to each other and how distinct from background.

To explore the robustness of the PFR-Sampler program, we used inputs of 10, 12, 14, 17, and 19 blastodermally expressed genes (see Table 1). In each of these inputs, the core set of 10 genes was the same, drawn from a previous TF-clustering based CRM prediction study (Berman et al. 2002). The larger inputs comprise this set of 10 plus a random selection of additional similarly expressed genes from the nine annotated in Lifanov et al. (2003). In each case, we define the locus for each gene as including 50 kb from both the start and stop sites. Where the 50 kb overlap two genes included in the analysis, PFRs in the overlapping region were assigned to the locus of the closest gene.
<table>
<thead>
<tr>
<th>Group</th>
<th>Gene locus</th>
<th>Locus length (bp)</th>
<th>PFRs (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a [set of 10]</td>
<td><strong>Abd-A</strong></td>
<td>122426</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>eve</td>
<td>101537</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>gt</td>
<td>101856</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>h</td>
<td>103280</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td><strong>hb</strong></td>
<td>106502</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>kni</td>
<td>103033</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Kr</td>
<td>102919</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>run</td>
<td>102885</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td><strong>salm</strong></td>
<td>111292</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Ubx</td>
<td>178250</td>
<td>60</td>
</tr>
<tr>
<td>b [set of 12]</td>
<td>(a) (from above)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>104206</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>prd</td>
<td>103459</td>
<td>16</td>
</tr>
<tr>
<td>c [set of 14]</td>
<td>(a) (from above)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>btd</strong></td>
<td>103385</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>ftz</td>
<td>101904</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>prd</td>
<td>103459</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>tll</td>
<td>102005</td>
<td>27</td>
</tr>
<tr>
<td>d [set of 17]</td>
<td>(a) (from above)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>btd</strong></td>
<td>103385</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>104206</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>ftz</td>
<td>101904</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>gsb</td>
<td>102314</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>oc</td>
<td>119310</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>prd</td>
<td>103459</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>tll</td>
<td>102005</td>
<td>27</td>
</tr>
<tr>
<td>e [set of 19]</td>
<td>(a) (from above)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>btd</strong></td>
<td>103385</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td><strong>Dll</strong></td>
<td>120333</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>ems</td>
<td>102765</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>104206</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>ftz</td>
<td>101904</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>gsb</td>
<td>102314</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>oc</td>
<td>119310</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>prd</td>
<td>103459</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>tll</td>
<td>102005</td>
<td>27</td>
</tr>
</tbody>
</table>
Table 1. Input sets of genes, composed of blastodermally expressed genes. (a) The core set of 10 genes, from Berman et al., 2002, which are included in all sets. (b-e) Additional sets of genes, of size 12, 14, 17, and 19 genes, respectively, randomly selected from a pool of blastodermally expressed genes (Lifanov et al., 2002). The gene loci include 50 kb upstream of the annotated start site and downstream of the annotated stop site for each gene (D. melanogaster Release 3.1 annotations; Misra et al. 2002), and the “locus length” column indicates the number of bp considered. The number of phylogenetic footprinted regions (for parameters defining PFRs, see methods) within each of these loci is reported in the PFR column.
Notably, each run using these inputs returned high scores and output that contained a high fraction of known blastodermal CRMs, indicating that the program is capable of successfully identifying CRMs from a range of inputs. The input set with the best score was set “c” (14 gene loci), which found 24 PFRs as the most similar set within the 457 total PFRs surrounding the 14 loci (Figure 3; Supplemental Figure 3). By leave-one-out analysis, the 24 PFRs had an average rank of 2.375/1000. To evaluate the significance of these results, we ran our PFR-Sampler program on 100 sets of 14 randomly selected genes, and constructed a distribution of average outcome rank. The output for set “c” (rank 2.375) places this result in the 98th percentile. Similarly, we analyzed the significance of the runs with various numbers of input genes. Each appeared at or above the 92nd percentile, with the significance level for set “c” again surpassing the others. Affirming the potential for this approach to identify similarly acting CRMs, the output for set “c” includes 17 which correspond or are in close proximity to the locations of known enhancers. See Table 2 and Supplemental Figure 2 for a summary of results from each input set. To evaluate the impact of alternative parameters for the simulated annealing optimization, we evaluated the results for set “c” using a starting temperature increased to 40, the same step-wise decrements, and an increased number of cycles (60) and updates (100) before halting. The results for set “c” were largely the same as reported for the shorter simulated annealing protocol (see Supplemental Figure 3).
<table>
<thead>
<tr>
<th>Input set</th>
<th>Num of genes</th>
<th>Score</th>
<th>Percentile</th>
<th>PFRs in output</th>
<th>Total PFRs searched</th>
<th>PFRs corresponding or near to known enhancers</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>10</td>
<td>5.63</td>
<td>92</td>
<td>19</td>
<td>343</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>12</td>
<td>3.19</td>
<td>96</td>
<td>21</td>
<td>417</td>
<td>11</td>
</tr>
<tr>
<td>c</td>
<td>14</td>
<td>2.375</td>
<td>98</td>
<td>24</td>
<td>457</td>
<td>17</td>
</tr>
<tr>
<td>d</td>
<td>17</td>
<td>4.6</td>
<td>94</td>
<td>30</td>
<td>555</td>
<td>18</td>
</tr>
<tr>
<td>e</td>
<td>19</td>
<td>8.53</td>
<td>93</td>
<td>30</td>
<td>653</td>
<td>17</td>
</tr>
</tbody>
</table>

**Table 2.** Summary of output from PFR-Sampler given the five input sets described in Table 1. The score ("average rank") is an assessment of the similarity of output, where leave-one-out cross validation is performed on each of the output PFRs, and the rank out of 1000 randomly selected genes is determined; the average rank for each of the PFRs in the output set is reported here. For each input set, percentile is determined from the distribution of scores of 100 sets of randomly selected genes.
Included in the output were PFRs for which there is no corresponding enhancer in the literature. These may represent false positives, which by chance have a short conserved sequence profile similar to the regions known to be subject to regulation by the transcription factors involved in blastoderm development. However, we note that the majority of these regions appear far from the gene’s transcriptional start site, and we hypothesize that true blastodermal CRMs residing in these locations may have been subject to less intense scrutiny and hence have eluded detection. We propose these PFRs correspond to candidates for novel blastodermal CRMs.

No PFR-Sampler run identified phylogenetic footprints underlying all of the known blastoderm CRMs regulating the input gene set. It may be that some the CRMs are subject to alternative regulation, and thus have an alternative short sequence signature, representing different transcription factor binding sites. For some characterized blastodermal CRMs, including those for gsb and ems, this is the case. The enhancers identified by PFR-Sampler are known to be responsive to gap genes, whereas the enhancers directing the blastodermal expression of gsb and ems are responsive to pair rule genes (Jones and McGinnis 1993; Li et al. 1993; Bouchard et al. 2000). In several other instances, the false negative result may be due to the restriction that at most two PFRs from a given gene locus can contribute to the results, a restriction we instituted to keep the model from being biased by a single locus’s genomic characteristics. We explored this as a possible source of false negatives, as described below.
Genome wide search for blastoderm CRMs based on sampling-derived model

Since the output of PFR-Sampler is a set of PFRs, we can combine their short sequence profiles to create a model, and we can compare any other phylogenetic footprint to this model to determine the extent of similarity. Effectively, this method provides a tool with which to perform a genome-wide search for similarly acting CRMs, and to move from a set of co-regulated genes and a pair of genomes to a genome wide catalogue of candidate CRMs.

To perform this search, we took the output from the PFR-Sampler run described above for the set of 14 blastodermal genes, and used our PFR-Searcher program, based on a Markov chain discrimination algorithm (see Methods). Each PFR in the genome gets assigned a score based on its similarity to the model as compared to background. The higher the score, the more the PFR region resembles the training set. By assuming the error from the leave-one-out analyses generalizes to the complete set, we can derive a threshold score to achieve a given sensitivity. Since 80% of the left-out PFRs were at rank 2/1000 or better and 88% at rank 4/1000 or better, we elected to take as a threshold the midway point of the score of the 3rd ranked PFR out of 1000. Querying the complete genome and selecting those with scores above the set point yielded 207 PFRs, of which 24 are the input PFRs (see Supplemental Table 1; http://arep.med.harvard.edu/enhancers/suptab1.xls).

An indirect method of evaluating candidate CRMs consists of checking whether genes flanking the candidate genomic regions are expressed in early blastoderm development.
While reporter construct assays provide stronger evidence, cross-referencing with in situ and literature databases offers a reasonable and rapid first-pass assessment. We surveyed two sources. First, we identified genes annotated in FlyBase (Consortium 2003) that are expressed in blastoderm development in the segmentation pathway. We also checked Release 2 of the BDGP in situ images database (P. Tomancak et al., manuscript in preparation) for staining during embryonic stages 4-6 in a spatial pattern consistent with segmentation. For candidate CRMs located in introns, we checked the expression patterns of the overlapping gene and neighboring upstream and downstream genes. For intergenic candidate CRMs, we checked the adjacent flanking genes, two on each side. Of the 207 PFRs with scores above our set point, we found relevant expression information from genes flanking 107 of the predictions. So far, 79 PFRs predicted to correspond to CRMs that drive expression in the early blastoderm are near at least one gene that is expressed in the expected spatio-temporal pattern. Using these numbers as a rough guide, and assuming that each predicted CRM near a blastodermal gene is a true CRM, we calculate that the combined PFR-Sampler/PFR-Searcher method for identification of blastodermal CRMs has a hit rate of ~74% and thus a best-case false positive rate of ~26%. To get a sense of how significant these results are, we generated five sets of 207 PFRs randomly selected from across the genome, and subjected them to the same classification procedure. In these five randomized samples, the mean hit rate of PFRs was 37% ± 2%, implying that our results, with a 74% hit rate, is considerably better than random.
We speculated above that the output of the PFR-Sampler runs might miss PFRs corresponding to true CRMs due to the restriction that at most two PFRs from each input gene locus can contribute to the final output set. Our PFR-Searcher program provides an opportunity to test this hypothesis, by evaluating whether the false negative PFRs possess word profiles similar to the PFR-Sampler output. We see in Figure 3 and Supplemental Figure 3 that seven additional PFRs corresponding to known blastodermal enhancers turn up positive by the PFR-Searcher run, thus supporting our suspected cause of false negatives and lending encouragement that other candidates across the genome also correspond to true CRMs.

Beyond those seven, additional candidate CRMs are near genes known to be blastodermally expressed such as cad, cas, dpn, nub, odd, opa, pros, rpr, slp1, slp2, tsh, and wg. For several of these genes, two or more regions scoring above the threshold were nearby. The set of candidates also includes CRMs near genes shown in the BDGP database to have blastodermal stripe-like expression such as CG1815, CG10176, CG33207, CG17383, and CG9924.
Figure 3: Several examples of blastodermally expressed genes denoted with results of PFR-Searcher, using the output from PFR-Sampler with input set c. Fuchsia arrows designate exons, blue triangles indicate known blastodermal enhancers (Lifanov et al. 2003), green semicircles indicate PFRs, red semicircles indicate PFRs comprising the output of the PFR-Sampler run, and yellow semicircles indicate PFRs with scores above threshold as determined by PFR-Searcher. Figure generated using gff2ps (Abril and Guigo 2000). Gene name abbreviations as in Figure 2. For full results, see Supplemental Figure 3.
State-transition composition of the model

The premise of the algorithms presented here relies on the correspondence between conserved sub-sequence within phylogenetic footprinted regions and transcription factor binding sites. We therefore sought to explore how well the model derived from the set of 14 blastodermal genes, which was used with PFR-Searcher as discussed above, corresponds to known transcription factor binding sites. In essence, the question we ask is whether the state transitions that contribute most to the identification of a region as being similar to the model overlap with known transcription factor binding sites.

We first identified the log-likelihood score associated with each of the 4096 state transitions that comprise the model (see Supplemental Table 2). This score is an indication of the likelihood that a given state-transition derives from the model or from background, and is the foundation of the algorithm’s discriminative power. The score for each region then is a combination of the state transitions present, their frequency, and their log-likelihood score (see Methods). We chose the PFR corresponding to the eve stripe 2 PFR as a case study because the binding sites within this region have been well characterized (Ludwig et al. 1998). The enhancer region is roughly ~800 bp, and lies mostly in the 5’ region of the 1600 bp phylogenetic footprint, which presumably also includes the promoter in the 3’, gene proximal, region. By multiplying each state transition’s frequency, including that of its reverse complement, by its score as set by the model, we can rank the contributions of each of this PFR’s state transitions according to how much each state transition led to identifying this PFR as similar to the model (a
combination of the state-transition’s score given the model and background and its frequency in the region under evaluation; see Methods and Supplemental Table 3).

We then mapped the top 15 state transition sequences and their reverse complements on a ClustalW alignment (Thompson et al. 1994) of the stripe 2 element from *D. melanogaster, D. pseudoobscura, D. erecta*, and *D. yakuba* (see Supplemental Figure 3). We find that a number of these state transitions are significantly contained (*p* = 0.0006 by hypergeometric distribution) within known binding sites for bicoid, Kruppel, hunchback, and sloppy paired 1 (see Figure 4, Supplemental Figure 4; Andrioli et al. 2002; Ludwig et al. 1998). That known binding sites for segmentation-related transcription factors are identified among the top 15 state transitions supports the approach taken by the algorithms described here. We see that some of these state transitions appear in unconserved locations in the sequence of all four species, suggesting the possibility of site turnover, and consistent with recent results (see Figure 4a, 4b; Dermitzakis and Clark 2002). We further observe that some of the state transitions that occur within known binding sites appear as well in highly conserved locations where no binding site has been characterized (see Figure 4b).
Figure 4. Top-scoring state transitions and their phylogenetic conservation in the PFR overlapping the *eve* stripe 2 element. Sequence from the stripe 2 element in *D. melanogaster* (mel), *D. erecta* (ere), *D. yakuba* (yak), and *D. pseudoobscura* (pse) were aligned by ClustalW (Thompson et al. 1994), and binding site information collated from Ludwig et al., 1998. Asterisks indicate columns of complete conservation in all four species. Regular expression matching identified locations of state transitions within the stripe 2 element sequence. 

**a.** Sites for state transition equivalent to six-mer GGGTTA (rank 14) and its reverse complement TAACCC appear within known stripe 2 Krüppel binding sites Kr-6, Kr-2, and Kr-1, along with two sites in the *D. pseudoobscura* stripe 2 region that are not well conserved in *D. melanogaster.* 

**b.** Sites for state transition equivalent to six-mer ATAATC (rank 8) and its reverse complement GATTAT appear within known stripe 2 Bicoid binding sites *bcd*-4 and *bcd*-3, as well as in a very well conserved block of sequence in the 3’ region of the stripe 2 enhancer. See Supplemental Table 3 for list of top scoring state transitions, and Supplemental Figure 4 for their mapping to the stripe 2 element sequence.
Discussion:

We model CRMs as functionally modular and separable sequences defined by the clusters of multiple transcription factor binding sites that populate the sequence. As the overall function of an enhancer is subject to stabilizing selection, in which a subset of transcription factor binding sites are evolutionarily conserved, we treat the conserved subsequences within phylogenetic footprints as closely approximating the distinguishing sequence profile of enhancers. Our implementation of algorithms to identify similar enhancers therefore employs phylogenetic footprinting as a way to define borders of enhancers and keys on the short, evolutionarily conserved words anchoring these footprints. This approach effectively circumvents the need for prior knowledge about both the constellation of transcription factors acting in concert at a given CRM, and the binding motifs of these transcription factors.

Our approach does not require that all phylogenetic footprints represent true enhancers; it is likely that the set of genomic PFRs includes many non-CRM sequences, from coding sequences to areas of randomly observed conservation. However, our assumption that the set of genomic phylogenetic footprints includes many corresponding to true CRMs is supported by the correlation observed between PFRs and early blastoderm enhancers. For the purposes of the algorithms presented here, the signal from PFRs corresponding to CRMs can overcome the noise generated by non-CRM PFRs.

We designed two programs for CRM identification based on looking for similarity of phylogenetic footprint word profiles. In the first, PFR-Sampler, we identify CRMs by
searching the phylogenetic footprints surrounding co-regulated genes for the subset of the most similar PFRs. The second, PFR-Searcher, constructs a model based on a set of input phylogenetic footprints, and then scores other PFRs for their similarity to the model as compared to background.

We show in this study the feasibility of this approach. When provided with an input of 10 to 19 blastodermally expressed genes and phylogenetic footprints from comparison of \textit{D. melanogaster} and \textit{D. pseudoobscura}, PFR-Sampler successfully identifies a subset of phylogenetic footprints that correspond to known blastodermal enhancers. Using the output PFRs from PFR-Sampler as the input training set for PFR-Searcher, the program located many known and candidate blastodermal enhancers, multiple of which are known to have \textit{in vivo} function. We estimate a false positive rate of around 26%, which compares favorably to other computational CRM prediction studies. However, we believe it important to point out that computational evaluations employed here and in the other studies are a reasonable first pass, but, as demonstrated in our previous work (Halfon et al. 2002), should not be construed as definitive for any specific enhancer prediction or as a replacement for experimental validation.

In our method we use AVID alignments of \textit{D. melanogaster}/\textit{D. pseudoobscura} sequences, following the precedent set in the global analysis by Bergman et al. (2002). Benchmarking tests done by Bergman and Kreitman (2001) for study of \textit{D. melanogaster} – \textit{D. virilis} alignments and by Pollard et al. (2004) on simulated data suggest that global alignment tools such as AVID and DiAlign perform well in the task of identifying
conserved blocks of non-coding sequences; however, as shown by Pollard et al. (2004), alignment tools have a range of success in detecting sequences subject to differing types of conservation pressure. In the CRM-prediction method proposed here, varying the alignment algorithm may impact two key junctures: (1) generation of PFRs; and (2) profiling of conserved subsequence composition of individual PFRs. Since PFRs are composed of multiple conserved blocks of sequence within proximity of one another, the characteristics of PFRs generated by an alignment algorithm depend not only on the extent of alignment and the distribution of lengths for conserved sequence blocks, but also on how the conserved subsequences cluster. Alternative PFR boundaries result in a de facto difference in Markov state transition profiles, since the same conserved subsequences used to define the PFR also underlie the state transition profile. One avenue for further study, then, will be benchmarking of alignment tools with respect to the boundaries of clusters of constrained sequences.

Our programs build on previous approaches to the problem of predicting cis-regulatory modules, combining studies to delimit sequence space in which CRMs may reside with models of CRMs as comprised of clusters of binding sites. Wasserman et al. (2000) demonstrated the feasibility of using phylogenetic footprints between human and rodent and a set of co-regulated genes to uncover relevant transcription factor binding sites. They noted that nearly all known binding sites were localized to conserved blocks, and that a Gibbs sampling algorithm successfully identified several known transcription factor binding motifs. However, they considered only short sequence stretches upstream of coding sequences, a limitation that prevents broad applicability of their approach, since
CRMs may exist many kilobases upstream or downstream of the transcriptional start site. The relationship between clusters of conserved non-coding sequences and cis-regulatory sequence has been noted in *D. melangoaster – D. pseudoobscura* alignments by Bergman et al. (2002), who demonstrated experimentally that a cluster of conserved non-coding sequences, what we refer to here as a PFR, near the gene *apterous* functions as a CRM. The link between clusters of non-coding sequences and CRMs, fundamental to the work presented in our study, provides a means by which to identify CRMs located beyond the promoter region. What this link does not establish, however, is the relationship between a CRM and the expression pattern it drives. The approach we present in this paper advances these avenues of research by providing a method to generate precisely this relationship and, using the principle highlighted in Wasserman et al. (2000), to identify key transcription factor binding sequences.

Several studies have endeavored to specifically predict early blastoderm enhancers on the basis of transcription factor binding motif clustering, using binding sites for *bcd, hb, Kr, kni*, and *cad*, among others. To gauge the extent of similarity of those approaches with the ones described in this study, we compared the overlap of results. One study which generated word profiles without considering evolutionary conservation predicted 146 regions (Rajewsky et al. 2002). There is only a small degree of overlap between our set and theirs, as only 14 of their regions were near our predictions. Another study, which searched for high density clusters of transcription factor binding sites, identified 28 candidate regions (Berman et al. 2002), of which 10 are near to our predictions, a much higher extent of overlap. A number of the overlapping predictions are near genes known
to have blastodermal expression in a stripe-like pattern, offering further encouragement that the different approaches to enhancer prediction might have correctly identified true cis-regulatory regions at these locations.

One recently developed tool for CRM prediction, *Stubbs*, uses an HMM approach in conjunction with phylogenetic conservation to identify CRMs (Sinha et al. 2003). An advantage over our method conferred by the HMM approach is the incorporation by the hidden model of potentially important relationships between binding motifs in the function of an enhancer, and hence may better define the architecture underlying CRMs. This method, however, differs from ours in that it requires prior knowledge of both transcription factor binding motifs and the constellation of transcription factors acting together in the regulatory network. Hence, it is limited both by the extent and biases of previously available information. Another tool, Argos, avoids this limitation by searching for motifs overrepresented within a window as compared to a background model, without requiring motifs as input (Rajewsky et al. 2002). This method uses a shifting window to scan the genome, rather than focusing as we do on potential CRMs predefined by phylogenetic footprinting. Further, Argos does not by itself link its CRM predictions with the expression patterns they are expected to drive, although, through the logic outlined in Rajewsky et al. (2002) and similar to the approach taken here, it should theoretically be possible to do so.

Important caveats remain. Here, phylogenetic footprinting with two species is sufficient to capture relevant information. However, it is likely that study of other regulatory
networks will require the addition of other species’ genomes. Multiple species comparisons will greatly assist in elucidating potential regulatory regions and resolving key sequences to which transcription factors may bind. More fundamentally, the model used in these algorithms rewards words that are over-represented in the training set as compared to background, and the frequency of appearance determines each word’s contribution to the overall PFR score. Such a scoring procedure presumes a model in which CRMs are densely packed with many binding sites for each transcription factor. While this appears to be the case for at least a subset of blastodermal enhancers, it is not clear that all enhancers work using the same model.

Conclusions:

This study proposes and gives preliminary results in support of a method that employs a current model of cis-regulatory module composition and evolution toward the goal of predicting CRMs. We show that analysis of phylogenetic footprinted regions surrounding co-regulated genes can identify regions sharing the most similar subsequence profiles, that these profiles can represent a functional link between sequences and regulatory activity, and that such analysis can be used to correlate regulatory sequence and the expression pattern it directs.

The approach is predicated on a number of tools and models that further development and study will continue to improve. Included among these are algorithms for genome alignment, phylogenetic footprinting, and comparisons of short sequence profiles; and models for rates and types of constraint within CRMs, for CRM composition, and for
CRM architecture. Advances along these lines in simulations, tools, and modeling will be critical in deciphering the logic behind how a given regulatory network works to coordinate the transcription of multiple genes.
List of abbreviations:

bp: base pair

CNS: conserved non-coding sequence

CNSS: conserved non-coding subsequence

CRM: cis-regulatory module

nt: nucleotide

PFR: phylogenetic footprint region

TF: transcription factor
**Acknowledgements:** The authors would like to thank John Aach, Jason Comander, and Matt Wright for critical readings of the manuscript; John Aach, Sung Choe, Jason Comander, Vidyasagar Koduri, Anthony Phillipakis, Jay Shendure, and Matt Wright for valuable discussions.
Figure and Table Legends:

Table 1. Input sets of genes, composed of blastodermally expressed genes. (a) The core set of 10 genes, from Berman et al., 2002, which are included in all sets. (b-e) Additional sets of genes, of size 12, 14, 17, and 19 genes, respectively, randomly selected from a pool of blastodermally expressed genes (Lifanov et al., 2003). The gene loci include 50 kb upstream of the annotated start site and downstream of the annotated stop site for each gene (D. melanogaster Release 3.1 annotations; Misra et al. 2002), and the “locus length” column indicates the number of bp considered. The number of phylogenetic footprinted regions (for parameters defining PFRs, see methods) within each of these loci is reported in the PFR column.

Table 2. Summary of output from PFR-Sampler given the five input sets described in Table 1. The score (“average rank”) is an assessment of the similarity of output, where leave-one-out cross validation is performed on each of the output PFRs, and the rank out of 1000 randomly selected genes is determined; the average rank for each of the PFRs in the output set is reported here. For each input set, percentile is determined from the distribution of scores of 100 sets of randomly selected genes.

Supplemental Table 1. Results from PFR-Searcher scan of PFRs in D. melanogaster genome using the PFR-Sampler output from set “c” (described in text; see Figure 2 and Table 2). Input PFRs are designated in red. Gene names in bold indicate that expression data is available for these genes. If the gene name is underlined, expression data was garnered from FlyBase; else expression data was gleaned from the BDGP in situ
database. The rightmost four columns indicate whether the gene is expressed in the early blastoderm in a segmentation-related pattern. A “y” indicates a positive; “n” negative; “?” unclear; and an “x” indicates that there are contradictory results between FlyBase annotation and the in situ images in the BDGP database.

**Supplemental Table 2.** Log-likelihood scores for all 4096 state transitions as used in model of 24 PFRs derived from PFR-Sampler output from set “c”. State transitions are represented here as 6-mers. Scores for each state transition are given in bits. A positive score indicates that the state transition is more likely to have come from the model, and a negative score indicates from background.

**Supplemental Table 3.** List of all 4096 state transitions (equivalent to all six mers) ranked according to their contribution to the score of the PFR overlapping the eve stripe 2 element. The frequency of appearance of a state transition – its counts – multiplied by the log-likelihood score give its contribution to the overall score of the PFR. Overlap with known and characterized transcription factors is also reported; for mapping back to the stripe 2 element, see Supplemental Figure 3.

**Figure 1.** Overview of PFR-Sampler and PFR-Searcher methods. *a.* PFR-Sampler overview. A set of co-regulated genes, together with the conserved sequence profiles (see Methods) for all genomic PFRs are input into PFR-Sampler, which identifies the set of PFRs surrounding the co-regulated genes, and identifies the subset with the most
similar conserved word profiles that is also most distinct from background.  

b. PFR-Searcher overview. The training set of PFRs along with the PFRs to search are input into PFR-Searcher, which constructs a model from the training set, and reports a ranked list of the scanned PFRs according to similarity to the model. Abbreviation: PFR – phylogenetic footprinted region.

**Figure 2:** Examples of results from the phylogenetic footprinting procedure for loci surrounding 19 blastodermally expressed genes (Abbreviations: *abdominal-A* – *abd-A; fushi tarazu* – *ftz; hunchback* – *hb; Kruppel* – *Kr; ultrabithorax* – *Ubx*). The loci extend 50 kb flanking the start and stopping site of each gene, or to a midway point between two adjacent genes. Fuchsia arrows designate exons, blue triangles indicate known blastodermal enhancers, and green semicircles indicate PFRs. Figure generated using **gff2ps** (Abril and Guigo 2000). For full results, see Supplemental Figure 1.

**Figure 3:** Several examples of blastodermally expressed genes denoted with results of PFR-Searcher, using the output from PFR-Sampler with input set c. Fuchsia arrows designate exons, blue triangles indicate known blastodermal enhancers (Lifanov et al. 2003), green semicircles indicate PFRs, red semicircles indicate PFRs comprising the output of the PFR-Sampler run, and yellow semicircles indicate PFRs with scores above threshold as determined by PFR-Searcher. Figure generated using **gff2ps** (Abril and Guigo 2000). Gene name abbreviations as in Figure 2. For full results, see Supplemental Figure 3.
**Figure 4.** Top-scoring state transitions and their phylogenetic conservation in the PFR overlapping the *eve* stripe 2 element. Sequence from the stripe 2 element in *D. melanogaster* (mel), *D. erecta* (ere), *D. yakuba* (yak), and *D. pseudoobscura* (pse) were aligned by ClustalW, and binding site information collated from Ludwig et al., 1998. Asterisks indicate columns of complete conservation in all four species. Regular expression matching identified locations of state transitions within the stripe 2 element sequence.  

*a.* Sites for state transition equivalent to TAACCC and its reverse complement GGGTTA appear within known Krüppel binding sites Kr-6, Kr-2, and Kr-1, along with two sites in the *D. pseudoobscura* stripe 2 region that are not well conserved in *D. melanogaster.*  

*b.* Sites for state transition equivalent to GATTAT and its reverse complement ATAATC appear within known Bicoid binding sites *bcd*-4 and *bcd*-3, as well as in a very well conserved block of sequence in the 3’ region of the stripe 2 enhancer.

*Bioinformatics*, 16(8), 743-4.


Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P.,
Huang, Y., Kaminker, J. S., Millburn, G. H., Prochnik, S. E., Smith, C. D., Tupy,
J. L., Whitfied, E. J., Bayraktaroglu, L., Berman, B. P., Bettencourt, B. R.,
Celniker, S. E., de Grey, A. D., Drysdale, R. A., Harris, N. L., Richter, J., Russo,
S., Schroeder, A. J., Shu, S. Q., Stapleton, M., Yamada, C., Ashburner, M.,
Drosophila melanogaster euchromatic genome: a systematic review.” Genome
Biol, 3(12), RESEARCH0083.

“Position specific variation in the rate of evolution in transcription factor binding
sites
Phylogenetically and spatially conserved word pairs associated with gene-expression
changes in yeasts.” BMC Evol Biol, 3(1), 19.

“Benchmarking tools for the alignment of functional noncoding DNA.” BMC


detection of genomic cis-regulatory modules applied to body patterning in the

approach to the identification of cis-regulatory modules and target genes in
whole-genome sequence data. Site clustering over random expectation.” *Proc Natl Acad Sci U S A*, 99(15), 9888-93.


conductance regulator gene reveals multiple levels of regulation.” *Biochem J*, 327(Pt 3)), 651-62.
