Tracing the Evolutionary History of Drosophila Regulatory Regions with Models that Identify Transcription Factor Binding Sites

Emmanouil T. Dermitzakis,*1 Casey M. Bergman,†2 and Andrew G. Clark*3

*Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University; and †Department of Ecology and Evolution, University of Chicago

Much of evolutionary change is mediated at the level of gene expression, yet our understanding of regulatory evolution remains unsatisfying. In light of recent data indicating that transcription factor binding sites undergo substantial turnover between species, we attempt to quantify the process of binding site turnover in regulatory regions of well-studied genes controlling embryonic patterning in Drosophila. We examine polymorphism and divergence data in Drosophila melanogaster and four related species from regulatory regions of five early development genes for which functional binding sites have been identified. This analysis reveals that Drosophila regulatory regions exhibit patterns of variation consistent with functional constraint. We develop a novel approach to binding site prediction which we use to characterize the process of binding site divergence in regulatory regions. This method uses sets of known binding sites to construct a model that predicts transcription factor specificity and bootstrap sampling to derive significance levels. This approach allows appropriate significance levels to be determined even in the face of skewed base composition in the background sequence. Using this approach, we show that, although functional elements exhibit conservation of sequence, there is substantial potential to gain new functional elements within the regulatory regions. Our results show that application of models that predict transcription factor binding sites can yield insights into the process and dynamics of binding site evolution within regulatory regions.

Introduction

Regulatory sequences have properties that make their study more difficult than that of coding sequences (Leung et al. 2000; Wasserman et al. 2000; Dermitzakis and Clark 2002). There are no straightforward properties in regulatory sequences analogous to the open reading frame and codons in coding sequences, making it difficult to define the position, amount, and strength of selective constraints on functional regulatory elements. In addition, the model of transcriptional regulation is not a simple one of activation or suppression by transcription factors, but rather, it includes competitive binding of proteins (Small et al. 1991), cooperative binding (Burz et al. 1998; Zhao et al. 2000), chromatin bending, and other molecular interactions that are not always reflected in the nucleotide sequence.

Understanding the evolutionary processes that regulatory sequences undergo will substantially improve our understanding of their functional constraints. Numerous comparative methods (termed “phylogenetic footprinting”) have been developed that use sequence conservation to infer function (reviewed in Hardison et al. 1997 and Hardison 2000; Miller 2001). It may be problematic, however, to assume that all functional sequences are conserved and all nonfunctional sequences have diverged. For instance, sequence comparisons of closely related species produce many false positive results, and comparison of distantly related species has low power to detect functional elements in the species compared (Dermitzakis and Clark 2002).

Moreover, it has been shown that regulatory sequences can maintain regulatory function despite structural reorganization as a result of species-specific loss and gain of transcription factor binding sites (Piano et al. 1999; Ludwig et al. 2000; Cuadrado, Sacristán, and Antequera 2001).

Given that the binding site is one of the fundamental units of regulatory structure, it is also likely to be one of the fundamental units of regulatory evolution. Therefore, methods that can predict the binding site composition of a sequence should provide a powerful means for quantifying regulatory sequence divergence in comparative analyses. Application of predictive tools such as probability weight matrices (PWMs) has been useful in the annotation of regulatory regions (Stormo 2000; Berman et al. 2002), and such tools have been shown to be useful for predicting the evolution of regulatory regions as well (Chuzhanova et al. 2000; Liu, Wu, and He 2000; Ludwig et al. 2000). Despite the preliminary reports, more research is necessary to determine which approaches to binding site prediction will be most useful for understanding the process of binding site turnover in regulatory regions.

Most methods for binding site prediction consider the background sequence as “null” and aim to identify a short sequence that has an outstanding match to the weight matrix (Berg and von Hippel 1987; Durbin et al. 1998; Eddy 1998). One of the main problems with such an approach is that the criteria for distinguishing a binding site from the background sequence are not well defined. For example, if the model uses nucleotide composition to generate the null hypothesis, and both the weight matrix and the sequence under study have the same nucleotide composition bias (e.g., it is A-T rich), it will be difficult to obtain a significant match to the weight matrix, because such a motif will have a high probability to occur by chance in the given sequence. Comparison of short sequences with the background sequence is also not biologically meaningful, because binding affinity does not depend on how different a binding site is from the background sequence, and the same sequence segment can be functional in many different
sequence environments. Therefore, weighing the significance of predicted binding sites relative to the properties of the background sequence may not be the most powerful approach to binding site prediction.

In this report we quantitatively describe the evolutionary changes that occur in regulatory sequences of several well-studied early development genes in Drosophila: the proximal promoter of engrailed, the even-skipped stripe-2 enhancer, the “zebra element” of fushi tarazu, the proximal promoter of hunchback, and the enhancer of kniirs. These five genes are embedded in a regulatory network that defines the anteroposterior pattern of the Drosophila early embryo. Because of the spatial and temporal sensitivity of the system, gene regulation is crucial for proper embryonic development (Driever et al. 1989; Struhl 1989; Rivera-Pomar and Jackle 1996). We initially present a description of the pattern of polymorphism and divergence in these sequences. We then introduce an approach for binding site prediction that does not take into account the background sequence, using PWMs and bootstrap sampling to define significance levels. This approach appears to be more biologically meaningful than several methods described previously, and it is more flexible for evolutionary analysis. We use this approach to obtain quantitative measures of divergence of the above regulatory regions and to dissect the process of binding site turnover in regulatory sequences.

Methods
Sequencing and Data from GenBank

Sequencing of the regulatory regions for ftz, en and kni was done using polymerase chain reaction (PCR) products as templates (PCR primers and conditions available on request). After the PCR reaction, products were either ethanol-precipitated or column-purified. A fraction of the PCR product was used in sequencing reactions using the commercial kit from Beckman Coulter. After the sequencing reaction, products were ethanol precipitated according to the instructions in the manual and were run in a Beckman CEQ2000 8-capillary automated sequencer. Base calling was done with the Beckman software and the calls were subsequently analyzed with the option SeqMan of Lasergene. For each region we sequenced both strands at least once and in most cases more than once. Variant sites were verified with careful inspection of the trace files, after the interspecific alignment was done, to confirm that they are real. Sequences were aligned with ClustalW, and divergence was calculated using P-distance (Nei 1987).

Sequences were submitted to the GenBank under accession numbers AY184070–AY184106. We also used data for the regulatory regions of eve and hb already available in GenBank (see Ludwig and Kreitman 1995, Ludwig, Patel, and Kreitman 1998; Tautz and Nigro 1998). Table 1 summarizes the regions and the lengths of the sequence used, and figure 1 shows the functional D. melanogaster binding site sequences retrieved from the TRANSFAC database (Wingender et al. 2000) analyzed in this article, and their alignment to orthologous sequences in other Drosophila species.

The Drosophila lines used in this study are these: (1) D. melanogaster 3rd chromosome extracted lines from Central Pennsylvania, and isofemale lines from Zimbabwe; (2) D. simulans isofemale line from California; (3) D. sechellia, D. yakuba, and D. orena from the Drosophila Stock Center.

Derivation of Ancestral Sequences

Ancestral sequences were derived using the maximum likelihood program PAML (Yang 1997) based on the approach of Yang, Kumar, and Nei (1995). This approach is not reliable for alignment gaps, so we derived the ancestral sequences of alignment gaps manually, applying parsimony criteria. For this approach D. yakuba and D. orena/D. erecta were considered equally distant from D. melanogaster, because our divergence analysis did not resolve reliably the relationships of D. melanogaster, D. yakuba, and D. orena/D. erecta (see also Jeffs, Holmes, and Ashburner 1994).

Binding Site Data and PWM Estimation

To obtain a reliable PWM that will represent the specificity of the transcription factor of interest, we have used a large sample of binding sites collected from the primary literature and personal communications. The two transcription factors studied here are the products of the genes bicoid (Bcd) and hunchback (Hh), for which we have used 51 and 97 binding sites, respectively. These data were initially reported in Ludwig et al. (2000) and have subsequently been published in full by Berman et al. (2002). Sequences, genomic locations, and primary references for the binding sites used in this study can be found at http://www.fruitfly.org/~cbergman/ or

---

**Table 1**

Summary of Data Used and Results of Nucleotide Polymorphism

<table>
<thead>
<tr>
<th>Gene Regulatory Region</th>
<th>Length (bp)</th>
<th>No. of Binding Sites</th>
<th>mel</th>
<th>sim</th>
<th>sec</th>
<th>yak</th>
<th>ore</th>
<th>ere</th>
<th>θ</th>
<th>Tajima’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>eve promoter</td>
<td>660</td>
<td>3</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0035</td>
<td>–1.50</td>
</tr>
<tr>
<td>eve stripe-2 enhancer*</td>
<td>670</td>
<td>17</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>0.005</td>
</tr>
<tr>
<td>ftz zebra element</td>
<td>600</td>
<td>11</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>0.0037</td>
</tr>
<tr>
<td>Hb promoter 2b</td>
<td>347</td>
<td>6</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0051</td>
<td>0.1</td>
</tr>
<tr>
<td>kniirs enhancer</td>
<td>350</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>


---

NOTE.—mel, D. melanogaster; sim, D. simulans; sech, D. sechellia; yak, D. yakuba; ore, D. orena; ere, D. erecta.
Alignments were performed with the Web-based motif extraction method YEBIS (Yada et al. 1998), which uses a hidden Markov model to obtain the optimal alignment and extract a motif from a given set of sequences. The length of the consensus Bcd binding site is 8 nucleotides, and the length of the consensus Hb binding site is 14 nucleotides.

Binding Site Prediction

We have devised a novel method for the prediction of transcription factor binding sites within a query sequence. In contrast to the usual way of setting up such a test, the null hypothesis of this test is that a short sequence is a binding site, and we reject the hypothesis that sequences are binding sites for a given factor when they deviate from a pattern consistent with the properties of the observed set of binding sites. The rationale behind the construction of this model is explained in the Results.

Initially, we align a set of binding sites for a given transcription factor and generate a PWM, which describes the probability of each of the four nucleotides at each position of the binding site sequence (see Supplementary Data) as expressed by their observed frequency in the respective position in the alignment of functional sites. We then use this matrix to generate a value $L$ that reflects how likely it is that a given sequence with the same length as the binding site is compatible with the matrix. The value $L$ is calculated as:

$$L_j = \sum_{i=1}^{n} \log(p_i)$$

where $n$ is the length of the binding site, $j$ is the first position of window $n$ on the nucleotide query sequence, and $p_i$ is the matrix frequency in matrix position $i$ of the nucleotide in position $j+i$ of the query sequence.

To obtain the critical threshold, we use the $L$ values that the individual sites used to generate the matrix acquire (Schneider 1997). Because a PWM is derived from a sample of binding sites, the confidence we have in the weight matrix is dependent on the size of the sample of sites. To determine how sensitive the weight matrix is to the particular sample of binding sites used to generate it, we drew 1,000 bootstrap samples of the binding sites.
(sampling with replacement), and from each sample we generated a corresponding weight matrix. Because this collection of weight matrices represents the sampling variability in our confidence in a determination of \( L \), the distribution of \( L \) values across these 1,000 matrices is taken to be the null distribution of \( L \) (fig. 2). We used the bootstrap to take into account the heterogeneity of the sample of binding sites when setting the critical threshold. Any sequence that obtains an \( L \) value below the 5% cut-off of the distribution will be considered "rejected" as a binding site of the respective transcription factor, because fewer than 5% of the weight matrices accept it as such. To obtain the predicted sites within a query sequence, we calculate the \( L \) values by taking a sliding window of sequence equal to the length of the binding site and stepping one nucleotide at a time from the 5' end to the 3' end of the query sequence. We also scan the query sequence with the matrix in its reverse orientation to identify binding sites in the opposite strand.

Simulations of Divergence

The simulations of divergence are forward simulations of accumulation of substitutions. We preassigned a divergence rate and obtained, for every initial sequence, a set of 1,000 sequences of equal length that differed from the original sequence by a proportion of nucleotides equal to the divergence rate. We performed simulations with and without mutation bias. Because there was no clear substitution bias in our data, we incorporated a mutation bias for both GC and AT of 54% and 46% in our simulations. All the computer programs were written in PERL and are available from the corresponding author on request.

Results
Patterns of Nucleotide Polymorphism and Divergence in Drosophila Regulatory Sequences

We have sequenced multiple alleles of \( D. melanogaster \) for two of the regulatory regions (\( ftz \) and \( en \), immediately upstream of the transcription initiation site) and one allele from \( D. simulans, D. sechellia, D. yakuba, \) and \( D. orena \) for three regulatory regions (\( ftz \) and \( en \), immediately upstream of the transcription initiation site and \( kni \), 1.6 kb upstream of the transcription initiation site). A summary of the sequences analyzed and polymorphism results is presented in table 1. Nucleotide polymorphism and divergence was generally low in most of the regulatory regions. Low nucleotide variation is in accordance with the pattern previously observed in \( eve \) stripe-2 enhancer and the \( hb \) proximal promoter (Ludwig and Kreitman 1995; Tautz and Nigro 1998). Levels of polymorphism in these regulatory regions are similar to average levels of polymorphism in coding sequences but lower than levels of polymorphism in silent sites or presumably unconstrained noncoding regions (Moriyama and Powell 1996). Divergence was generally lower for all regulatory regions on average than previously reported for synonymous divergence in coding sequences in these species (\( K_s \); Dunn, Bielawski, and Yang 2001). As shown in table 1, calculation of the normalized difference between heterozygosity estimated from average pairwise differences (\( \pi \)) and heterozygosity from the number of segregating sites (\( \theta \)) Tajima's \( D \) indicates that there is a tendency for excess of rare polymorphisms in three of the four regulatory regions (\( en, eve, \) and \( ftz \)) with polymorphism data, which could result from purifying selection, recovery from a selective sweep, or population growth (Tajima 1989). Comparison of the spatial distribution of divergence and polymorphism along the sequence may provide some clue about the similarity in the selective constraints within and between species. We compared the patterns of polymorphism within \( D. melanogaster \) and substitution between \( D. melanogaster \) and \( D. simulans/D. sechellia \) using a modification of the randomization test described in Dermitzakis and Clark (2001). Under neutrality the expectation is that the two patterns will be random and not significantly different from each other. We applied the test in four of the five genes surveyed here for which polymorphism data were available. In no case did the spatial patterns of polymorphism and substitution differ significantly. We noticed, however, that the \( P \) values
generated from our random permutations tended to be close to 1. We decided therefore to consider the opposite null hypothesis and use the other tail of the permuted null distribution to test the hypothesis that the polymorphic sites are excessively close to the interspecific substitutions. Using this hypothesis, one of the four genes (eve) showed significant proximity (clustering) of substitutions and polymorphisms. We subsequently applied Fisher’s combined probability (Sokal and Rohlf 1997) for all four genes, and the result was significant with $P = 0.0319$. Note that the test results for the other two genes reflected a degree of clustering that was close to significant (en, ftz, proximal hh).

The significance of the clustering of polymorphic and divergent sites can be explained by the fact that these sequences contain many functional regulatory elements. If the vast majority of the sequence is constrained as a consequence of functional importance, then the remainder will only be a small fraction of the sequence, and that is where there is freedom for the accumulation of mutations and substitutions. When this pattern is compared to a completely random pattern assuming no constraint (as our permutations assume) it will show significant clustering of the two types of nucleotide change. The pattern of polymorphism and divergence suggests that the regulatory fraction of the sequence is higher than estimated by the binding site sequences characterized so far.

Many binding sites have been previously identified in these regions in D. melanogaster (TRANSFAC; Wengender et al. 2000), some of which contain polymorphisms and substitutions or indels in the other species (see figure 1 for binding site alignments). The distribution of divergence values in these binding sites between D. melanogaster and other species is shown in figure 3. In some cases the mutational events are so radical that the binding sites are most likely not functional in other species. For instance, one of the caudal sites in the ftz zebra element has obtained a two-base insertion in D. simulans, leading to a probable loss of this essential site. Application of the modified McDonald-Kreitman test for regulatory regions (Jenkins, Ortori, and Brookfield 1995; Ludwig and Kreitman 1995) did not show any significant deviation from neutral expectation. This result is consistent with the observation that polymorphic sites and substitutions tend to cluster.

Low levels of interspecific divergence and intraspecific variation, as well as an excess of rare variants, are findings consistent with the notion that these sequences are functionally constrained. Clustering of polymorphic and divergent sites, as well as proportional levels of polymorphism and divergence in functional binding site sequences, suggests that functional constraints on these sequences are similar in these species. Therefore, purifying selection is probably acting to maintain the correct structure of functionally characterized elements within these regulatory sequences. These results are consistent with the conclusions of Ludwig and Kreitman (1995) and Tautz and Nigro (1998), but they contrast with those of Jenkins, Ortori, and Brookfield (1995), who argue that positive selection is acting on the ftz zebra element. Despite these relatively constant functional constraints, it is possible that substantial binding site turnover may still be occurring in these sequences (Piano et al. 1999; Ludwig et al. 2000; Cuadrado, Sacristan, and Antequera 2001). In addition, it is possible that many functional sequences in these regions have not been identified, and thus there may be “near miss” binding site motifs close to attaining function which can facilitate such evolutionary change.
We generated models for Bcd and Hb because these transcription factors play an important role in the regulation of the genes examined here. Many experimentally verified binding sites of these two transcription factors have been previously identified in their regulatory regions. As a first step, we wanted to obtain an estimate of the power of our method to predict transcription factor binding sites. We approached this by performing a jack-knife, each time excluding one of the Bcd or Hb sites, generating the matrix and the thresholds from the remaining sites, and testing whether the excluded site was predicted as functional. This method predicted as functional 92.7% of the observed Bcd sites (47 of 51) and 94.8% of the observed Hb sites (92 of 97).

These power calculations are not typical because our test is not designed to reject the hypothesis that a sequence is not a site, but rather it is designed to reject the null hypothesis that a sequence is a binding site (see Methods for justification). In essence, the above calculations are for type-I error, the probability that we will reject the hypothesis that a site is functional when in fact it is. The reassuring result is that these calculations produce values very close to 95%, which is what we expect given the way the test is performed.

Another issue is the frequency of false positive results generated by our prediction method. This is the real value of the test, because it shows how many times we include a nonfunctional site in the functional range—i.e., how frequently we fail to reject that a site is functional when it is not functional. This is a difficult issue to address because we do not have a clear picture of the structure of a nonfunctional Bcd or Hb binding site. There have been no systematic experiments showing what are the types of sequences that Bcd or Hb do not bind and therefore we cannot generate models assuming nonfunctionality.

To obtain an idea of how high our false positive rate is, we compared our results to those obtained using a binding site prediction method that tests the null hypothesis that a sequence is not a binding site. We used the method developed by Yada et al. (1998) implemented in the Web-based software YEBIS (http://www.btls.jst.go.jp/yebis). We use the D. melanogaster sequences and the Bcd and Hb weight matrices obtained from our binding site data. Two LOD score thresholds were used for the binding site prediction with YEBIS. When the cut-off LOD score was set to 3, YEBIS predicted 26% fewer Bcd binding sites and 35% more Hb binding sites than our method, but all the YEBIS Bcd sites were also predicted with our method. For a cut-off LOD score of 2, the YEBIS results were almost identical to our method, with slight variations across genes for Bcd, but nearly twice as many more Hb sites were predicted relative to our method (see table 2 for summary of these results). This demonstrates one of the useful properties of our method. The weight matrix for Hb is more degenerate than that of Bcd, and we see that a conventional method predicts many more Hb sites than our method, even with a strict threshold, indicating the increased power of our method for de-generate weight matrices.

Because methods such as YEBIS, whose null hypothesis is a lack of function, produce very similar

---

**Table 2**

**Comparison of the Number of Binding Sites Predicted by Our Method Versus YEBIS**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Gene</th>
<th>Our Method</th>
<th>YEBIS, LOD = 3</th>
<th>YEBIS, LOD = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcd</td>
<td><em>eve</em> stripe-2</td>
<td>30</td>
<td>20 (20)</td>
<td>28 (28)</td>
</tr>
<tr>
<td></td>
<td><em>Hbp2</em></td>
<td>14</td>
<td>12 (12)</td>
<td>15 (14)</td>
</tr>
<tr>
<td></td>
<td><em>knirps</em> enhancer</td>
<td>16</td>
<td>12 (12)</td>
<td>16 (16)</td>
</tr>
<tr>
<td>Hb</td>
<td><em>en</em></td>
<td>5</td>
<td>7 (5)</td>
<td>9 (5)</td>
</tr>
<tr>
<td></td>
<td><em>eve</em> stripe-2</td>
<td>3</td>
<td>5 (3)</td>
<td>5 (3)</td>
</tr>
<tr>
<td></td>
<td><em>ftz</em></td>
<td>3</td>
<td>4 (3)</td>
<td>4 (3)</td>
</tr>
<tr>
<td></td>
<td><em>Hbp2</em></td>
<td>1</td>
<td>1 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td></td>
<td><em>knirps</em> enhancer</td>
<td>2</td>
<td>2 (2)</td>
<td>2 (2)</td>
</tr>
</tbody>
</table>

* In parentheses we show the number of binding sites that were predicted by both YEBIS and our method.

---

**A Novel Binding Site Prediction Method**

To capture the process of binding site evolution, we use a binding site prediction model as an indicator of potentially functional regulatory sequence components. Our binding site prediction approach relies on the initial assumption that a short sequence is a binding site. We generate a distribution of a test statistic ($L$), which quantifies the functional binding potential of a sequence by the transcription factor of interest. Our method, described in *Methods*, tests whether the observed $L$ value falls within the 95% confidence interval of known functional sites (fig. 2) of the same transcription factor. The rationale behind the null hypothesis of this approach follows:

1. For reasons mentioned in the *Introduction*, we wanted to avoid using the background sequence as a standard, so we had to rely just on the properties of known functional sites of the factors we are studying (Bcd and Hb).
2. Little is known about the properties of short sequences that are *not* functional binding sites of a given transcription factor. Therefore it is difficult to construct models that have as null hypothesis that a short sequence is not a binding site.
3. Some transcription factors are stricter in their sequence requirements for binding than others, and a null hypothesis that a short sequence is a binding site can more efficiently accommodate the special properties of each transcription factor. Even if we had good information on the features of sequences that are not binding sites of a given transcription factor, as sequence requirements for binding become more flexible the properties of short sequences that are not functional binding sites become harder to define.
4. We are interested in strong binding sites as well as weak ones. Our model is more likely to identify weak sites because they are generally less pronounced in terms of how well they fit a consensus, but are still within the range of the binding requirements of the transcription factor of interest. Note also that the “strength” of a binding site, as inferred from the weight matrix, is a rough approximation owing to the functional dependence on cooperative binding of other transcription factors.
results to our method, which has the opposite null hypothesis, it is safe to assume that the rate of false positive results produced by our method is low. In addition, our method relies highly on biological aspects because we take into account the distribution of $L$ values derived from functional binding and ignore the properties of the background sequence. It is also important to point out that for flexible weight matrices (high nucleotide degeneracy in the binding requirements of a transcription factor), the power to detect binding sites with a null hypothesis of no function will decrease, whereas the power of our test to predict the same sites will not decrease because the weight matrix degeneracy is taken into account in obtaining the cut-off values. It remains to be seen what proportion of our predicted binding sites will be experimentally validated with in vivo or in vitro binding assays.

Predicting the Pattern of Binding Site Evolution

Using the matrices and cutoff determined above, we searched the regulatory regions in the different species for putative Bcd and Hb binding sites. The objective of this analysis was to determine whether the nucleotide substitutions that have occurred within the experimentally verified Bcd or Hb binding sites have potentially eliminated or altered function. Additionally, we sought to evaluate whether nucleotide changes outside the known binding sites provide any evidence of gain and loss of Bcd and Hb binding sites. In figure 4 we show the pattern of predicted Bcd binding sites in regions that are regulated by Bcd, in all five Drosophila species studied. An important aspect of these data is that all but one of the known functional Bcd binding sites (Bcd_3 in eve stripe-2 enhancer in D. yakuba and D. erecta), and all but one of the Hb binding sites (Hb_1 in eve stripe-2 enhancer in D. erecta) are predicted as functional in all species. This suggests that within the range of divergence between D. melanogaster and D. yakuba/D. orena/D. erecta, purifying selection was strong enough to maintain most sites that are essential for gene regulation.

In contrast, many of the predicted (but not yet functionally supported) sites are not shared between species, which suggests that several potential binding sites with less essential roles can be gained and lost independently without strong effects on gene regulation. In figure 5 we present data describing the potential process of gain and loss of putative Bcd and Hb binding sites. It appears from this figure that the process of gain and loss of putative binding sites is present, with many events occurring in short evolutionary time. Sites that turn over would also be more tolerant to mutations, because the essential ones are already present and may even be responsible for variation in expression levels of the genes they regulate. It is interesting to note that although the lineages of D. yakuba and D. orena/D. erecta appear to have diverged at about the same rate from D. melanogaster, the change in the $L$ values caused by substitutions within binding sites in D. yakuba is higher than that in D. orena/D. erecta, which may indicate a unique mode of evolution of these regions in D. yakuba. In addition, the level of nucleotide divergence (table 3) is too low to explain the proportion of predicted sites that are not shared (fig. 5). The large number of gains and losses may be explained by the high proportion of indels, which can provide new sequences which may contain putative Bcd and Hb binding sites.

Another aspect of binding site divergence is whether the nucleotide substitutions between species are likely to affect quantitatively the ability of the transcription factor to bind the sequence. We chose experimentally verified Bcd
sites that have substitutions in *D. yakuba* and *D. orena/D. erecta* relative to the *D. melanogaster* sequence. For each site, we counted the number of nucleotide changes within the binding site sequence that differentiate the *D. melanogaster* sequence from the sequence of *D. yakuba* or *D. orena/D. erecta* and then generated all possible sequences that can be obtained by applying the same number of changes randomly to the *D. melanogaster* binding site sequence. We then scanned the mutated sites with the Bcd weight matrix and obtained for each original site the distribution of $L$ values of these randomly mutated binding site sequences. Finally, we compared the $L$ values from the random distribution to the $L$ values the sites obtain in their observed substituted form in the other species. Of five Bcd sites with substitutions, three sites showed changes significantly toward the conservative (look more like a Bcd site) extreme, one showed changes within the random range, and one showed changes significantly toward the radical (looks less like a Bcd site) extreme. This result suggests that although most of the nucleotide changes are maintaining the properties of the binding site, there are still changes that are not constrained and can potentially result in the loss of the site.

### Simulation of the Evolutionary Potential of Regulatory Regions

The basis of this study is that binding sites within regulatory regions undergo turnover, which results in the frequent gain and loss of functional elements but maintains the function of the regulatory region as a whole on the expression pattern of the gene (Hancock et al. 1999; Piano et al. 1999; Ludwig et al. 2000; Cuadrado, Sacristan, and Antequera 2001; McGregor et al. 2001). If this is true, then we expect to find traces of intermediate steps of this turnover process within these regions. These traces could either be lost binding sites affected by mutational events or sites that are close to becoming functional. The latter possibility is likely because the binding site sequences are short, and the probability of obtaining a new functional binding site by chance is relatively high, given sufficient time for the accumulation of mutations. It is also possible that some regulatory regions have properties such as nucleotide composition bias or repetitive sequences that allow the emergence of new binding sites for some transcription factors with high probability.

We simulated a range of divergence rates with or without mutation bias (see Methods; Stone and Wray 2001) in the regulatory regions of *knirps* enhancer and *hunchback* proximal promoter and asked whether there are regions where a new Bcd site, not predicted in the original *D. melanogaster* sequence, appears consistently at the same position in the simulated diverged sequences. We did 1,000 simulations of divergence of the *D. melanogaster* sequences terminating with an overall divergence of 10%, 20%, 30%, and 40% for each case. We then searched the simulated diverged sequences with the weight matrix for putative Bcd binding sites and obtained positions that allow the emergence of new binding sites for some transcription factors with high probability.

We simulated a range of divergence rates with or without mutation bias (see Methods; Stone and Wray 2001) in the regulatory regions of *knirps* enhancer and *hunchback* proximal promoter and asked whether there are regions where a new Bcd site, not predicted in the original *D. melanogaster* sequence, appears consistently at the same position in the simulated diverged sequences. We did 1,000 simulations of divergence of the *D. melanogaster* sequences terminating with an overall divergence of 10%, 20%, 30%, and 40% for each case. We then searched the simulated diverged sequences with the weight matrix for putative Bcd binding sites and obtained positions that allow the emergence of new binding sites for some transcription factors with high probability.

We simulated a range of divergence rates with or without mutation bias (see Methods; Stone and Wray 2001) in the regulatory regions of *knirps* enhancer and *hunchback* proximal promoter and asked whether there are regions where a new Bcd site, not predicted in the original *D. melanogaster* sequence, appears consistently at the same position in the simulated diverged sequences. We did 1,000 simulations of divergence of the *D. melanogaster* sequences terminating with an overall divergence of 10%, 20%, 30%, and 40% for each case. We then searched the simulated diverged sequences with the weight matrix for putative Bcd binding sites and obtained positions that allow the emergence of new binding sites for some transcription factors with high probability.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>mel</th>
<th>sim/sech</th>
<th>sech/yak</th>
<th>ore/ere</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim/sech</td>
<td>0.025 ± 0.0071</td>
<td>0.0063 ± 0.0102</td>
<td>0.064 ± 0.00846</td>
<td></td>
</tr>
<tr>
<td>yak</td>
<td>0.0064 ± 0.00846</td>
<td>0.053 ± 0.0099</td>
<td>0.056 ± 0.00839</td>
<td>0.051 ± 0.00782</td>
</tr>
</tbody>
</table>

Note.—*mel*, *D. melanogaster*; *sim*, *D. simulans*; *sech*, *D. sechellia*; *yak*, *D. yakuba*; *ore*, *D. orena*; *ere*, *D. erecta*. Estimated from the divergence values in the five different regulatory regions.
10% and 30% (fig. 6). This indicates that there are sequences within the regulatory regions that are close enough to becoming a functional site that a few substitutions can make then potentially functional. The original sequences (before divergence) can initially obtain substitutions and become weak sites, which may then be favorably selected if another functional Bcd site is lost or if the addition of the new binding site is advantageous. If favorable selection for new binding sites operates, the rate at which some of these are fixed will be higher. We also observed that, from the known functional sites, some are more robust to substitutions than others, and therefore the rate of loss of function is different for different sites. This is a result of the degeneracy of Bcd binding requirements and the sequence evolutionary starting point. Another interesting observation is that, at 10% divergence, the occurrence of a new site at the same position was lower than at 20% divergence, but then the occurrence reached a plateau at higher divergence values. This suggests that with 10% divergence there were not enough substitutions to generate the site, but above 20% divergence the number of substitutions was high enough to reach the maximum probability of gaining a site.

Discussion

Microevolutionary analysis of five Drosophila regulatory regions reveals that functional binding sites show patterns of variation consistent with functional constraint. However, methods that identify binding site locations suggest that there are many putative Bcd and Hb binding sites within regulatory regions, which may serve as the raw material for regulatory evolution. We demonstrate that a high proportion of these predicted sites are gained and lost in the evolutionary history of the Drosophila species, indicating extensive potential for functional binding site turnover. In addition, nucleotide changes occurring within the known and essential functional binding sites sometimes result in radical changes in the binding potential of the sites as this is described by the magnitude of the L value for Bcd and Hb sites or by the number of mutational events. Finally, by simulating divergence from the D. melanogaster sequence, we showed that regulatory sequences have a high potential for obtaining new functional binding sites (Stone and Wray 2001). Our results suggest an underlying dynamic process in the evolution of regulatory sequences, where there is an evolving background of potential binding sites ready to acquire an essential role when selected. In short, with this novel approach we have captured the intermediate steps of

Fig. 6.—(a) Counts of predicted Bcd sites at each position in 1,000 simulated divergent sequences for knirps and hunchback with divergence rate of 20%. (b) Counts of predicted Bcd sites at each position in 1,000 simulated divergent sequences for knirps under a mutation bias model (54:46 GC and AT bias shown). Black bars represent Bcd sites that occur in more than 100 of the 1,000 simulated sequences and are not predicted at the same position in the original sequence (new sites). Gray bars are experimentally verified Bcd sites in the original D. melanogaster sequence.
binding site turnover and have demonstrated that regulatory evolution, although difficult to study, may indeed be traceable if we use the appropriate methods.

Many studies have demonstrated high sequence conservation and conserved function of regulatory regions across distant Drosophila species (e.g., Langeland and Carroll 1993; Lukowitz et al. 1994). However, recent studies indicate that the system is more complicated and fluid, with regulatory regions having an underlying pattern of evolution not directly visible from simple sequence comparisons (Hancock et al. 1999; Piano et al. 1999; Ludwig et al. 2000; see Tautz 2000 for review). One of the reasons for such a contradiction is that the rules that govern the evolution of regulatory elements are not well known. It is therefore difficult to determine which changes are contributing to functional differences. One of the problems is that we do not know exactly how selection operates on regulatory elements. Furthermore, it is difficult to quantify the effect of a nucleotide change just with sequence comparisons, even within known functional sequences, because different changes will alter the properties of a binding site in different ways.

In this study we devised a binding site prediction method that uses PWMs generated from known binding site data and that uses the distribution of a similarity statistic (L) to obtain reliable thresholds that determine whether a tested sequence is likely to be functional. The use of such matrices allows for discrimination of nucleotide changes that affect the binding of the transcription factor from changes that are not likely to alter the binding affinity of the protein. This method is a special extension of the comparison of rates of synonymous versus nonsynonymous changes in coding sequences, but with two important distinctions: (1) it is not a binary value (synonymous–nonsynonymous) but rather it is a measure of the extent of deviation as described by the difference in L values in the same sequence in two different species; (2) it accounts for compensatory changes because the same likelihood value can be obtained even if several nucleotide changes have occurred.

This quantification of divergence of regulatory elements allows for analysis of the dynamics and potential of regulatory regions from the standpoint of the relevant structural unit, the binding site. We were able to identify putative binding sites for Bcd and Hb in these regions that are predicted with high enough L values that they are likely functional or potentially functional. What we do not know is how essential these putative binding sites are for the regulation of the gene. It is possible that these sites are used only in a subset of the tissues in which Bcd and Hb regulate these genes, and that is the reason why they are not identified in functional studies. They may also have little contribution to the regulation of the gene but could serve as a reserve in cases of evolutionary transitions. Finally, because Bcd and Hb are morphogens, these may be weak sites that contribute to a finer sensing of the nuclear regulatory environment by the gene so that downstream pathways are activated or suppressed in a more accurate way (Burz et al. 1998). This possibility is relevant because these segmentation genes are usually expressed in a striped fashion, and thus an accurate sensing of the regulatory environment is necessary to define the precise boundaries of the expression domains.

Theoretical studies have predicted that the early development system of genes in Drosophila is highly fluid and allows for many evolutionary changes without serious distortion of the embryonic patterning (Gibson 1996; von Dassow et al. 2000). Transgenic experiments have shown that excessive expression of bicoid initially distorts the anterior-posterior pattern but is later corrected by fusion and cell death (Namba et al. 1997). Our analysis shows that there is possibly underlying fluidity in the system, not readily observed, as indicated by the properties of nucleotide sequence within these regulatory regions. This is also supported by a recent study that looked at the co-evolution of components of the same system (Shaw et al. 2002). We have shown with simulations that there are short sequences within these regions that have a high potential of becoming Bcd binding sites by many different combinations of substitutions. This potential could derive from “dead” sites that were lost not long ago, or because the probability of emergence of a binding site is increased by a skewed nucleotide composition or features such as mononucleotide runs of As or Ts that may provide good background sequence for Bcd and Hb sites to eventually emerge. It is finally important to mention that gene regulation is not carried out solely through the interaction of transcription factors with their binding sites. Several other aspects such as chromatin structure, flanking sequences, and other properties of the sequence environment are critical for the efficient interaction of proteins with regulatory DNA.

We have shown that regulatory regions have the potential to undergo turnover. However, the main development in this study is that we have devised a way to follow this evolutionary process and explain how these transitions can occur. Based on our results, it is not surprising that recent studies have demonstrated extensive gain and loss of binding sites (Cuadrado, Sacristán, and Antequera 2001; Dermitzakis and Clark 2002). We have shown that there is a finer organization in these regulatory regions that, together with the redundancy of binding site requirements, allow the reassignment of roles to some regulatory elements and the occasional emergence of new regulatory properties. This process is achieved at a slow rate and possibly involves radical transitions rather than a continuous evolutionary process.

Literature Cited


Thomas Eickbush, Associate Editor
Accepted December 12, 2002