Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information

Vsevolod J. Makeev, Alexander P. Lifanov¹, Anna G. Nazina² and Dmitri A. Papatsenko²,*

Scientific Center `Genetika', Moscow 113545, Russia, ¹Engelhardt Institute of Molecular Biology, Moscow 119991, Russia and ²Department of Biology, New York University, New York, NY 10003-6688, USA

Received June 6, 2003; Revised July 31, 2003; Accepted September 2, 2003

ABSTRACT

We explored distance preferences in the arrangement of binding motifs for five transcription factors (Bicoid, Krüppel, Hunchback, Knirps and Caudal) in a large set of Drosophila cis-regulatory modules (CRMs). Analysis of non-overlapping binding motifs revealed the presence of periodic signals specific to particular combinations of binding motifs. The most striking periodic signals (10 bp for Bicoid and 11 bp for Hunchback) suggest preferential positioning of some binding site combinations on the same side of the DNA helix. We also analyzed distance preferences in arrangements of highly correlated overlapping binding motifs, such as Bicoid and Krüppel. Based on the distance analysis, we extracted preferential binding site arrangements and proposed models for potential composite elements (CEs) and antagonistic motif pairs involved in the function of developmental CRMs. Our results suggest that there are distinct hierarchical levels in the organization of transcription regulatory information. We discuss the role of the hierarchy in understanding transcriptional regulation and in detection of transcription regulatory regions in genomes.

INTRODUCTION

Initiation of tissue-specific or spatio-specific transcription in multicellular organisms requires binding of multiple transcription factor molecules to transcription regulatory regions, such as promoters and enhancers (cis-regulatory modules; CRMs). Multiple binding motifs and even multiple binding sites for the same motif presented in the regulatory regions are often described as ‘regulatory clusters’ (1-6). Statistical models, based on motif clustering, are helpful for finding novel CRMs in the genome, but very often they consider only site density (cluster significance) and relative site affinity (such as a weighted matrix score) (6). However, it is known that specific arrangements of binding sites within the regulatory regions (regulatory clusters) are necessary to achieve proper biological function. Incorporating such architectural features into formal clustering models might facilitate computational recognition of CRMs and interpretation of their biological function (7).

Specific arrangements between binding sites are known from many examples in biology. For instance, recent quantitative studies of basal transcription (8) revealed a striking dependence of basal promoter activity on both the distance and the orientation of an artificial activator binding site (Gal4) and the TATA box. An optimal spacing between binding sites (NF-Y and SRE motifs) has also been demonstrated in the human SREBP-2 promoter (sterol regulatory element-binding protein) (9). In vitro analysis of binding site arrangements in the rat collagenase-3 promoter (10) has revealed that a 10 bp (‘helical’) phasing in binding site distribution provides maximal transcriptional activity. The importance of the ‘helical phasing’ and specific binding site arrangement was also demonstrated in vivo for the murine CD4 promoter (11). In some cases, even a very small difference in the distance between binding motifs results in dramatically different transcriptional outcomes. One of the most striking examples of this kind is the binding of the POU domain transcription factor Pit-1 to its target sites, differentially spaced (2 bp difference) in growth hormone and prolactin gene promoters (12). The ‘helical phasing’ (10 bp) has also been demonstrated computationally (13) using a large number of proximal eukaryotic promoters (14) and the list of binding motifs available from the TRANSFAC database (15). In many of the described quantitative experimental studies, the disruption of specific spacing (phasing) between binding sites resulted in reduction, but not abolishment, of transcription. This fact, together with some known cases of successful promoter reconstruction (16-18), also supports the presence of a certain flexibility in site arrangement.

The biological reasons leading to a specific arrangement of sites in promoters are clear: the transcription factors, bound to promoter DNA, are also involved in specific protein–protein interactions (19,20); therefore, the binding motifs must be distributed in the promoter in a non-random fashion. In other words, the arrangement of binding motifs can control the formation of 3D protein complexes involved in initiation of specific transcription.

Attempts to reveal and describe specific site arrangements resulted in a very interesting concept of composite elements (CEs) (21). In the simplest case, a CE corresponds to a pair of
individual binding motifs located at a particular distance and involved in formation of specific tertiary (DNA–protein–DNA) complexes. Identical CEs may perform related functions in different genes. Further development of this concept resulted in construction of a dedicated database TRANSCompel (22), combining sequences for 256 (Release 6.0) CEs from different organisms. Currently, the CE concept is widely used for finding co-localized, synergistic (antagonistic) binding motif pairs (23,24) or combinatorial arrays of motifs responsible for the formation of similar gene expression profiles (25–27).

In the current work, we explored preferential site distances in CRMs of Drosophila developmental genes. CRMs are transcription regulatory units (~1 kb range), often located far from the transcription start site and responsible for spatiotemporal expression of their cognate developmental genes (3). We recently built a database containing known functional Drosophila CRMs (see our web resource: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm) together with a list of matrices for a number of transcription factors and known transcriptional interactions (6,28). Selection of relevant binding motifs in a particular functionally related group of transcription regulatory regions minimizes the risk of false positives, which is known to be a problem for large-scale analysis of highly diverse data sets.

In this work, we have shown that the binding sites in CRMs of Drosophila are arranged in particular ways, indicating the presence of specific developmental CEs. We also discuss a general model describing hierarchical levels in organization of transcriptional information and the role of CEs in understanding the responses of developmental genes to transcriptional signals.

**METHODS**

In order to identify binding site cores, we calculated information content I (bits) in the ith column of a binding motif alignment as the Shannon entropy for this alignment column (29):

\[ I_i = 2 + \sum_{\alpha} q_i^\alpha \log_2 (q_i^\alpha) \]

In this equation, \( q_i^\alpha \) represents the frequency of the letter \( \alpha \) (\( \alpha \in \{A,C,G,T\} \)) in the ith position of the alignment. To calculate the score of a binding motif match, we constructed position weighted matrices (PWMs) for each motif using the equation with a pseudocount parameter:

\[ S_i^\alpha \log \left( \frac{n_i^\alpha + aq_\alpha}{(n+a)q_\alpha} \right) \]

\( S_i^\alpha \) is the score of letter \( \alpha \) in position \( i \), \( n_i^\alpha \) is the number of letters \( \alpha \) in column \( i \) of the motif alignment, \( q_\alpha \) is the frequency of letter \( \alpha \) in the Drosophila genome, and \( a \) is the pseudocount parameter, which we set as equal to 1. In our previous publications, we discussed how to select the PWM cutoff for the binding motifs in this particular system (28). For each PWM cutoff value, we estimated the site frequency \( E_S \) as the total number of motif matches above the PWM cutoff in the entire Drosophila genome normalized to the length of the genome.

To estimate randomness of binding site distribution in the CRMs, we have found all distances between neighboring sites and compared the observed distance distribution with the expected distance distribution in a random Bernoulli sequence (see also equation 7). For each \( j \)th interval of 10 distances \((j = 5, 15...225)\), we calculated Z-scores (see Fig. 1):

\[ Z_j = \frac{(N_j^{\text{obs}} - N_j^{\text{exp}})}{\sqrt{N_j^{\text{exp}}}} \]

In this equation, the expected number of distances \( N_j^{\text{exp}} \) for each distance interval \( j \) in genomic samples was calculated taking into account the site frequency in the genome \((E_i = 5E-4)\). The corresponding PWM cutoff values for Bicoid (Bcd) and Krüppel (Kr) are shown in Table 1. In the case of the CRM data set, we calculated \( E_S \) from the number of sites having the same frequency in the genome (the same PWM cutoff), but actually found in the CRM data set \((E_S^{\text{CRM}} = \text{the total number of sites in CRMs/total CRM length})\). Due to the limited size of CRMs, we calculated the total number of distances between the neighboring sites \( N_j^{\text{CRM}} \) only in the range of 1–300 bp. With correction for the maximal possible number of distances (conservative estimation), we calculated the expected number of sites in the \( j \)th distance interval as:

\[ N_j^{\text{exp}} = \frac{N_j^{\text{CRM}}}{300} \sum_{n=1}^{300} E_n(1 - E_n)^{n-1} \]

To obtain comparable statistical values, the Z-scores for genomic sequences were calculated for the sample size of CRMs \((N_j^{\text{obs}} = N_j^{\text{CRM}})\) using the same equation. Statistical noise caused by the small sample size prevented further reduction of the selected distance intervals (see Fig. 1A). Larger intervals would result in lower resolution by distance.

Spectra of distance distributions in the frequency domain were built using the Matlab (Mathworks, Inc.) signal processing module. We used a filtered fast Fourier transform (FFT) algorithm, implemented in the multiple signal classification method (MUSIC). The order of FFT was set to the maximal, as well as the signal dimension (149 for 300 input points, analyzed range of distances), thus keeping all putative signals without noise reduction. Input signals (Z-scores) for periodic analysis were generated from distance distributions (histograms smoothed by three distance points) using equation 3, taking into account the site frequency observed in the CRM data set (total number of sites in CRMs/total CRM length, see above).

To extract Bcd/Kr functional elements, we calculated PWM scores for Bcd and Kr, respectively, for each position (with the +2 bp shift) of CRM sequences and large genomic samples (1 Mb total). Then, for each jth PWM score zone (i.e. Bcd 5.4–5.8, Kr 7.4–7.8, see Table 2), we found the number of matches in the CRM data set \( N_j^{\text{obs}} \) and compared this number with the number obtained from genome samples \( N_j^{\text{exp}} \). Notice that in the described test, \( N_j^{\text{obs}} \) and \( N_j^{\text{exp}} \) were different (j here is a PWM score zone) from those given above.
To minimize errors caused by the small size of the sample (the number of overlaps in CRM sequences), we calculated a conservative estimator for the number of overlaps, $N_{\exp}$. To find this number, we evaluated separately the conditional probability of observing a Kr match given a Bcd match $P_{\exp}^{E} (KR_{i+2} | BCD_{i})$ and the conditional probability of observing a

![Figure 1](https://academic.oup.com/nar/article-abstract/31/20/6016/1039515)

**Figure 1.** Distribution of distances built for the neighboring binding sites. (A) The histogram of distances between the neighboring sites for five binding motifs in developmental CRMs. The red line shows the random (geometric) distribution calculated for sites with the same frequency in CRMs. (B) Normalized deviations (Z-scores; see equation 3) between the observed and the expected number of distances. The Z-scores were calculated for distance intervals of 10 bp. The binding sites in CRMs (green bars) are closer than expected from the random distribution and the distribution detected in the genome (blue bars).

<table>
<thead>
<tr>
<th>PWM score</th>
<th>5.0</th>
<th>5.4</th>
<th>5.8</th>
<th>6.2</th>
<th>6.6</th>
<th>7.0</th>
<th>7.4</th>
<th>7.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site $E_{\text{genome}}$</td>
<td>0.002</td>
<td>0.001</td>
<td>8E-04</td>
<td>5.2E-04</td>
<td>3.3E-04</td>
<td>2E-04</td>
<td>1.2E-04</td>
<td>1E-04</td>
</tr>
<tr>
<td>CRMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Bcd sites</td>
<td>324</td>
<td>240</td>
<td>172</td>
<td>118</td>
<td>88</td>
<td>67</td>
<td>38</td>
<td>16</td>
</tr>
<tr>
<td>No. of Kr sites</td>
<td>264</td>
<td>207</td>
<td>160</td>
<td>109</td>
<td>87</td>
<td>70</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>No. of overlaps</td>
<td>92</td>
<td>72</td>
<td>45</td>
<td>30</td>
<td>22</td>
<td>12</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>$P_{\text{CRM}}^{E} (KR_{i+2}</td>
<td>BCD_{i})$</td>
<td>0.28</td>
<td>0.30</td>
<td>0.26</td>
<td>0.25</td>
<td><strong>0.25</strong></td>
<td><strong>0.18</strong></td>
<td><strong>0.13</strong></td>
</tr>
<tr>
<td>$P_{\text{CRM}}^{E} (BCD_{i}</td>
<td>KR_{i})$</td>
<td>0.35</td>
<td>0.35</td>
<td>0.28</td>
<td>0.25</td>
<td>0.25</td>
<td><strong>0.17</strong></td>
<td><strong>0.14</strong></td>
</tr>
<tr>
<td>Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{\text{E}}^{E} (KR_{i+2}</td>
<td>BCD_{i})$</td>
<td>0.32</td>
<td>0.28</td>
<td>0.20</td>
<td>0.18</td>
<td>0.14</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>$P_{\text{E}}^{E} (BCD_{i}</td>
<td>KR_{i})$</td>
<td>0.31</td>
<td>0.29</td>
<td>0.22</td>
<td>0.21</td>
<td>0.20</td>
<td>0.09</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The table shows changes in the motif interdependence with increasing site probability (PWM cutoff). The number of sites detected in the CRM data set is given for Bcd and Kr motifs for a range of considered site frequency [site $E_{\text{genome}}$]. Motif interdependence in the CRM data set was estimated from these numbers with respect to Bcd: $P_{\text{CRM}}^{E} (BCD_{i} | KR_{i})$ and Kr: $P_{\text{CRM}}^{E} (KR_{i+2} | BCD_{i})$. The motif interdependence in CRMs is compared with the motif interdependence in the genome (see the bottom two rows). The interdependence between Bcd and Kr is much higher in CRMs for high-affinity sites.
Comparison between observed and expected numbers of Bcd/Kr overlaps found for the CRM data set. The conservative estimation for the expected number of overlaps was calculated from the number of overlap occurrences in the genome sequences (see Methods). The first row and the first column contain the PWM score cutoff for the Bcd and Kr motifs, respectively. Only in certain score zones is the number of overlapping sites found in CRMs higher than expected (numbers are in bold). Presumably, these score zones contain functional Bcd/Kr overlaps.

Bcd match given a Kr match $P_j^G (BCD_{i+2} | KR_i)$). Both conditional probabilities were estimated from the entire genome. To obtain the conservative estimation, we took the maximum over the two expectations for each $j$th PWM score zone:

$$N_j^{BcdCRM} = \max [N_J^{BcdCRM}, P_j^G (KR_{i+2} | BCD_i)]; \quad N_j^{KrCRM} = P_j^G (BCD_{i+2} | KR_i)]$$

Here $N_j^{BcdCRM}, N_j^{KrCRM}$ are the numbers of Bcd and Kr matches, respectively, found in CRMs for the $j$th PWM score zone. Notice that the conditional probabilities can also be calculated for any PWM score zone, e.g. for sites having site probability above a selected cutoff $E$ (shown in Table 1):

$$P_j^G (KR_{i+2} | BCD_i) = \sum_{E_{Bcd} = E_{Kr} > E} P_j^G (KR_{i+2} | BCD_i)$$

RESULTS

Mapping binding sites and defining distances

To perform an analysis of distances between binding sites, it is necessary (i) to select and map binding motifs; (ii) to delineate a relevant sequence data set; and (iii) to formulate a working definition of distances between the binding site matches.

We limited our choice to the five best binding motifs for the transcriptional regulators Bcd, Caudal (Cad), Hunchback (Hb), Kr and Knirps (Kni), having a relatively large number of occurrences in our CRM database. To map these binding motifs, we employed a PWM search with parameters described earlier (6,28). In general, we considered only high-affinity sites with probabilities not exceeding $10^{-3}$, as estimated from the Drosophila genome (see Methods).

To generate a representative sequence data set, we considered only CRMs regulated by any of the selected transcription factors and containing multiple high-affinity binding sites for these proteins. The positions of binding site clusters previously identified in the context of these CRMs (6) provided a formal criterion for establishing boundaries of the selected early developmental CRMs. The total size of analyzed CRMs after the described pre-screening procedure combined >68 kb of sequence data in 33 non-overlapping contigs. The sequence data can be obtained from our web resource (see New York University website: http://homepages.nyu.edu/~dap5/PCL/pseudoobscura/train_plus_contigs.zip).

Since the binding motifs for selected transcription factors have different widths, we measured the distances between the centers of binding site cores made by site alignment columns with a high informational content (see Methods). Table 3 illustrates the procedure of distance measurement. Notice that distances even for the same binding motif may require coordinate adjustment due to the asymmetry of the motif.

With the described rules, the distances can be measured between sites that belong to the same binding motif or between sites belonging to different binding motifs; between sites located on the same DNA strand (in tandem) or sites on the opposite strands (in palindrome).

Non-random arrangement of binding motifs in CRMs

To test whether the binding site arrangement in the Drosophila CRMs is non-random, we calculated all distances between neighboring binding sites for the five motifs, and compared the obtained distance distribution with the random expectation. In a random sequence, the probability of observing distance $n$ between two neighboring PWM matches can be calculated from the geometric distribution (30):

Table 3. Definition of the distance between two binding sites

<table>
<thead>
<tr>
<th>Motif/orientation</th>
<th>Shift</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicoid &gt;</td>
<td>1</td>
<td>0.08</td>
<td>0.08</td>
<td>0.98</td>
<td>1.36</td>
<td>1.48</td>
<td>0.96</td>
<td>1.48</td>
<td>0.78</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicoid &lt;</td>
<td>0</td>
<td>0.09</td>
<td>0.78</td>
<td>1.48</td>
<td>0.96</td>
<td>1.48</td>
<td>1.36</td>
<td>0.98</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caudal &gt;</td>
<td>2</td>
<td>0.17</td>
<td>0.20</td>
<td>0.96</td>
<td>0.96</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.96</td>
<td>0.33</td>
<td>0.20</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>Caudal &lt;</td>
<td>2</td>
<td>0.11</td>
<td>0.07</td>
<td>0.20</td>
<td>0.33</td>
<td>0.96</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.20</td>
</tr>
<tr>
<td>Hunchback &gt;</td>
<td>1</td>
<td>0.15</td>
<td>1.10</td>
<td>1.59</td>
<td>1.66</td>
<td>1.34</td>
<td>1.32</td>
<td>0.57</td>
<td>0.54</td>
<td>1.33</td>
<td>0.33</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Hunchback &lt;</td>
<td>1</td>
<td>0.07</td>
<td>1.33</td>
<td>0.54</td>
<td>0.57</td>
<td>1.32</td>
<td>1.34</td>
<td>1.66</td>
<td>1.59</td>
<td>1.10</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knirps &gt;</td>
<td>1</td>
<td>0.62</td>
<td>0.87</td>
<td>0.37</td>
<td>0.54</td>
<td>0.39</td>
<td>0.60</td>
<td>1.21</td>
<td>1.36</td>
<td>1.36</td>
<td>0.74</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Knirps &lt;</td>
<td>2</td>
<td>0.21</td>
<td>0.74</td>
<td>1.36</td>
<td>1.36</td>
<td>1.21</td>
<td>0.60</td>
<td>0.39</td>
<td>0.54</td>
<td>0.37</td>
<td>0.87</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Krüppel &gt;</td>
<td>1</td>
<td>0.24</td>
<td>0.76</td>
<td>0.85</td>
<td>0.63</td>
<td>0.97</td>
<td>0.97</td>
<td>0.25</td>
<td>0.56</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krüppel &lt;</td>
<td>0</td>
<td>0.48</td>
<td>0.56</td>
<td>0.25</td>
<td>0.97</td>
<td>0.97</td>
<td>0.63</td>
<td>0.85</td>
<td>0.76</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
where \( E \) is the site occurrence in the selected data set and \( n \) (bases) is the distance between the neighboring motif matches. In this test, PWM cutoff values (site likelihood ratio values) were set to achieve the same site occurrence for each binding motif in the Drosophila genome, equal to \( E = 5 \times 10^{-4} \) (6). We excluded close distances from this consideration (see ‘Overlapping and correlated motifs’ below).

This statistical test (see Fig. 1 and Methods) demonstrated that the distances between sites in CRMs are smaller than expected from the described random model. We also analyzed the distribution of distances between the neighboring binding sites in Drosophila genome samples (1% of genome total). In fact, a similar distance distribution was observed (see Fig. 1B), although the significance of the short distances in genome samples was much smaller. The existence of microsatellites, repetitive sequences and other correlations in DNA (31) can explain the deviation of the observed from the expected distance distribution in genome samples.

The described analysis demonstrated that binding sites in Drosophila CRMs are distributed in a non-random fashion and the fraction of sites having spacing in the range 50–60 bp is larger than expected. This distance range might correspond to CEs containing several closely spaced binding sites. Notice also that even 'stand alone' sites may represent parts of CEs, which were not detected in our search using five binding motifs.

### Periodic signals in arrangement of a single binding motif

To explore whether binding sites are distributed periodically in CRMs, we calculated distances between any two binding sites (all possible binding site pairs in a CRM) belonging to the same binding motif or to a binding motif combination. In this case, the distance expectation in a random sequence is independent of the distance itself and has a uniform distribution (30). Therefore, we compared the empirical (observed) distribution of distances for any binding site pair with the uniform distribution.

To minimize interference between periodic signals, specific to different binding motifs (or binding motif combinations), we focused our attention on analysis of distances between sites belonging to one or two binding motifs. Periodic signals present in the resulting distance distributions were assessed using Fourier analysis (see Methods).

First, we built distance distributions and the corresponding Fourier spectra for the three most frequent binding motifs from our data set, Bcd, Hb and Kr. The most striking result, confirming the hypothesis of ‘helical phasing’ (see Introduction), was obtained for Bcd (see Figs 2A and B, 3A and B, and 5A). The vast majority of high-affinity Bcd sites are positioned at distances close to 10, 20, 30 etc. bp. The periodicity in the arrangement of Bcd sites drops rapidly with decreasing site affinity, supporting the biological importance of this specific signal. A similar, but not identical periodic signal was observed in the distribution of distances between binding sites for Hb (Figs 2C and 3D). In this case, however, the period was equal not to 10 but to 11 bp. This difference in periodicity might be explained by a slightly different DNA conformation (twist) of the two binding motifs (compare CCTAATCCC, the consensus for Bcd, and TTTTTTTG, the consensus for Hb). Surprisingly, the distribution of another binding motif, Kr, showed no periodic signal corresponding to the ‘helical phasing’ (see Fig. 3E). The different structure of the Kr DNA-binding domain together with the different mechanism of Kr binding might explain the absence of the ‘helical phasing’ in the distribution of Kr sites. Bcd is known to be involved in cooperative DNA binding (32), which typically requires several closely spaced binding sites. Instead, Kr seems to be involved in competitive rather than cooperative DNA binding (see ‘Overlapping and correlated motifs’ below) (33).

The arrangement of binding motifs for another transcriptional activator, Cad, also displayed ‘helical phasing’ (data not shown). The Kni motif has a low number of occurrences in our data set and was not considered in this type of analysis.

### Periodic signals in arrangement of a binding motif combination

To extract periodic signals corresponding to a specific combination of binding motifs (potential synergistic pairs or CEs), we analyzed distance preferences for pairs (any two matches) of the most frequent motifs from our database, Bcd–Hb, Bcd–Kr and Kr–Hb, and the corresponding Fourier spectra. Expression patterns of Bcd, Hb and Kr in the early embryo have substantial overlaps and the transcription factors are expected to be involved in direct synergistic or antagonistic interactions (33)

Analysis of the Bcd–Hb pair revealed two phasing signals (10 and 11 bp), corresponding to Bcd–Bcd and Hb–Hb combinations (Fig. 3G). The high amplitude of the signal corresponding to the double ‘helical’ period (22 bp) is the result of signal interference from Bcd–Bcd and Hb–Hb pairs (compare positions of peaks corresponding to 2P period in Fig. 3A and D). We also generated the differential Bcd–Hb spectrum (data not shown) by removing distances for Bcd–Bcd and Hb–Hb pairs from our consideration, but detected no high-amplitude periodic signals. Given the periodicity detected in distributions of Bcd and Hb motifs separately, even the presence of a single specific distance for the Bcd–Hb pair would result in a periodic signal. The absence of specific distances between Bcd and Hb suggests that they perform their functions rather independently and perhaps their binding motifs never belong to the same CE, or the potential Bcd–Hb CEs have a flexible structure and are difficult to detect using our type of analysis.

We also explored distance preferences in the distribution of another interesting motif pair, Bcd–Kr. These motifs are very similar (CCTAATCCC, the Bcd consensus, and TAAACCCTTT, the Kr consensus) and the corresponding transcription factors are involved in antagonistic interactions by competing for the same binding sequences in regulatory regions (34). We analyzed both the short- (<5 bp, see ‘Overlapping and correlated motifs’, below) and the long-range Bcd–Kr distances. Periodic analysis of the Bcd–Kr distribution (long-range distances) revealed the presence of a new signal, having a period rather opposite to the ‘helical’ (17 bp, see Fig. 3H). The signal was absent in Fourier spectra, built for either Bcd or Kr motif distribution (Fig. 3A and E). The differential Fourier spectrum, generated for Bcd–Kr distances only (data not shown), confirmed the presence of the 17 bp periodic signal and of an additional signal, with a period close to that of the...
More specifically, ‘helical’ (11 bp). This second signal is expected, as Bcd and Kr motifs are highly correlated (2 bp shift) and the Bcd sites are distributed periodically (10 bp). However, the signal corresponding to 17 bp is new and its presence suggests that the non-overlapping Bcd and Kr sites have a tendency to be placed on the opposite sides of the DNA helix (bound proteins are facing in opposite directions). In this case, the non-overlapping Bcd and Kr sites may belong to distinct CEs, performing independent functions (see Discussion).

We also extracted periodic signals specific to some other motif combinations (Hb–Kr, no new signals) as well as signals presented in the combination of all five binding motifs (Bcd, Hb, Kr, Kni and Cad). In the latter case, we detected the same ‘helical’ signal (11 bp), although with somewhat lower amplitude. In addition, we measured periodicity in the distribution of experimental sites (not PWM matches) in one of the best studied enhancer regions, even-skipped stripe 2, from six Drosophila species (35). In this case, we also detected the major signal with a period close to 10 bp and the ‘opposite phase’ signal (17 bp), presumably corresponding to the Bcd–Kr motif combination (see Fig. 3C). Table 4 summarizes data for detected periodicities in distributions of the considered binding motif combinations.

These data clearly demonstrate that the arrangement of non-overlapping binding motifs in regulatory regions cannot be described by the simple ‘helical phasing’ formula. Instead, each binding motif as well as each binding motif combination exhibits its own periodicity, sometimes quite different from the major ‘helical’ signal (10–11 bp).

**Overlapping and correlated motifs**

In the analysis described above, we considered only non-overlapping sites separated by distances exceeding the binding motif lengths. Nevertheless, the overlapping sites are of interest, especially when the binding motifs correlate and the transcription factors compete for the same binding sequences. As described above, the Bcd–Kr (activator–repressor) motif combination is a characteristic example of this quite common biological situation. Bcd (consensus CCTAATCCC) and Kr (consensus TAACCCTTT) motifs may overlap by chance (consensus CCTAAYCCYTTT), but some of the overlaps do correspond to functional antagonistic elements (composite sites) and some do not. We calculated the possible fraction of the functional Bcd/Kr overlaps and extracted these putative antagonistic elements from our database of developmental CRMs.
In the first step, we estimated what distance (shift) between Bcd and Kr matches causes maximal motif correlation. For each possible shift, we calculated PWM scores for the Bcd and Kr motifs in every position of a test DNA sequence. Figure 4 shows that the maximal correlation value (CC = 0.7) between the two motifs is observed if they are placed on the same DNA strand (according to the orientations shown) and shifted by two bases. Correlation values obtained for other Bcd/Kr motif shifts were low (no overlapping high-scoring matches). In the second step, we compared the frequency of words containing overlapping Bcd and Kr sites (with the 2 bp shift) in the Drosophila genome with the frequency of the same words in the developmental CRMs (see Methods). Table 2 shows a comparison of the numbers of Bcd/Kr overlaps found in the CRM data set and the corresponding numbers found in the genome and normalized to the sample size (number of sites) of
the CRM data set. Words corresponding to Bcd/Kr overlaps and overrepresented in CRM sequences (see numbers in bold in Table 2) were extracted and aligned as shown in Figure 5B. The alignment contains many of the known functional Bcd±Kr elements, for instance those found previously in the even-skipped stripe 2 region.

The described analysis demonstrates how the test for motif interdependence might help in extraction of antagonistic elements, such as the Bcd/Kr composite binding site. Given a set of transcription regulatory sequences and a list of binding motifs, it seems to be possible to reveal the presence of potential antagonistic relationships among the motifs (competitive binding) without any additional information.

**DISCUSSION**

**Preferred site arrangement in developmental enhancers**

The preferential arrangement of binding sites for transcription factors in regulatory modules might be considered as a specific type of functional information encoded in regulatory DNA. In the current work, we demonstrated how to extract preferential site arrangements and potential CEs from large data sets using periodic analysis and a test for motif interdependence.

We have shown that the distribution of Bcd and Hb motifs, considered alone, fits well to the known ‘helical phasing’ rule (10–11 bp), and so the bound transcription factors are placed on the same surface of DNA. According to existing models (10), this preferential binding site arrangement facilitates protein–protein interactions and promotes formation of specific tertiary complexes (DNA–protein–protein–DNA), involved in activation of specific transcription. The phenomenon of ‘helical phasing’ is also known from the distribution of nucleosome positional signals (13), and thus it is very likely that multiple DNA–protein contacts, caused by any large protein complex, have a good chance of following the ‘helical phasing’ rule. The example in Figure 5A represents actual sequences, containing periodically distributed binding sites for Bcd. Alignment of these sequences reveals a common element, containing at least 2–3 high-scoring Bcd matches, which are present in many developmental CRMs.

It is more important, however, that the ‘helical phasing’ rule is not sufficient to describe arrangement of any binding motif or binding motif combination in any regulatory sequence. This fact is demonstrated by periodic signals detected in the distribution of the Kr motif and of the Bcd–Kr binding motif combination (17 bp, see Fig. 3E and H).

**Hierarchical levels in the organization of transcription regulatory information**

CRMs represent independent functional units, responsible for the formation of specific expression patterns in developing fly embryo. This functional independence of CRMs suggests that they constitute one of the upper levels in the informational hierarchy. Conversely, binding motifs for transcription factors represent the bottom level, as they cannot be divided further, e.g. into smaller functional ‘subwords’. Combinatorial arrangements of binding motifs such as CEs and antagonistic

![Figure 4](https://academic.oup.com/nar/article-abstract/31/20/6016/1039515)

**Table 4. Periodic signals detected in binding motif distributions**

<table>
<thead>
<tr>
<th>Motif</th>
<th>Bcd</th>
<th>Hb</th>
<th>Kr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcd</td>
<td>10, 21</td>
<td>11, 22 (18, 19, 28)</td>
<td>17 (11, 26)</td>
</tr>
<tr>
<td>Hb</td>
<td>22 (10, 11)</td>
<td>(11, 18, 19)</td>
<td>23, 26 (19)</td>
</tr>
</tbody>
</table>

The signal corresponding to the ‘helical phasing’ (10–11 bp) is present in the distance distribution for Bcd–Bcd and Hb–Hb motif combinations, but absent in the distribution of distances for Kr–Kr. The Bcd–Kr combination displays the signal opposite to the ‘helical phasing’ (17 bp). Low-amplitude signals are shown in parentheses.

![Figure 5A to D](https://academic.oup.com/nar/article-abstract/31/20/6016/1039515)

**Figure 4. Correlation between occurrences of Bcd and Kr motifs.** (A–D) Scatter plots show the PWM scores calculated for each position of a test sequence to Bcd (x-axis) and Kr (y-axis). Different shifts between positions of the Bcd and Kr matches (shown on the top of each panel) result in different correlation between the motifs. (A) No shift; (B) +1 bp shift; (C) +2 bp shift; (D) +3 bp shift. Dots in the shaded area correspond to overlapping high-affinity Bcd/Kr sites. (E) Values of the correlation coefficient calculated with different motif shifts. The maximal correlation is observed for a +2 bp shift (CC = 0.7).
pairs of overlapping motifs (e.g. Bcd/Kr) might represent yet another, ‘middle’ level in the informational hierarchy.

In this work, we demonstrated that binding motifs are distributed in Drosophila CRMs in a non-random fashion, with a large fraction of sites belonging to small closely spaced groups (50–70 bp range). Moreover, within these small groups (putative CEs), the distances between the binding sites are far from random and comply with some specific spacing requirements. We believe that these two findings confirm the presence of the ‘middle’ hierarchical level, defining (i) order, (ii) orientation and (iii) spacing in small groups of binding sites. If so, then the ‘upper’ informational level, represented by entire promoters and CRMs, combines several such small functional groups or several CEs, acting independently in their response to the spectrum of native transcriptional signals. For example, repression of the same promoter (CRM) by two transcription factors might be achieved through an independent response of two or more corresponding CEs to the concentrations of these proteins.

One can see that maximal spatial independence of adjacent CEs might be achieved through positioning of corresponding protein complexes on opposite sides of the DNA helix. In this respect, our finding of the 17 bp phasing (opposite to the ‘helical’) in the distribution of the Bcd–Kr motif combination (see Fig. 3C and H) fits the proposed model (see Fig. 6). Three hierarchical levels, binding motifs, CEs and CRMs (as well as

**Figure 5.** Structure of Bcd–Bcd and Bcd–Kr elements. (A) The alignment of sequences containing specifically spaced Bcd-binding sites (10 bp peak in Fig. 2A). (B) Antagonistic Bcd/Kr element, built from sequences over-represented in CRMs. The list contains some functional Bcd/Kr overlaps found previously in the *even-skipped* stripe 2 enhancer. The names of CRMs containing the sequences are given at the right side of the alignments.

**Figure 6.** Hierarchy of transcriptional signals. Periodically spaced groups of binding sites (marks on the blue bar) comprise composite elements. Each composite element is responsible for the formation of a cognate protein–protein complex (the connected blocks). A set of independent composite elements bound by different protein complexes (shown by color) comprises a functional cis-regulatory module (CRM).
proximal promoters), appear to describe the distribution of binding motifs and explain the biological function of motif combinations.

**Genomics approaches and promoter analysis**

Specific signals, detected in the distribution of binding motifs at the ‘middle’ level, could be helpful in finding similar CEs in the genome and improving promoter recognition algorithms. On the basis of our periodic analysis and analysis of overlapping motifs, we generated models (alignments) for two putative CEs, containing Bcd–Bcd (synergistic) and Bcd–Kr (antagonistic) motif combinations (see Fig. 5). The motifs corresponding to the CEs are wider, and they provide better grounds for a specific search than the binding motifs for transcription factors themselves. Pre-screening with CEs might also facilitate identification of functional binding motif clusters and functional CRMs in the genome. Periodic arrangement (e.g. ‘helical phasing’) of binding motifs may also help in finding unknown binding motifs in regulatory sequences. This idea can be implemented in motif-extracting software, such as Gibbs Sampler.

CEs can become indispensable for understanding the transcription regulatory code and for reconstructing entire CRMs and promoters. In this respect, identification of CEs in promoters and CRMs and consequent analysis of these CEs software, such as Gibbs Sampler, might also facilitate identification of functional binding motif clusters and functional CRMs in the genome. Periodic arrangement (e.g. ‘helical phasing’) of binding motifs may also help in finding unknown binding motifs in regulatory sequences. This idea can be implemented in motif-extracting software, such as Gibbs Sampler.

**ACKNOWLEDGEMENTS**

We thank Stephen Small for discussion, valuable remarks and help with manuscript preparation. We would also like to thank our reviewers for their careful work and helpful comments. This work was supported by a grant from National Institutes of Health (EM064864) to Claude Desplan. V.J.M. and A.P.L. were also supported in part by grants from the Ludwig Institute for Cancer Research, HHMI East Europe (55000309) and RFBR (02-04-49111). The CRM annotation and related resources are available at the following address: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm.

**REFERENCES**


