**A PROGRAM FOR CALCULATING GAPPED DINUCLEOTIDE CORRELATIONS IN NUCLEIC ACID SEQUENCES AND ITS APPLICATION TO REVEAL SPECIFIC SEQUENCE FEATURES OF X-LINKED PROMOTERS IN DROSOPHILA.**

ARKHIPOVA IRINA R.1, POKROVSKI SERGEI V.2

1Dept. of Molecular and Cellular Biology, Harvard University, Cambridge MA 02138

2Dept. of Physics, MIT, Cambridge, MA 02139 (Present address: ADI Corp., Framingham MA)

Keywords: gapped nucleotide correlations, X-linked promoters, Drosophila, specific sequence motifs, Drosophila melanogastr promoter sequences

The promoter regions of X-linked genes in Drosophila melanogaster differ from those of autosomal genes at the level of nucleotide sequence. While no conventional consensus can be deduced for the promoter elements, correlation analysis reveals significant over- representation for several classes of relatively simple sequence motifs in certain positions in the X-linked but not autosomal subsets of the aligned promoter database.

These elements are broadly distributed over a distance of several hundred base pairs, extending from the 5' upstream flanking region far downstream into the transcribed region. In addition, TATA-containing promoters are strongly under-represented in the X chromosome subset. Taking into account the strand-specificity and positional distribution of X chromosome-specific sequences, it appears likely that the combined action of these multiple motifs might provide the cis-acting elements necessary and/or sufficient for creating the unique chromatin structure of the male X chromosome in Drosophila and mediating the dosage compensation response.

This is the first report of specific sequence motifs in a set of diverse genes connected only by their linkage relationship. These results might greatly facilitate design of experiments aimed towards investigation of chromosome-specific regulatory elements. The developed programs could prove extremely useful in analyzing data obtained in the course of genome/EST sequencing projects for model eukaryotic organisms.

**Methods of analysis:**

The data sets were first analyzed by "word profiling" as described in [1]. However, this technique allowed us to analyze only the nearest-neighbor correlations. To detect longer-range correlations within nucleotide sequences and to reduce statistical noise, a program for gapped dinucleotide correlation analysis was developed.

The program was written in Mathematica programming language (Wolfram Research, Inc.) and implemented as a series of Mathematica packages. This design provided sufficient flexibility in processing the data and organizing various batch scenarios by combining different packages into the main script. Mathematica as an excellent analytical tool provided us with an extensive library of mathematical, string-parsing and statistical functions, a versatile and powerful graphical library, and especially the comprehensive packages for the matrix calculus and Fourier-transform.

The program is designed to explore medium-range correlations of nucleotide pairs in DNA sequences for persistent patterns. Since such patterns can vary in composition as well as location, we calculated the probabilities with which different nucleotide combinations can be found at different positions of the nucleotide sequence. Specifically, relative occurrence frequencies for 16 different pairs of nucleotides (AA, AC, etc.), not necessarily adjacent to each other, were counted in a set of mutually aligned DNA sequences for various locations of both nucleotides relative to the reference point of the sequence alignment. The inclusion of complementary pairs allowed us to detect any strand-specificity. The normalized counts (frequencies) for each pair were plotted as functions of the distance L from the reference point and the separation D between the nucleotide constituents of the pair considered. Both three-dimensional and level (topographic) maps were generated. The distance between the nucleotides usually varied in the range from 1 to 75 (1 to 30 in most of the calculations). The length of the typical sequence in a set varied from 600 to 2000 in different sets.

To eliminate such effects as a lack of data or omissions in many of the available sequences (especially close to the end of the sequence), the pair counts were normalized locally for each distance from the reference point and for each distance between nucleotides with respect to the total number of pairs containing the actual data.

In some cases, the base composition bias can obscure the dinucleotide correlations. To avoid this, pair counts were normalized by the base frequencies. Namely, the program counted the quantities FAC(XA,XC)/(FA(XA)\*FC(XC)), where FAC(XA,XC) is the frequency with which the pair of nucleotides A and C was found at the location XA for nucleotide A and XC for nucleotide C, respectively, whereas FA(XA) and FC(XC) denote base frequencies for A and C at the same locations. The base frequencies were also normalized locally with respect to the total counts containing the actual data, as described above for pairs.

Due to the relatively small number of sequences in the set, it was necessary to reduce the statistical noise in order to obtain acceptable visual representation of the results. We have applied a standard convolution technique with a normalized Gaussian kernel for smoothing the oscillating dependence of frequencies on the distance from the reference point. The calculations were performed for several choices of the smoothening length in order to optimize visual representation and to reveal characteristic features of the correlation functions in question. Empirically, the optimal length proved to be of the same order as the maximal separation between nucleotides. Technically, convolution has been performed as a combination of a discreet fast Fourier and inverse fast Fourier transforms with an intermediate multiplication by the Fourier-transformed Gaussian kernel (which becomes a diagonal matrix).

We have also applied a roughening mapping to the arrays of smoothed frequencies, mainly because the alignment of different sequences in any of the sets examined could not be established exactly. The displacement of reference points could reach 5-10 units. Therefore, several adjacent points having the same separation between nucleotides were combined into clusters. Each cluster was replaced by a single point with a value of the correlation function at this point equal to the average of the function over the cluster. This roughening procedure also facilitated plotting of the function, reducing the number of points. The optimal cluster size was found to be 5-10 units. Since the cluster size was always smaller than the characteristic length of the initial convolution, it did not significantly affect the general appearance of the plots.

**Results and discussion**

The analysis was carried out for 300 Drosophila melanogaster promoter sequences (expanded dataset [1]) including 500 bp upstream and 1000 bp downstream from the RNA start site. The database was divided into four chromosome-specific subsets of comparable size (60-70 entries): X, 2L+2R, 3L, and 3R. A subset of 60 X-linked promoters displayed several features of sequence organization which were clearly different from that of the second and third chromosomal subsets of approximately the same size, whereas the differences between the latter subsets were much less significant. These features, each of which is detectable in more than a half of X-linked promoters, are as follows:

(1) X-linked promoters tend to have non-canonical TATA boxes or not to have any. Two- thirds of the entries can be definitely classified as TATA-less, since no AT-rich sequence whatsoever can be detected at a distance 20-35 bp upstream from the RNA start site. In the remaining one-third, those AT-rich motifs which are present do not bear much resemblance to the TATA box consensus, and only three promoters (achaete, sgs4, and yellow) contain the canonical TATAAA motif. This is in striking contrast to the autosomal subsets. Stretches of adenine three and more nucleotides in length seem to be particularly disfavored in the region that should correspond to the TATA box.

(2) At least twofold enrichment in AT-rich sequences is observed in a wide area centered at 300 bp upstream from the RNA start site. Correlations between alternating A and T are revealed over distances not exceeding 15 bp. None of the autosomal subsets display such pattern.

(3) An enrichment in GT-containing sequences is observed predominantly in the downstream region. These sequences are concentrated strand-specifically (GT not AC; TG not CA; TT not AA; GG not CC) in several preferred locations throughout the promoter, upstream as well as within the transcribed region (centered at positions -320, +110, +240, +700). Although they are frequently organized into alternating (GT)n stretches in individual sequences, removal of several entries with the longest stretches does not significantly change the overall pattern, indicating that the observed over-representation does not result from a few long stretches. The regularity can be revealed over distances not exceeding 30 bp. Other arrangements of G and T, such as TTG, GTT or TTTG, are also frequent in these regions, and may be spaced at certain intervals. To a lesser extent, a non-periodic strand-specific enrichment in AG is observed in adjacent but non-overlapping downstream regions (centered at +140 and +280).

(4) Another impressive difference is a correlation between C and G at the positions +100 (CG, CNG and CN6G) and +700 (CNG). This correlation appeared so strong that it stimulated the analysis of an additional subset of X-linked genes. This subset, less reliable with respect to alignment, was created by extracting GenBank Drosophila entries which consist only of the transcribed regions (usually but not necessarily including introns) and in which the position of the RNA start site was not determined by primer extension but rather taken as the 5' end of the most upstream cDNA. Therefore, some of these sequences may be 5'-truncated or intronless, but none of those would contribute to any positioned motifs. Surprisingly, the analysis of this additional non-overlapping subset of 70 transcribed regions revealed exactly the same pattern of correlations for CG, CNG and CN6G.

The above observations indicate that the method of analysis is valid, even for such a relatively small size of data sets, since the same patterns were obtained for two independent non- overlapping X-linked subsets. In addition, previously described correlations with a period of 3 resulting from triplet occurence within coding regions [2] are readily visible for certain dinucleotide combinations. The fact that analysis of the 5' cDNA ends can yield useful information should make this approach applicable to genome/EST sequencing projects.

A special interest in analysis of X-linked promoters is well motivated. Transcription of most X-linked genes in Drosophila is known to occur with a twofold intensity in a single male X chromosome, as compared to the two female X chromosomes (dosage compensation). It is associated with unique features of chromosomal architecture, such as the diffuse cytological appearance of the male X and the existence of a number of proteins which bind to numerous sites on the hyperactive male X chromosome. However, little is known about the cis-acting elements which contribute to creation of these structural features and generation of the dosage compensation response, although it is well-documented that such elements do exist and there is also evidence that they might be represented by multiple sequences [3,8].

There is evidence in support of the involvement of the above-mentioned sequence motifs in the architecture of Drosophila chromosomes. Three of the ten possible mono- and dinucleotide stretches (GT/AC, GA/TC, and C/G) occur on the dosage-compensating chromosomes with a twofold higher frequency than on the autosomes [4,7]. The same three motifs also exhibit the correlations described above. Moreover, they appear to be strand-specific and positioned with respect to the RNA start site. In addition, the highly recombinogenic GT/AC stretches are not found in heterochromatin, and neither is dosage compensation and meiotic recombination known to occur there [7]. The GAGA factor might be a good candidate for binding to GA/TC motifs. An enrichment of the 3' untranslated regions with oligo(T) stretches, which may serve in RNA as binding sites for the Sex-lethal gene product, was also observed by other authors in 20 X- linked genes and proposed to constitute part of an alternative dosage compensation pathway [5].

The CNG and CN6G correlations detected in this study are of particular interest, since they do not appear to be as repetitive as other simple motifs and were not previously described as significant in DNA(RNA)-protein interactions. It is worth noting that analysis of the dosage- compensated Arr-B gene of D. miranda revealed a heptanucleotide TGGGCNR, repeated five times, which is absent from its non-compensated D. melanogaster counterpart and was pointed out as a potential cis-acting element [6].

It remains to be seen whether the products of any of the msl (male-specific lethal) or other genes controlling dosage compensation in Drosophila are able to bind to any of the motifs described above. Simple sequence motifs are often implied in chromosomal architecture, and many nuclear proteins can bind to such motifs. The strand-specificity observed in the present study for some of the motifs might be interpreted as their involvement in RNA-protein interactions or in rotational positioning of DNA-protein complexes.

The role of the basal promoter elements, such as the TATA box, and the basis for its under-representation in X-linked promoters remains unclear. One might speculate that the TATA-box-binding protein somehow disfavors interactions between basal promoter elements and proteins mediating the dosage compensation response. Nor it is clear how certain sequence motifs accumulate in specific promoter regions in the course of chromosome evolution on a gene-by-gene basis, especially upon translocation of entire chromosomal arms involving hundreds of genes.

Development of experimental approaches to analysis of higher-order chromatin structure in vivo should make it possible to determine whether any or all of the sequence motifs described in this report are necessary and/or sufficient for creation and/or maintenance of unique features of X-chromosomal architecture in Drosophila. The correlation analysis technique described here also opens perspectives for analysis of aligned data sets for different groups of genes in attempts to understand long-range correlations in DNA sequences potentially associated with specific in vivo structures during development and differentiation.

**References:**

1. Arkhipova, I.R. (1995) Genetics 139: 1359-1369.
2. Arques, D. and C. Michel (1992) J. Theor. Biol. 156: 113-127.
3. Cooper, M.K. et al. (1994) Genetics 138: 721-732.
4. Huijser, P. et al. (1987) Chromosoma 95: 209-215.
5. Kelley, R.L. et al. (1995) Cell 81: 867-877.
6. Krishnan, R. and R. Ganguly (1995) Gene 160: 185-190.
7. Pardue, M.L. et al. (1987) EMBO J. 6: 1781-1789.
8. Qian, S., and V. Pirrotta (1995) Genetics 139: 733-744.