BMC Bioinformatics

Research article



# Statistical extraction of *Drosophila cis*-regulatory modules using exhaustive assessment of local word frequency

Anna G Nazina and Dmitri A Papatsenko*

Address: Department of Biology, New York University, New York, USA

Email: Anna G Nazina - agn202@nyu.edu; Dmitri A Papatsenko* - dap5@mail.nyu.edu

* Corresponding author

Published: 22 December 2003

*BMC Bioinformatics* 2003, **4**:65

Received: 19 July 2003

Accepted: 22 December 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/65



## Abstract

**Background:** Transcription regulatory regions in higher eukaryotes are often represented by *cis*-regulatory modules (CRM) and are responsible for the formation of specific spatial and temporal gene expression patterns. These extended, ~1 KB, regions are found far from coding sequences and cannot be extracted from genome on the basis of their relative position to the coding regions.

**Results:** To explore the feasibility of CRM extraction from a genome, we generated an original training set, containing annotated sequence data for most of the known developmental CRMs from *Drosophila*. Based on this set of experimental data, we developed a strategy for statistical extraction of *cis*-regulatory modules from the genome, using exhaustive analysis of local word frequency (LWF). To assess the performance of our analysis, we measured the correlation between predictions generated by the LWF algorithm and the distribution of conserved non-coding regions in a number of *Drosophila* developmental genes.

**Conclusions:** In most of the cases tested, we observed high correlation (up to 0.6–0.8, measured on the entire gene locus) between the two independent techniques. We discuss computational strategies available for extraction of *Drosophila* CRMs and possible extensions of these methods.

## Background

Recognition of transcription regulatory sequences is one of the most important and challenging problems in modern computational biology. In the case of higher eukaryotes, there are proximal transcription regulatory units, located close to 5' ends of coding sequences and called 'proximal promoters', and distant transcription regulatory units, located further upstream or downstream of the gene and called 'enhancers' or '*cis*-regulatory modules' (CRMs). It is clear that identification of a 'proximal' transcriptional unit can be based on its relative position to the coding sequence and the presence of specific transcriptional signals such as TATA box, CAAT box, transcription start site consensus (TSS) and, perhaps, other specific signals (such as downstream promoter elements, DPE). Typical CRMs (or enhancers) possess no such specific features; therefore their annotation in genome is much more difficult.

Currently existing methods dedicated to the recognition of transcription regulatory regions can be subdivided into three main categories: (*i*) search by signal, (*ii*) search by content and (*iii*) phylogenetic footprinting [1-4]. Modern 'search by signal' techniques are based on identification of known transcriptional patterns in DNA sequences, such as clustered binding motifs for known transcriptional regulators [5-10]. Extraction of clustered recognition motifs is among the most reliable current techniques, but it is

limited to recognition of *similarly regulated cis*-regulatory modules in a genome.

Another strategy of CRM extraction from genome is phylogenetic footprinting. Methods of this class assume that regulatory regions contain highly conserved segments and they can be extracted by means of sequence comparison from evolutionary related genomes [11-18]. Performance of the phylogenetic footprinting greatly depends on the evolutionary distance between chosen species and on the conservation level of particular genes from these organisms. Phylogenetic footprinting have become especially important in recent days as more than one genome represents the sequence data for most of the main model organisms. However, it is not clear yet whether phylogenetic footprinting alone is sufficient for *precise* and *exhaustive* mapping of CRMs and how many related genomes it will require to achieve this goal. Non-coding conserved regions might also include not only promoter and enhancer regions, but also other functional sequence classes, such as origins of replication, matrix-attached regions etc, so an independent method of CRM extraction might be necessary.

'Search by content' (*ab initio*) methods are often based on the difference in the local base composition and in the local word composition between the regulatory and non-regulatory DNA [10,19-21]. It is assumed that the difference is caused by presence of transcriptional signals, such as binding motifs for transcriptional regulators in the regulatory DNA. For example, the presence of multiple copies of the same binding site may change local frequency of short words in promoter regions. This idea was explored by analysis of most frequent hexamers (differential hexamer frequency) [22], other short words and motifs [23,24] in regulatory sequences. More recent implementations of the 'search by content' strategy take into account base interdependence in transcription regulatory regions and exploit interpolated Markov chains [19] as well as local word frequency [25]. General-purpose techniques based on the 'search by content' are of great interest as they provide *an independent* line of evidence for the recognition of transcription regulatory sequences in genome.

Recognition of regulatory sequences using 'search by content' is still a difficult problem due to the presence of distinct signals in different regulatory sequences as well as high divergence of these signals themselves. In this respect, each particular promoter as well as a given large training set may never contain even a small fraction of all specific words. Beside the binding motifs for transcription factors, the regulatory DNA may also possess patterns with specific physical properties [2,21] or other functional patterns, such as nucleosome positioning signals [26,27].

The described diversity of promoters and transcription regulatory signals suggests that methods based on local word composition/local word frequency, regardless of the words themselves, might be more suitable for the CRM and promoter recognition. In this work we describe an *exhaustive* local word frequency analysis and apply this new technique to recognition of *cis*-regulatory modules of *Drosophila* developmental genes. We based our recognition strategy on the assumption that biological signals in any regulatory region possess similar redundant properties, independent on a particular gene, promoter or CRM. According to this assumption, words corresponding to binding motifs for transcription factors might possess one level of redundancy, specific words involved, for instance, in promoter bending might possess another etc.

To describe all these various specific patterns in the context of one DNA sequence segment (resolution window), we represent *exhaustive* statistical analysis of word frequency, where the local frequency of each word in the DNA segment is taken into consideration, *regardless* of the word itself. For instance, two DNA segments having similar local word frequencies would produce similar scores, even if the words comprising the two DNA segments were different. In this respect, the proposed strategy aims to identify sequence segments containing specific *word distributions* rather than sequence segments containing specific words or arrays of specific words. In this sense our new method resembles promoter recognition techniques based on local assessment of most frequent words (hexamers) [22], but it is superior as it takes into consideration the *full range* of word frequency and scores even unique words.

Typically, the efficiency of a promoter recognition algorithm is evaluated with the help of an annotated promoter database, such as EPD [28]. The database contains proximal promoter elements, and is very helpful for evaluation of methods based on recognition of *specific* words, such as proximal transcriptional signals represented by the TATA box, transcription start site (TSS) and others. However, a vast majority of eukaryotic transcription regulatory regions located far upstream or far downstream of TSS contain no proximal signals. Therefore, to test the performance of the proposed new technique we developed a unique training set of sequence data, based on *cis*-regulatory modules (CRMs) of *Drosophila* developmental genes, transcription regulatory regions from higher eukaryotes, located far from gene coding sequences and TSS.

The developmental genes of *Drosophila* comprise spatio-temporal cascade (network) of regulatory interactions, responsible for early pattern formation of the developing fly embryo [29-32]. This model system has several advantages that make it unique in computational sequence

analysis: (*i*) The majority of the genes encode transcription factors that are connected into the network of *direct* transcriptional interactions. (*ii*) For most of the developmental genes (Bicoid, Hunchback, Krüppel, etc) large amount of experimental data is available at the genetic, biochemical and evolutionary (comparison between species) levels, including positions of their cis-regulatory modules and description of binding motifs for upstream regulators.

We compared the results of our word frequency analysis of *Drosophila melanogaster* developmental genes with a reference data, generated by phylogenetic footprinting (percent identity profiles, PIP) of the same genes using recently sequenced genome of *Drosophila pseudoobscura*. This comparison has revealed a striking agreement between predictions generated by the two independent methods.

## Results

### *Construction of interactive CRM annotation*

Recognition of regulatory sequences using 'search by content' approach requires representative training set of sequence data, ideally, a set of *functional* transcription regulatory regions. For this reason we assembled and annotated all published experimental data for the early developmental enhancers of *Drosophila*. These data include three major types of information: (*i*) sequence data for experimentally tested CRM regions, (*ii*) binding motifs for known transcriptional regulators, and (*iii*) known regulatory interactions between the early developmental genes.

To assemble the CRM sequence data, we retrieved published deletion analysis data for more than 60 CRMs from 20 different developmental genes. To navigate through this sequence collection, we created an interactive database containing two major structural levels (see Figure 1). The top level comprises a list of genes, and references to the corresponding sources in the literature. The bottom level consists of an interactive map, describing the position of all known functional elements (coding sequences and enhancers) in each gene locus (loci ranged in size from 16 to 120 Kb). The exact location of each functional element is presented in a table below the map, along with a description of any known regulatory interactions mediated by the element. The bottom level also contains an annotated text of the locus sequence with highlighted functional regions. In addition, we aligned footprint data for 28 binding motifs, representing the majority of the maternal, gap, pair rule, and some segment polarity genes. For each alignment, we established the optimal motif width, and outlined a well-defined core formed by positions with high information content. This data is also available from our interactive database.

In comparison with existing dedicated databases, such as GeNet [33,34], our compilation is focused on available deletion and footprinting data and it is convenient for: (*i*) fast navigation and retrieving of CRMs and (*ii*) fast retrieving of the binding motifs involved in a given regulatory interaction. All annotated data are publicly accessible from New York University web site http://homepages.nyu.edu/~dap5.

### *Construction of positive and negative training sets*

We considered it irrelevant to construct positive training set from the entire CRM collection directly as the positions of the CRM boundaries identified by empirical deletion analysis are quite arbitrary. In most of the classical deletion studies the boundaries of a deletion fragment were defined according to the presence of convenient restriction sites. This fact, along with the limit on a possible number of deletion combinations (number of transgenic constructs) resulted in a lack of precise resolution of the deletion analysis technique. In many cases the identified by the deletion technique minimal regulatory elements (such as MSE – minimal stripe enhancers) still provide correct spatial distribution of expression patterns, but the rescued patterns contain only a fraction of the endogeneous gene expression levels.

To minimize the effect of possible errors caused by the insufficient resolution of the deletion technique and to develop a formal principle of the positive training set construction, we defined CRM boundaries based on position of clusters of binding sites for transcription factors, involved in the CRM regulation (see Figure 2). In our previous work [9], we have demonstrated that positions of binding site clusters for these regulators correlate well with the positions of the CRMs, identified previously using deletion analysis. In a number of related studies [7,8,10] it has also been demonstrated that the binding motif clusters can be effectively used for mapping CRMs in the genome. Therefore, we constructed our positive training set from sequences containing the most significant clusters of binding sites for five transcription factors (Bicoid, Hunchback, Krüppel, Knirps and Caudal), involved in the regulation of many genes in our CRM database. The vast majority of these clusters overlaps the deletion analysis data, as it has been demonstrated earlier [9].

Despite the fact that we used only five transcription factors to evaluate sequences for a positive training set, the resulting training sequences contained many other binding motifs (yet unknown), present in CRMs. Thus, the five chosen motifs served us as markers only, indicating positions in the sequences that belong to *cis*-regulatory modules. The total size of the described positive training set comprised more than 68 Kb of sequence data and

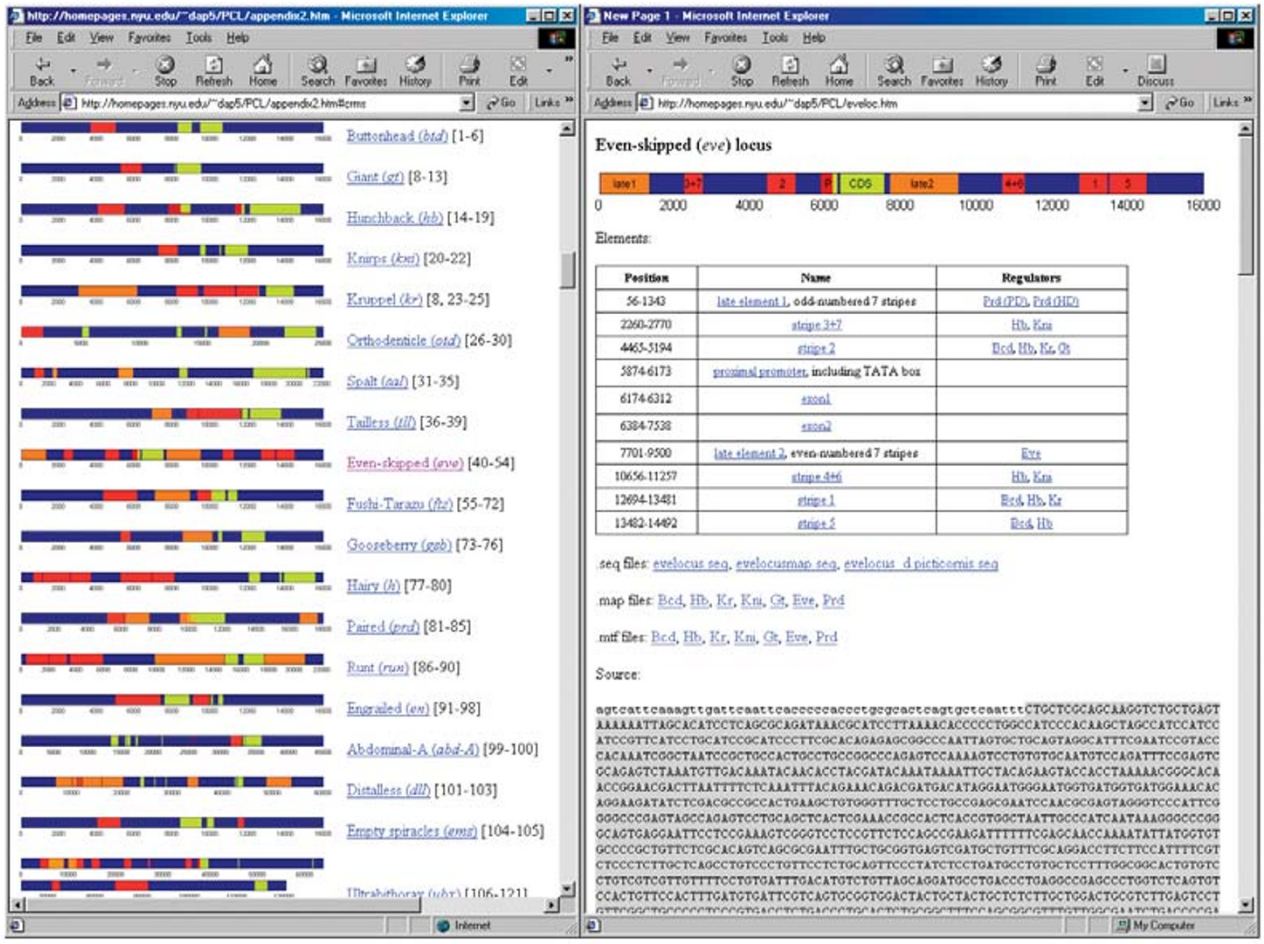


**Figure 1**
**Interactive collection of CRMs.** Window on the left shows the upper level of the annotation with the list of genes and references, window on the right displays the bottom level containing interactive functional map of a gene locus and the sequence of the locus with highlighted functional regions. Red bars correspond to early and orange bars to late CRM regions; yellow bars mark exons. Exact positions of the highlighted functional regions are given in the table below the interactive map. Binding motifs are available from the upper level of the annotation.

contained 58 homotypic clusters in 33 non-overlapping contigs. Sequences of the positive training set, including identified homotypic clusters in the gene locus regions are available from our database: http://homepages.nyu.edu/~dap5/PCL/pseudoobscura/train_plus_contigs.zip.

We also generated several negative training sets, one containing random samples (50 Kb each sample) from genome of *Drosophila melanogaster*, one containing random samples from *Drosophila* CDS collection [35] and one containing non-coding sequences only. Each negative training set combined >2 Mb of sequence data (>1.5% of the *Drosophila* genome).

### *Multiresolution analysis of word frequency*

One can describe frequency $F$ of a word in $i$th position of a DNA sequence through a number of matches found for this word inside a $w$-window (detection window), centered on the word (See Figure 3, 3A). Presence of binding motifs and other transcriptional signals affects the word frequency, but, in most cases, it is not known which functional signal corresponds to what frequency level. Therefore, consideration of only words with a given frequency level, for instance 'most frequent' words (high $F$), may result in a loss of functional information. To collect the information exhaustively, we take into consideration local frequency $F$ of each word (or each position) in a

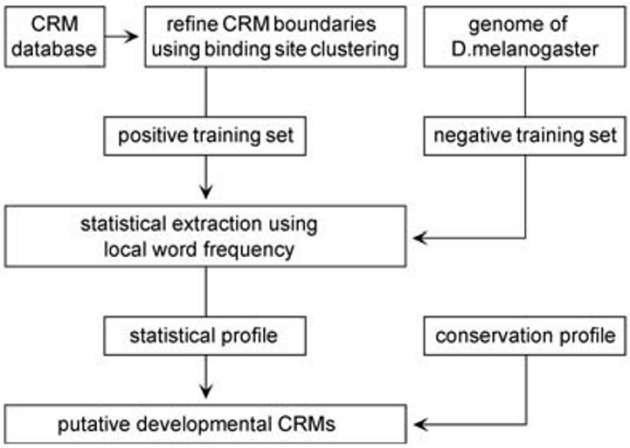


**Figure 2**
**Strategy of CRM extraction.** Input training data for word frequency analysis represent genome of *D. melanogaster* and CRM sequences refined by search for regulatory clusters using binding motifs for transcription factors: Bicoid, Hunchback, Krüppel, Caudal and Knirps [9]. The results of word frequency analysis are compared with results of phylogenetic footprinting.

sequence segment. Then, we group (sort) all words in a sequence segment according to their frequency $F$ into separate *frequency channels*, where the *feature score* $S_j(i)$ in a frequency channel $j$ represents the total number of words (positions) in a window $l$ (resolution window) having a frequency from a fixed range of the frequency value: $j \le F < (j + n)$, $(j, n \in N)$ (See Figure 3, 3B). For instance, if $j = 1$, $n = 0$, the feature score $S_j(i)$ of the resulting frequency channel represents the total number of unique words (having $F = 1$) in a resolution window $l$. In contrast, high values of $j$ describe frequency channels that account for highly repetitive words. High values of $n$ define 'wide' frequency channels, combining words with distant frequency values. Combination of all non-overlapping frequency channels in the range of $w > F > 0$ covers full spectrum of word frequency detected in the window $w$. Independent consideration of the feature score $S$ in each frequency channel represents *multiresolution* analysis of our main feature, word frequency.

A partition of the spectrum into non-overlapping frequency channels may be set differently and highly depends on other parameters such as the size of the resolution window and the size of the word. In the current work we considered only short words (2–4 bases), assuming that a functional transcriptional signal (for instance, transcription factor binding site), must have a conserved core, which is often only a small fraction of the functional word [36]. Another advantage of short words is that they have much greater local frequencies, which facilitates statistical analysis and allows construction of spectra with larger number of wide non-overlapping frequency channels. For the same reason, wider windows ($w$, $l$) are preferred, but the upper size of the window is limited by the desired resolution of the method. In the case of our model system, *Drosophila* CRMs, we considered equal detection and resolution windows ($w = l$) that are close to the size of the minimal known CRM sequences and vary within the range of 0.5–1 Kb.

### *Statistical discrimination of sequences with distinct word frequency*

Based on the feature score values $S$, one can built a statistical model describing DNA sequence segments belonging to a specific functional class, for example, to a CRM or to a coding/non-coding sequence. To generate the statistical model for a functionally related class of sequences (for positive or negative training sets) we built distribution $E$ of the feature score $S$ in each frequency channel, where $E_s$ is the fraction of all windows $N$ in a training set, having the same score $S$. If $N_s$ is the number of windows with the score $S$, then:

$$E_s = N_s / N \quad (1)$$

Mathematical expectations $M(S)$ and distributions of the feature score $E(S)$ obtained for positive and negative training sets are given in Figure 4A, 4B. In some training sets we obtained distributions that are very close to *Poisson* in several frequency channels. However, that was not common, especially in the case of frequency channels accounting for unique or very frequent words. For this reason, we have used the $E$-values, obtained directly from the distributions that were smoothed to reduce statistical noise.

Given two functional sequence classes $\omega_1$ and $\omega_2$ (positive and negative training sets) and a feature score value $S_j$, from a test sequence, we now solve the recognition (classification) problem for each frequency channel independently, using simple log-likelihood ratio test [37]:

$$L_j = log\left(\frac{E(\omega_1 \mid S_j)}{E(\omega_2 \mid S_j)}\right) \quad (2)$$

Thus, in each frequency channel $j$, each position of the test sequence $i$ obtains a log-likelihood score $L_{ji}$, which displays chances that the $i$-th window of the test sequence belongs to a positive ($\omega_1$) or a negative ($\omega_2$) training set. In this work, we consider only the simplest case, where the contribution of each frequency channel into the recognition is equal (equal weights). Therefore, we calculated the resulting score $R_i$ for the $i$-th window as a sum of scores $L_{ji}$

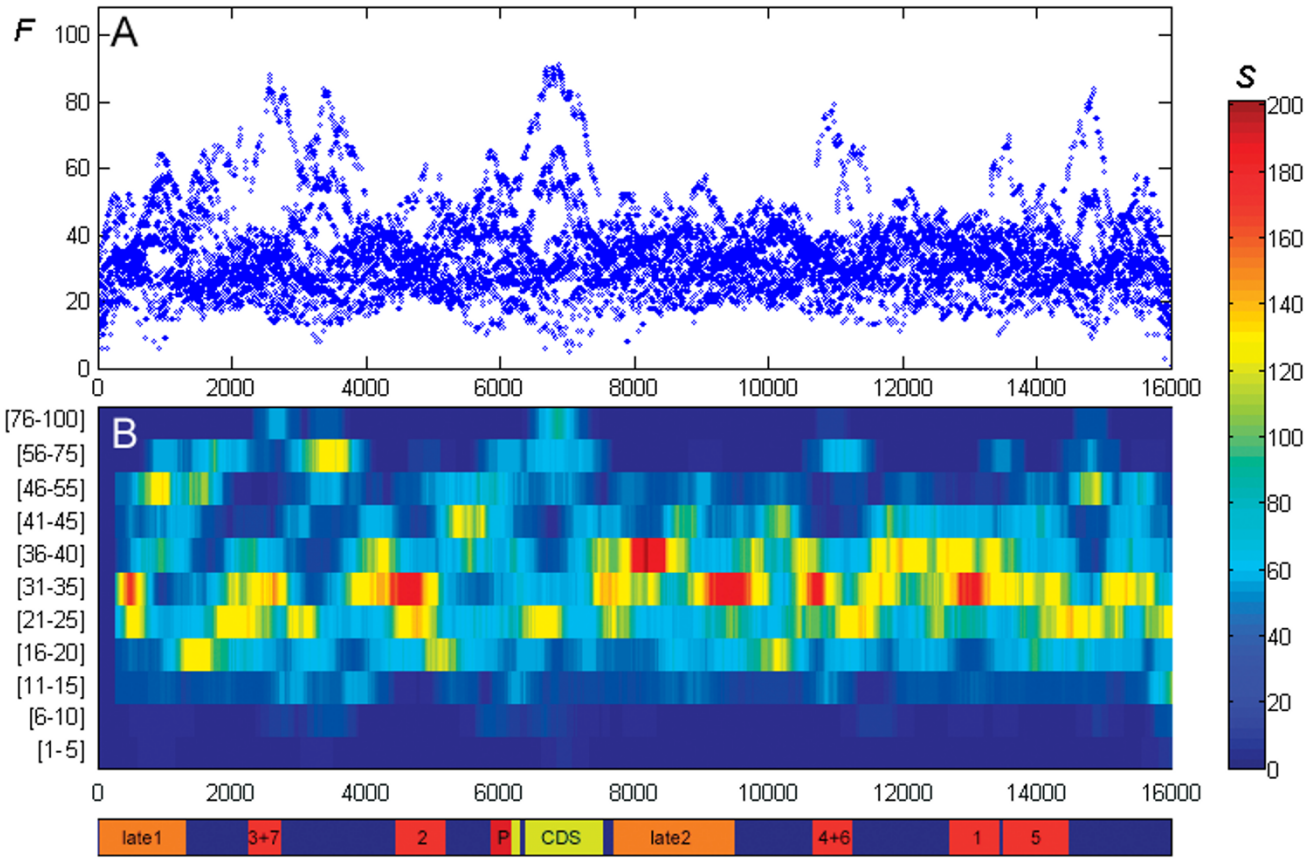


**Figure 3**
**Local word frequency algorithm.** (*A*) Degree of local (detection window *w* = 500) word frequency *F* (*Y*-axis) in every position of the *even-skipped* locus, generated for words n = 2. (*B*) Feature score values *S* (see colorbar) obtained from the LWF analysis of the same DNA fragment. Each frequency channel (*Y*-axis) accounts for words in a limited range of word frequency only (shown in brackets). Feature score values represent total number of words, having similar *F* in a resolution window *l* (in the case shown *l* = 500).

for all frequency channels. On practice, we used more sophisticated formula (see below) to calculate and correct the score $R_i$, but even the simplest approximation produced relevant results.

### *Extraction of* Drosophila *CRMs using local word frequency*

To assess the quality of our word frequency classification algorithm on the same system of early *Drosophila* genes, we adopted an additional reference set of data that *independently* marks positions of transcription regulatory modules in the analyzed sequences.

To construct this independent reference dataset we took advantage of the existing tools and strategies [13,16] and generated conservation profiles (percent identity profiles, PIPs) for wide genomic regions of 16 early developmental gene loci, representing the majority of genes from our database. We retrieved from the recently published genome of *D. pseudoobscura* (Human Genome Sequencing Center at Baylor College of Medicine) all contigs corresponding to the 16 selected genomic regions of *D. melanogaster*, and produced the conservation profiles using available standard procedures. The assembled sequences for the developmental genes of *D. pseudoobscura* along with the graphical data for alignments and numerical data for conservation profiles are publicly available from New York University web site http://homepages.nyu.edu/~dap5/PCL/pseudoobscura/pseudoobscura.htm.

We measured correlation (Pearson Association Coefficient) between the likelihood profile obtained from the

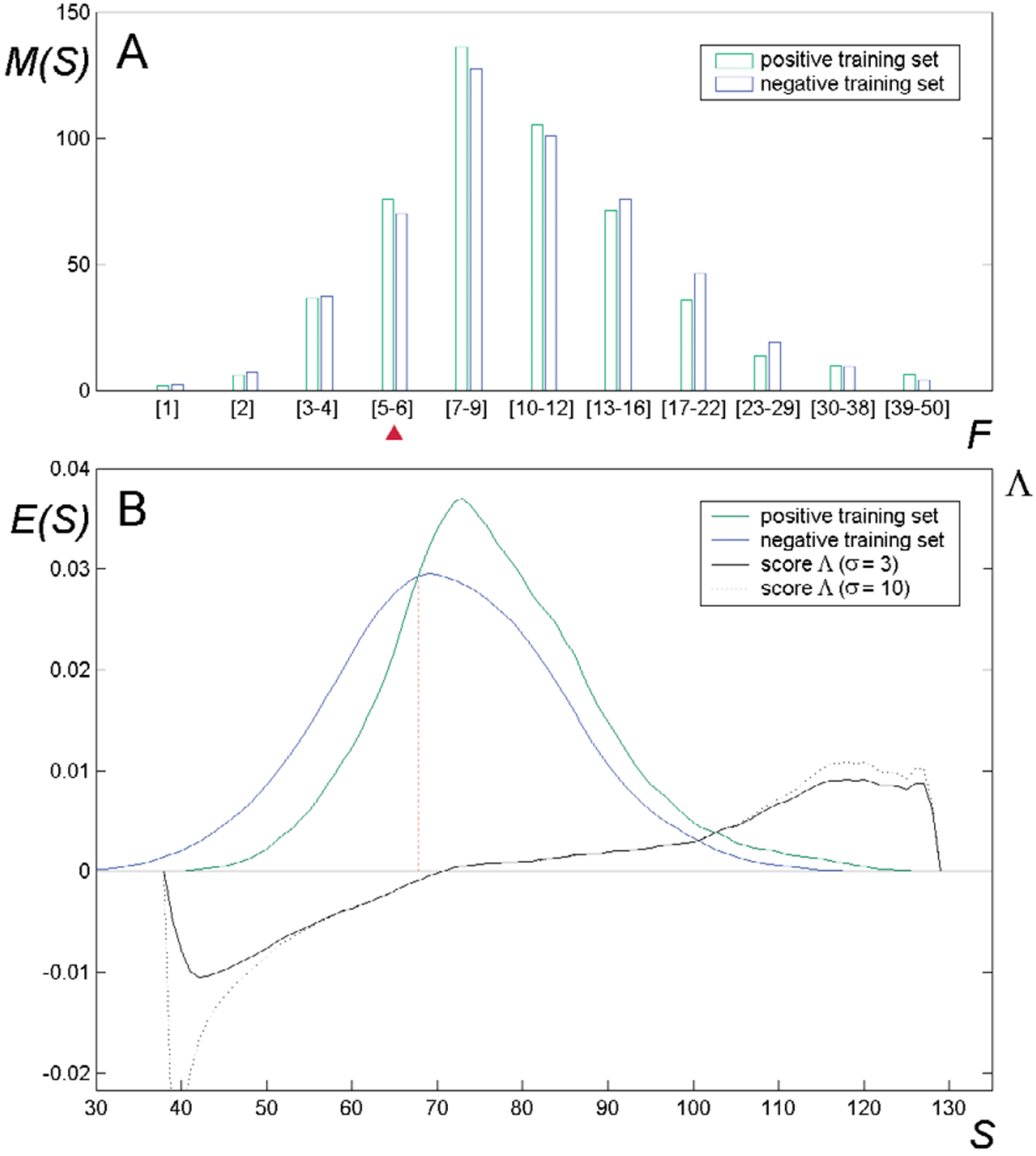


**Figure 4**

**Classification of a sequence segment.** (*A*) Mathematical expectation for the feature score *S* in different frequency channels (*X*-axis, shown in brackets), calculated as $M(S) = \Sigma\, S_k\, E\,(S_k)$ from distribution $E(S)$ for words n = 3. (*B*) Distribution $E(S)$ of the score *S* in the frequency channel corresponding to $5 \le F < 7$ and the resulting likelihood score $\Lambda\,(S)$. Effect of corrections applied to the final likelihood score is shown for two values of σ.

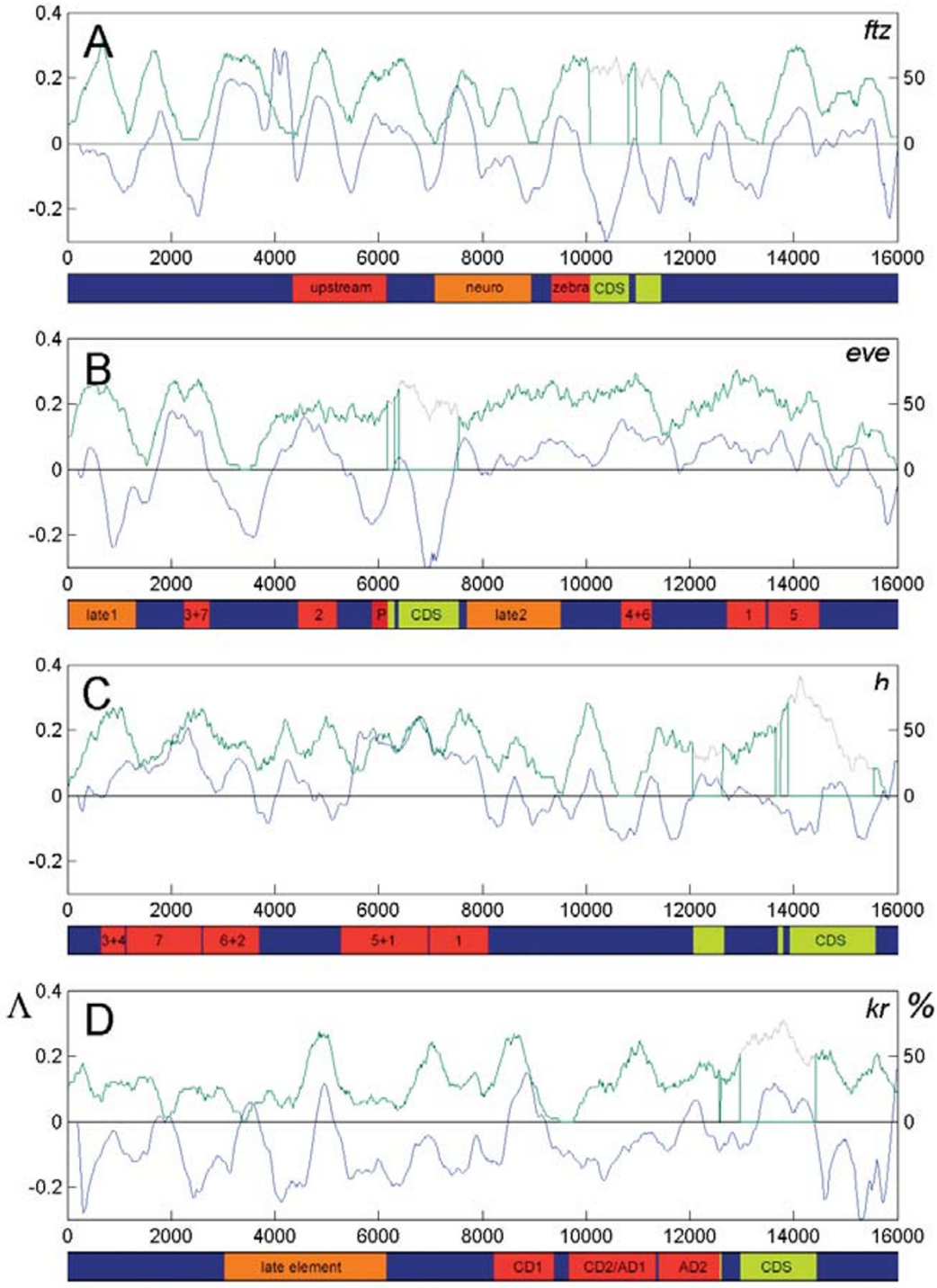


**Figure 5**

**Extraction of CRM sequences using local word frequency algorithm.** (*A-D*) The statistical LWF profiles (Λ) are compared with the conservation profiles (PIP) constructed for loci of several developmental genes of *D. melanogaster*. Map of functional CRM regions (deletion analysis data) is given on the bottom of each panel. Correlation (*cc*) between the statistical (blue line) and the conservation profiles (green line) was measured for the entire locus sequence (see Table 1). Notice that we put percent of sequence identity to zero in exons (yellow bars in the functional map), thus penalizing the correlation value if the word frequency algorithm produced positive score in these regions (see *B*).

**Table 1: Correlation between the LWF and the conservation profiles. The table shows values of Pearson Association Coefficient (*cc*) [41], measured on the entire gene loci sequences (gene names are in the first column). The best correlation (fourth column) was observed between the statistical profiles (Λ) obtained from word frequency analysis and the conservation profiles (PIP). Lower correlation between ether the statistical (Λ/deletion) or conservation (PIP/deletion) profiles with the deletion data supports low resolution of deletion analysis (see results, 'construction of positive and negative training sets'). Jackknife test results are shown (in red typeface) for selected loci that contributed the largest fraction of sequences to the positive training set. The corresponding profiles are shown in Figure 7 [see Additional file 1].**

| Locus | size (KB) | Λ/deletion | *Cc* Λ/PIP | PIP/deletion |
|---|---|---|---|---|
| *ftz* | 16 | 0.12/0.18 | 0.66/0.81 | 0.14 |
| *gt* | 15 | 0.36 | 0.65 | 0.30 |
| *eve* | 16 | 0.30/0.25 | 0.63/0.57 | 0.56 |
| *kni* | 14 | 0.32 | 0.54 | 0.24 |
| *prd* | 16 | 0.18 | 0.52 | 0.04 |
| *h* | 16 | 0.73/0.64 | 0.47/0.41 | 0.46 |
| *sal* | 22 | 0.31 | 0.40 | 0.31 |
| *ems* | 16 | 0.26 | 0.37 | 0.08 |
| *gsb* | 16 | 0.16 | 0.36 | 0.02 |
| *tll* | 16 | 0.22 | 0.35 | 0.20 |
| *en* | 16 | -0.13 | 0.34 | 0.23 |
| *otd* | 25 | 0.23 | 0.32 | 0.11 |
| *run* | 22 | 0.23 | 0.31 | 0.15 |
| *hb* | 16 | 0.48 | 0.31 | 0.06 |
| *btd* | 16 | 0.18 | 0.27 | 0.11 |
| *kr* | 16 | 0.11/0.08 | -0.02/-0.01 | 0.17 |

word frequency analysis and the conservation profile, both constructed *independently* for each of the 16 developmental gene loci from our annotation. We also measured correlation of the likelihood profile and the conservational profile with the positions of annotated CRMs (deletion data). Some of the described results are shown in Figure 5; correlation values for the loci of 16 developmental genes are given in Table 1. One can see that the both independently constructed profiles correlated with each other very well in almost all cases. This important finding suggests that genomic DNA segments possessing word frequency of *regulatory clusters* (positive training set) correspond to *highly conserved regions*.

We performed the described word frequency analysis using three different negative training sets: random samples (50 Kb sample size, total of 2 Mb sequence data) from genome of *Drosophila*, random samples from *Drosophila* cDNA collection and genomic samples containing non-coding DNA. Genomic samples and non-coding sequences resulted in better correlation of the likelihood profile with the conservational profile and with the deletion data (annotated CRMs) than coding sequences. Figure 7 [see additional file 1] shows comparison of performance of LWF algorithm, trained on different datasets. We also compared prediction accuracy of LWF algorithm with prediction accuracy of phylogenetic footprinting. Results of a benchmarking test shown in Figure 6 demonstrated better performance of LWF in extraction of known CRM regions (deletion data). A combination of the two independent approaches allowed further improvement of the prediction accuracy (see Figure 6, 6C).

## Discussion

### *Cross validation of cis-regulatory modules using different computational techniques*

Exhaustive mapping of the exact CRM distribution in developmental genes of *Drosophila* is an extremely important biological task. In its turn, evaluation of novel regulatory sequences in this particular system leads to generation of better training sets for further genome-wide CRM recognition and facilitates computational mapping of binding motifs and other transcriptional signals comprising CRMs.

In this work we compared outputs of two independent computational techniques, the word frequency analysis and the phylogenetic footprinting and have shown that both methods generate highly correlated predictions. Most surprising is that the correlation between the two independent approaches is much higher than the correlation of either one with the deletion analysis data (annotated CRMs). In some cases lower correlation between positions of predicted and the known CRMs (deletion analysis) is explained by the conservative design of our formal tests (see Table 1 and Figure 6). We assumed that in each locus (*i*) all CRMs are known and (*ii*) their

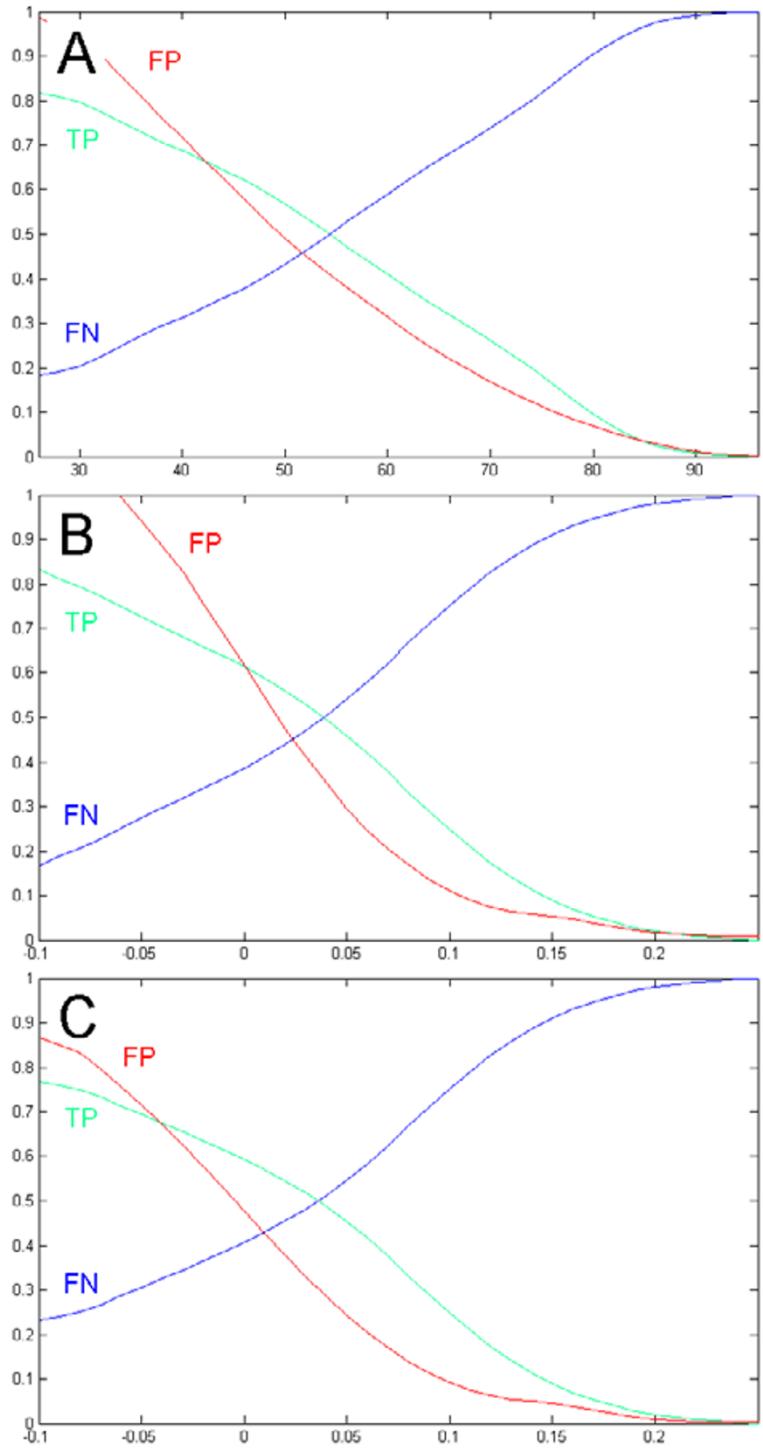


**Figure 6**
**Prediction accuracy on a wide genomic region.** Prediction accuracy (ROC) measured for the loci shown in Figure 5. Amount of true positives (TP, green line) is equal to the fraction of correctly predicted DNA positions in CRMs, calculated for different cutoff values. Blue and red lines show the rate of false-negatives (FN) and false-positives (FP) correspondingly. (*A*) Prediction accuracy based on sequence conservation (PIP) between the two *Drosophila* species, (*B*) prediction accuracy based on word frequency analysis, (*C*) an example of a combined approach; moderate filtering for non-conserved regions (below 30%) allows further improvement of the prediction quality as in (*B*).

boundaries are mapped precisely. In reality a vast fraction of the sequences comprising the considered loci (16–25 Kb) were never tested for CRM activity and the CRM boundaries represent an interpretation of empirical biological tests that are often difficult to formalize. At the same time word frequency algorithm produced very high positive scores for many of the best known 'classical' developmental CRMs, such as *eve* stripe 2, *eve* stripe3+7, *eve* stripe 4+6, *kr* CD1, *ftz* zebra and many others (see Figure 5). Correct recognition of these 'classical' elements, analyzed by many independent experimental groups, strongly supports relevance of our word frequency analysis.

The good agreement between the two independent techniques (word frequency and phylogenetic footprinting) and the successful recognition of the most known CRMs demonstrates the power of our strategy in extraction of CRMs from developmental genes (see Figure 2). It is important that using positive training set based on clusters of early developmental transcription factors we also extracted some of the *late* (expressed at later stages of fly development) CRM elements that were not among the training sequences. This suggests that our learning method accounts for *general* features (such as word distributions) inherent to regulatory DNA, rather than for particular motifs and words, specific to particular promoters/CRMs.

It is early to say what strategy will finally dominate in recognition of transcription regulatory regions. Apparently, at the current stage of this field, cross-validation using several independent techniques seems to be the most appropriate solution.

## Conclusions

### *Multiresolution analysis of transcription regulatory sequences*

To analyze the functional properties of transcription regions, we assign frequency value $F$ to each word of a sequence segment. This procedure results in a signal, where every position of DNA contributes one sample. At the following step we partition the frequency spectrum into frequency bands (frequency channels) containing words with similar frequency properties and analyze each frequency band (channel) independently. We take into account the full spectrum of word frequencies in our detection window; therefore for a chosen window size our assessment is exhaustive. Major advantage of this multiresolution analysis is its ability to produce highly informative pattern models, which minimize loss of information and facilitate recognition of very uncertain DNA patterns. In our case, for instance, a combination of non-overlapping frequency channels collects maximal information about the local word frequency in a sequence segment. All this information contributes to the final likelihood score. Direct extraction of words with defined frequencies (single frequency channels) may also have its specific field of application, different from promoter recognition. For example, one can filter out undesirable signals (noise) by combining or subtracting different frequency channels. This preprocessing may facilitate further extraction of biological signals such as binding motifs or nucleosome positioning signals using other techniques.

In this work we select word frequency $F$ as our main feature, however there is a good number of various transformation functions that were already implemented for the analysis of biological sequences [38-40]. Most of these functions are based on analysis of local base composition and local base frequency. It is interesting to emphasize here, that if we accept size of the word equal to 1, then the described word frequency analysis becomes 'local base frequency analysis'.

For many types of signals, such as speech and images, most of the information is localized in certain resolution (scaling) levels. In the case of biological sequences the resolution level or size of the resolution window can be selected equal to the size of known functional regions, such as *cis*-regulatory modules, analyzed in this work. Nevertheless, in many cases it is not known if the selected resolution window is optimal. In our recent work [9] we evaluated biological signal (clustered binding motifs) in a broad scaling range and developed a procedure that allows establishing the size of the optimal resolution window. That application also might be considered as a multiresolution analysis, but relatively to the scaling range (resolution window $l$ in our case).

Applications based on the multiresolution analysis are relatively new in the field of promoter study and they still require careful feature selection and thorough biological interpretation of obtained patterns. The last type of problems, however, can be solved in the context of well-known biological systems such as *Drosophila* developmental enhancers (CRMs), where accumulated biological information and annotated sequence data create very comfortable environment for such exploration.

## Methods

### *Noise suppression and error correction*

To construct our final likelihood score, we adopted two types of error correction. First correction accounts for chances that the observed in a single frequency channel value of $S$ possesses an error. One can see that according to formula (2), the log-likelihood score is equal to zero in the point where $E(\omega_1|S) = E(\omega_2|S)$. This point is marked by the vertical red line in Figure 4B. However, this value $S_0$

is much closer to mathematical expectation of $E$ for the negative training set, than for the positive one. One way to account for this error is to weight cumulative $E$-values of both training sets for $S < S_0$ and $S > S_0$:

$$L = log\left(\frac{E(\omega_1 \mid S)}{E(\omega_2 \mid S)}\right) + log\left(\frac{\left(\sum_0^s E(\omega_1 \mid S)\right)\left(1 - \sum_0^s E(\omega_1 \mid S)\right)}{\left(\sum_s^{+\infty} E(\omega_2 \mid S)\right)\left(1 - \sum_s^{+\infty} E(\omega_2 \mid S)\right)}\right) \quad (3)$$

This formula takes more simple form if the distributions are known [37], which is not exactly our case (see results).

Our second correction is intended to fix possible domination of one frequency channel over the others in some cases. Indeed, if the observed in a test sequence value of $S$ in any frequency channel is very far from the math expectation of a positive or a negative training set ($E(S) = 0$), than the log-likelihood score of this channel takes very high values, and it predominates over the score of other channels. As a result, the final decision is maid according to this channel only (sum of scores in all channels). To cope with this problem, we add a multiplier, suppressing possible high likelihood values of a $j$-th frequency channel:

$$\Lambda_j = L_j * \nu(L_j, M, \sigma) \quad (4)$$

This multiplier ν is a normal distribution of the likelihood score $L$ with expectation $M = 0$ and standard deviation σ. For very high positive or very high negative values of $L$, the value of the multiplier approaches zero, suppressing possible dramatic difference between channels. Behavior of the corrected likelihood score $\Lambda_j$ at the different values of parameter σ is given in Figure 4B.

## Authors' contributions

A.N. participated in annotation of *Drosophila cis*-regulatory modules and computational analysis. D.P. carried out programming and database construction. Both authors read and approved the final manuscript.

## Additional material

### Additional File 1

*CRM extraction using alternative training datasets. Comparison of LWF profiles obtained using different training sets of sequence data. CRM extraction using alternative training datasets. (*A-D*) Comparison of the two negative training sets, one (blue line) representing genome of* D. melanogaster*, another one representing non-coding samples only from genome of* D. melanogaster *(cyan line). Exclusion of coding sequences from the negative training set (cyan line) resulted in higher error rate in exons (compare values of the highest peaks on* D*). (*E-F*) Changes in the positive training set (jackknife test, red line) had little effect on the overall prediction accuracy (see also Table 1). One striking effect was detected in the* ftz *locus (*E*) in the region containing highly repetitive AT-rich tract in the position ~4000.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-65-S1.png]

## Acknowledgements

We thank Konstantin Severinov and Nikolaus Rajewsky for their critical remarks and help with the manuscript preparation. This work was supported by grants from National Institutes of Health GM51946 to Stephen Small and GM064864 to Claude Desplan. The CRM annotation and related resources are available from New York University Web site: http://homepages.nyu.edu/~dap5

## References

1. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition.** *Genome Res* 1997, **7:**861-78.
2. Ohler U, Niemann H: **Identification and analysis of eukaryotic promoters: recent computational approaches.** *Trends Genet* 2001, **17:**56-60.
3. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-81.
4. Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH: **Patchy Interspecific Sequence Similarities Efficiently Identify Positive cis-Regulatory Elements in the Sea Urchin.** *Dev Biol* 2002, **246:**148-61.
5. Hehl R, Wingender E: **Database-assisted promoter analysis.** *Trends Plant Sci* 2001, **6:**251-5.
6. Klingenhoff A, Frech K, Werner T: **Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach.** In *Silico Biol* 2002, **2:**S17-26.
7. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:**757-62.
8. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci U S A* 2002, **99:**763-8.
9. Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA: **Homotypic regulatory clusters in Drosophila.** *Genome Res* 2003, **13:**579-88.
10. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:**30.
11. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12:**739-48.
12. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9:**211-23.

13. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker – a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10:**577-86.
14. Elnitski L, Riemer C, Petrykowska H, Florea L, Schwartz S, Miller W, Hardison R: **PipTools: A Computational Toolkit to Annotate and Analyze Pairwise Comparisons of Genomic Sequences.** *Genomics* 2002, **80:**681-90.
15. Rajewsky N, Socci ND, Zapotocky M, Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons.** *Genome Res* 2002, **12:**298-308.
16. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I: **Strategies and tools for whole-genome alignments.** *Genome Res* 2003, **13:**73-80.
17. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GM, Celniker SE: **Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome.** *Genome Biol* 2002, **3:**RESEARCH0086.
18. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-4.
19. Ohler U, Harbeck S, Niemann H, Noth E, Reese MG: **Interpolated markov chains for eukaryotic promoter recognition.** *Bioinformatics* 1999, **15:**362-9.
20. Ohler U: **Promoter prediction on a genomic scale – the Adh experience.** *Genome Res* 2000, **10:**539-42.
21. Ohler U, Niemann H, Liao G, Rubin GM: **Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.** *Bioinformatics* 2001, **17:**S199-206.
22. Hutchinson GB: **The prediction of vertebrate promoter regions using differential hexamer frequency analysis.** *Comput Appl Biosci* 1996, **12:**391-8.
23. Lewis EB, Knafels JD, Mathog DR, Celniker SE: **Sequence analysis of the cis-regulatory regions of the bithorax complex of Drosophila.** *Proc Natl Acad Sci U S A* 1995, **92:**8403-7.
24. Scherf M, Klingenhoff A, Werner T: **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.** *J Mol Biol* 2000, **297:**599-606.
25. Levitsky VG, Katokhin AV: **Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis.** In *Silico Biol* 2003, **3:**8.
26. Ioshikhes I, Trifonov EN, Zhang MQ: **Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure.** *Proc Natl Acad Sci U S A* 1999, **96:**2891-5.
27. Levitsky VG, Podkolodnaya OA, Kolchanov NA, Podkolodny NL: **Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis.** *Bioinformatics* 2001, **17:**998-1010.
28. Perier RC, Junier T, Bonnard C, Bucher P: **The Eukaryotic Promoter Database (EPD): recent developments.** *Nucleic Acids Res* 1999, **27:**307-9.
29. Kassis JA: **Spatial and temporal control elements of the Drosophila engrailed gene.** *Genes Dev* 1990, **4:**433-43.
30. Nasiadka A, Krause HM: **Kinetic analysis of segmentation gene interactions in Drosophila embryos.** *Development* 1999, **126:**1515-26.
31. Holloway DM, Reinitz J, Spirov A, Vanario-Alonso CE: **Sharp borders from fuzzy gradients.** *Trends Genet* 2002, **18:**385-7.
32. Andrioli LP, Vasisht V, Theodosopoulou E, Oberstein A, Small S: **Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms.** *Development* 2002, **129:**4931-40.
33. Kolpakov FA, Ananko EA, Kolesov GB, Kolchanov NA: **GeneNet: a gene network database and its automated visualization.** *Bioinformatics* 1998, **14:**529-37.
34. Serov VN, Spirov AV, Samsonova MG: **Graphical interface to the genetic network database GeNet.** *Bioinformatics* 1998, **14:**546-7.
35. Rubin GM, Lewis EB: **A brief history of Drosophila's contributions to genome research.** *Science* 2000, **287:**2216-8.
36. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23:**4878-84.
37. Theodorides S, Koutroumbas K: *Pattern recognition London: Academic Press*; 1998.
38. Arneodo A, Bacry E, Graves PV, Muzy JF: **Characterizing long-range correlations in DNA sequences from wavelet analysis.** *Physical Review Letters* 1995, **74:**3293-3296.
39. Dodin G, Vandergheynst P, Levoir P, Cordier C, Marcourt L: **Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences.** *J Theor Biol* 2000, **206:**323-6.
40. Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y, Thermes C: **Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes.** *J Mol Biol* 2002, **316:**903-18.
41. Waterman MS: *Introduction to Computational Biology Chapmen & Hall*; 1995.