BMC Bioinformatics

BioMed Central

Research article

Open Access

Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs

Bartek Wilczynski, Norbert Dojer, Mateusz Patelak and Jerzy Tiuryn

Address: [1]Institute of Informatics, University of Warsaw, Warsaw, Poland

E-mail: Bartek Wilczynski - bartek@mimuw.edu.pl; Norbert Dojer - dojer@mimuw.edu.pl;
Mateusz Patelak - m.patelak@students.mimuw.edu.pl; Jerzy Tiuryn* - tiuryn@mimuw.edu.pl
*Corresponding author

Published: 10 March 2009

BMC Bioinformatics 2009, **10**:82 doi: 10.1186/1471-2105-10-82

This article is available from: http://www.biomedcentral.com/1471-2105/10/82

Received: 26 August 2008

Accepted: 10 March 2009

Abstract

Background: Finding functional regulatory elements in DNA sequences is a very important problem in computational biology and providing a reliable algorithm for this task would be a major step towards understanding regulatory mechanisms on genome-wide scale. Major obstacles in this respect are that the fact that the amount of non-coding DNA is vast, and that the methods for predicting functional transcription factor binding sites tend to produce results with a high percentage of false positives. This makes the problem of finding regions significantly enriched in binding sites difficult.

Results: We develop a novel method for predicting regulatory regions in DNA sequences, which is designed to exploit the evolutionary conservation of regulatory elements between species without assuming that the order of motifs is preserved across species. We have implemented our method and tested its predictive abilities on various datasets from different organisms.

Conclusion: We show that our approach enables us to find a majority of the known CRMs using only sequence information from different species together with currently publicly available motif data. Also, our method is robust enough to perform well in predicting CRMs, despite differences in tissue specificity and even across species, provided that the evolutionary distances between compared species do not change substantially. The complexity of the proposed algorithm is polynomial, and the observed running times show that it may be readily applied.

Background

Deciphering mechanisms of gene regulation is currently one of the key problems in molecular biology. The number of sequenced and annotated genomes is increasing rapidly, but we do not fully understand the regulatory networks underlying gene regulation. A few datasets approaching a genome-wide understanding of gene regulation in relatively simple organisms such as *E. coli* [1] or *S. cerevisiae* [2] exist, but especially for higher eukaryotes our understanding of gene regulation is far from complete. Experimental reconstruction of regulatory interactions is possible for relatively small systems [3], but it is impossible to scale this approach to all the available genomes. Therefore, computational methods are currently the best tool for improving our understanding of genome-wide gene regulation.

Biological background

The process of transcriptional regulation is facilitated by proteins called transcription factors which bind to DNA sequences to help or prevent the initiation of transcription by RNA polymerase. This binding is selective, i.e.

trans-factors bind only to specific DNA sequence motifs (called cis-elements) [4]. In higher eukaryotes, many genes need to exhibit complex spatio-temporal expression patterns. The key to achieving such complexity is the combinatorial transcription regulation [5], i.e. different combinations of similar *cis*-elements may yield different expression profiles. Sequence elements, whose main function is driving complex expression patterns, are often referred to as *cis-regulatory modules* (CRMs). Throughout this paper, we will use this term, but it should be noted that our method is limited to finding CRMs that are relatively close the transcription start site (TSS) of a gene of interest (in the range of 10 kb up- or down-stream of its TSS) whereas in general the term "CRM" may also be used to refer to distant enhancers which cannot be found using our method.

Previous work

The earliest computational approaches to discovering CRMs in non-coding DNA were based on two observations:

- CRMs contain unusually high concentration of binding sites [6],
- CRMs are more conserved across species than other non-coding sequences [7].

These early approaches sparked a number of studies which utilize different computational approaches to find CRMs based on these two presumed properties [8-19]. However, in the light of more recent analyses of the statistical properties of CRMs [20], neither assumption appears to be a reliable foundation for CRM prediction. After analyzing over 500 experimentally verified CRMs from *D. melanogaster*, Li et al. claim that the clustering of motifs may reliably predict only a few CRMs (most notably the ones involved in the early blastoderm formation). Similarly, evolutionary conservation of CRMs appears to be less stringent and much more nuanced than previously thought. Firstly, CRMs are significantly more conserved than the rest of non-coding DNA only if measured by the density of short (7 bp) blocks conserved between species, rather than by simple sequence identity over larger windows. This is supported by recent findings that the evolution of CRMs is driven by gain and loss of whole binding sites rather than point mutations [21]. Secondly, even though the set of investigated CRMs was statistically conserved, the authors conclude that most CRMs are not distinguishable from other non-coding sequences based solely on conservation. These findings are not specific to *D. melanogaster* and are supported by a very recent study [22] based on comparing TF binding signatures in human and mouse liver.

However, there are two published studies addressing these issues at least partially. Hallikas et al. [23] propose the EEL algorithm for finding alignments of significant motif occurrences instead of the sequences themselves. This method is very efficient and does not rely on raw sequence similarity but it assumes that the motifs in conserved CRMs occur exactly in the same order. On the other hand, the BLISS method [24] approaches the same problem by analysis of a matrix containing occurrences of all motifs along both homologous sequences after Gaussian smoothing. This relaxes the assumption of conserved motif occurrence order but at the very high cost of computations. These two approaches fall into the category of *non-tissue-specific* methods. The approach reported in the present paper also falls into this category. The other group of methods, which could be called *tissue-specific*, are tuned for a particular type of CRMs, using either a set of several known specific motifs [9], or by learning such motifs from the known tissue-specific CRMs [8].

Contributions of the present paper

We present a novel approach to finding CRMs in non-coding sequences associated with homologous genes. It is based on a simple method of scoring likelihood of the occurrence of a conserved combination of binding sites in a fixed-size window. This measure is constructed in such a way that it does not rely on strict criteria for neither sequence conservation, nor for motif clustering. We show that we are able to use the same parameters to discover motifs in human, rat, mouse and fruit fly using a universal, non-tissue-specific set of known motifs.

The overall procedure of the proposed method may be divided into the following steps:

- Finding occurrences of transcription factor binding site motifs obtained from a database in a set of DNA sequences proximal to transcription start sites of genes of interest.
- Calculating the alignment scores of motif sets in windows of fixed size using a novel method for homologous sequences from related species
- Measuring the rarity score of best alignments against a randomized set of promoters from the same species to filter out non-specific alignments.

The output of this procedure is a ranking of alignments of sequences along with the motifs contributing to this alignment. In the following section, we describe each of these steps in detail.

Methods

Identifying motif occurrences

In the present study we use models of transcription factor binding sites from the JASPAR CORE database

[25]. It consists of 138 non-redundant motifs represented by frequency matrices. Each matrix has to be assigned a threshold of the log-likelihood of binding site affinity.

As explained in the following sections, our method of CRM identification takes into account both positive and negative signals from the promoter sequence. Thus, the control of both error types in the motif prediction has to be balanced, in the sense that the number of false positives should be of the same order of magnitude as the number of false negatives. As proposed by Rahmann et al. [26] we choose thresholds maintaining the balance between the control of type I and type II errors (see supporting materials for details).

In order to identify occurrences of a motif, sequences are scanned for words with log-likelihood above the related threshold.

Comparing window contents

As discussed earlier, the presented approach is based on a slight relaxation of the motif order in the promoter region. We incorporate this idea by utilizing the concept of a fixed-size window with the assumption that the order of motifs which occur within the window does not matter, i.e. we treat the motif occurrences in one window as forming a multiset rather than a sequence. For this, we introduce a parameter W denoting the window length. We say that a motif occurs in a window if its left border occurs in that window. In order to compare contents X, Y of two windows we reward the common motif occurrences and penalize motifs which occur in one window but not in the other, as well as windows with empty motif sets. This leads to the following cost function:

$$c(X, Y) = \alpha \cdot |X \cap Y| - \beta \cdot |X \div Y| - \gamma,$$

where $|X \cap Y|$ denotes the number of motif occurrences which are in common (intersection), while $|X \div Y|$ is the number of motif occurrences in one window but not in the other (symmetric difference). If a motif M occurs M_X times in X and M_Y times in Y, then its contribution to the term $|X \cap Y|$ is $\min(M_X, M_Y)$, while its contribution to the term $|X \div Y|$ is $|M_X - M_Y|$. The constants α, β, γ are parameters of the cost function. We assume that $\alpha > 0$ to give preference to windows containing common motif occurrences. Since multiplying the cost function by a positive constant does not change the relative assessment of the window content, it follows that we may assume, without loss of generality, that $\alpha = 1$. Then, the role of $\beta > 0$ is to penalize for motifs occurring in one of the sequences but not in the other. It follows from our experiments that in case of a general motif database β should be much smaller than 1. The role of $\gamma > 0$ is to penalize pairs of windows with empty content which cannot be affected by changes of α and β. In our experiments we have verified that γ should be close to 0 and, in general, increasing it overly decreases the sensitivity of the method. For details of β and γ estimation see the section on Parameter estimation – case study of muscle specific CRMs.

Given two promoter sequences S_1 and S_2, we are looking for window stretches of the same length: one in S_1 and one in S_2, so that the cumulative cost for consecutive pairs of windows yields an "unusually high" score. Computing the score is illustrated in Fig. 1(a) where the total cost for two pairs of windows is $6 - 6\beta - 2\gamma$. However, it may happen, as shown in Fig. 1(a), that the

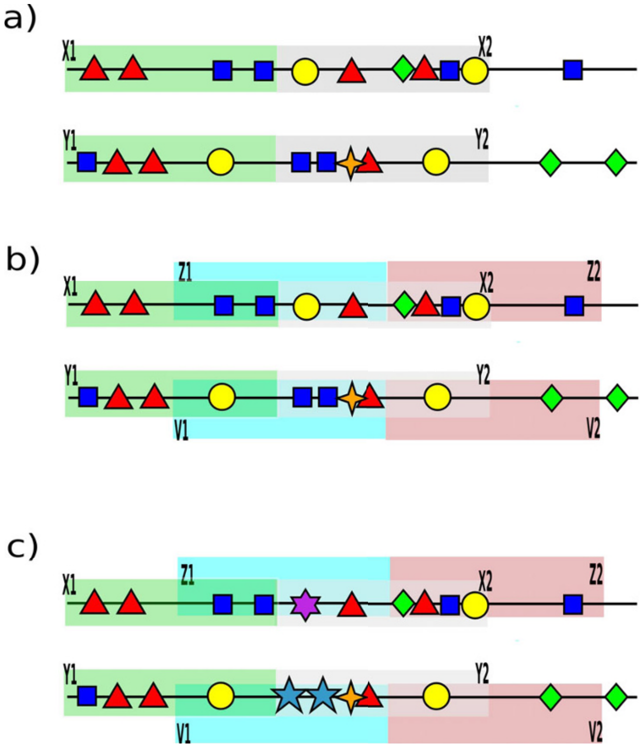

Figure 1

Comparing window contents. Part (a) shows two pairs of non-overlapping windows. Here we have $c(X1, Y1) = 3 - 2\beta - \gamma$ and $c(X2, Y2) = 3 - 4\beta - \gamma$. Observe that motif occurrences near the border between the windows show high similarity. In part (b) we show two extra pairs of windows which partly overlap the windows from (a). The cost of extra windows is $c(Z1, V1) = 4 - \beta - \gamma$ and $c(Z2, V2) = 2 - 3\beta - \gamma$. In part (c) there is presented a similar situation, except the motifs near the border have changed. We obtain the cost for the new contents: $c(Z1', V1') = 1 - 7\beta - \gamma$ and $c(X2', Y2') = 3 - 6\beta - \gamma$. The other contents in (c) remain the same as in (b). Hence, the total cost of alignment in (c) is smaller by $3 + 8\beta$.

border between two windows may cut through an area of the promoter with high similarity of occurring patterns of motifs. In this case we have in Fig. 1(a) two occurrences of one motif (square) and one occurrence of another motif (circle). In order to accommodate for this situation we allow windows to overlap, i.e. we introduce a step J and assess fragments of promoter sequences as seen from the contents of a window which moves every J nucleotides. This is illustrated in Fig. 1(b) where J is taken as a half of W. In this case, the total score of the fragment becomes $12 - 10\beta - 4\gamma$. On the other hand, when the border area between two windows does not show high similarity, the additional (shifted) windows contribute less to the total score as shown in Fig. 1(c). In this case the total score is $9- 18\beta - 4\gamma$, which is smaller than the former score by $3 + 8\beta$.

In order to reduce computation time we consider only window positions starting from the left end of a promoter sequence and jumping every J nucleotides. For a promoter sequence of length L this yields about L/J positions to investigate, instead of L positions if we considered all possible window positions. Also the smaller J is, the more window positions we have to consider. As changing the length of J produces only quantitatively different, but qualitatively results (see Additional file 1) we set the J parameter in our method to be half of the window length.

Finding optimal window alignments

The next step is to find contiguous stretches of window pairs with maximal accumulated cost. Let us recall that we are considering two sequences S_1 and S_2 which are promoter regions for genes g_1 (in species A_1) and g_2 (in species A_2). We assume that g_1 and g_2 are homologs. Let the window length W and step J be fixed. Starting from the first position in each of the regions we move the window every J nucleotides until we reach the end of the region. Let N_1 (resp. N_2) be the number of such window positions in S_1 (resp. in S_2). For each $1 \le i \le N_1$, let X_i denote the contents of the window which starts in S_1 at window position i. Similarly, let Y_i be the contents obtained for S_2 (for $1 \le i \le N_2$). We build a $N_1 \times N_2$ matrix V which stores in entry (i, j) a maximal value of the sum along the diagonal of the cost matrix c, which ends in position (i, j). More formally $V(i, j)$ is computed as follows

$$V(i,j) = \begin{cases} \max[0, c(X_i, Y_j) + V(i-1, j-1)] & \text{if } i, j > 1 \\ \max[0, c(X_i, Y_j)] & \text{otherwise} \end{cases}$$

Having computed the matrix V, we are ready to do a pre-selection of CRMs. For each $i \le N_1$ let $j \le N_2$ be the index such that $V(i, j)$ is maximal among all the values in the i-th row of V. Let (i, j), $(i - 1, j - 1)$,...,$(i - k, j - k)$ be the maximal contiguous fragment of a diagonal of V which consists of a strictly decreasing sequence of values in V. A CRM pre-selected in S_1 is the region from window position $i - k$ through i (recall that positions in V correspond to window positions in the promoter sequences). At this stage we declare the CRM preselected since we have to further assess the likelihood of its occuring in the genome by chance.

Assessing the rarity of CRMs

Even though the most natural guideline for selection of CRMs is the accumulated cost of the pre-selected CRMs, we have to adjust this statistic in order to avoid a bias due to possible species-specific abundance of "random" occurrences of certain motifs. In the present paper we propose a simple heuristic approach to this issue, considering computational feasibility of the whole method. It is based on the idea that CRM in principle should be specific to some group of genes, but not too abundant in the genome. If not accounted for that, the results tend to contain mostly results from the core promoter area and from repetitive elements. Since it is not desirable to filter out these sequences (particularly filtering out core promoters could lead to serious problems), we set out to propose a quantitative measure of alignment *rarity*, which is supposed to promote specific alignments. The approach we take stems from simulation methods for obtaining random promoters, but since we need to retain the features of relatively long stretches of sequences, instead of randomly shuffling the sequences, we select a sample of random promoters from the considered genome and calculate how often the aligned region gets an alignment score at least as high as when it is aligned with its homologue. If this occurs frequently we consider such an alignment not interesting as a putative CRM.

We describe this procedure of randomization little more precisely. First of all our approach assumes that the user knows which area of promoter sequence is of interest to her/him. For instance, in our experiments with the muscle and liver genes we consider the promoter area to be -10 kb through +5 kb of the start of transcription, but other values can be considered as well. In order to assess the rarity in a given experiment we sample a set of 99 genes from the genome of the other organism and take their promoter regions which are of the same size and similar position with respect to the start of transcription as assumed by the approach. Then, for each gene g whose predicted CRM rarity we want to assess, we compute the score of g against each of the above mentioned 99 promoters. For a given position in the matrix, instead of the raw score $V(i, j)$, we consider the position of this score in the ranking of 100 sequences (and hundred matrices). This gives us an estimate of the relevance of this prediction. Detailed procedure is described in the following paragraph.

Let us keep the notation of the previous section (i.e. the matrix V, the genes, g_1 and g_2, the sequences S_1, S_2). We first explain how we compute the S_1-*view rarity of* (i, j), for a given position (i, j) in V. We retrieve 99 randomly selected sequences $R_1,...,R_{99}$ which are promoter regions of genes in species A_2 such that none of the genes is homologous to g_2, and each sequence R_m is of the same length as S_2 (i.e. of length N_2). Next, for each $m = 1,...,99$ we build the accumulative cost matrix V_m by computing window contents of S_1 against those from R_m. Matrix V_m is computed in the same way for S_1 and R_m as V was computed for S_1 and S_2. For $m = 1,...,99$, let $v_m(i)$ be the maximal element in the i-th row of V_m. Let p be the position of $V(i, j)$ in the set $\{v_1(i),...,v_{99}(i), V(i, j)\}$ counted in ascending order. *Ex aequo* occurrences of $V(i, j)$ with the other values are resolved to the benefit of these other values, i.e. $V(i, j)$ always occupies the last position in the block of equal values. We set S_1-view rarity of (i, j) as the ratio $p/100$.

Assume we want to assess rarity of a preselected CRM in promoter sequence S_1 and assume that this CRM was obtained from the fragment of V for positions (i, j), $(i - 1, j - 1)$,...,$(i - k, j - k)$. The rarity of this CRM in S_1 we define as the minimum S_1-view rarity of these positions.

Evaluating the quality of CRM predictions

In order to assess the performance of any CRM prediction method one has to choose a proper objective function. In this paper, we test our method on the data with known CRM data. It can be viewed as a classical prediction problem and scored with measures of sensitivity and specificity [9]. However, two facts should be accounted for in the case of CRM prediction:

(i) the expected number of negative examples is by far greater than the expected number of positive ones. For example, in our experiments for each promoter sequence in the muscle set, there are 300 windows of size 100, out of which only a few comprise CRMs. Similarly, Philippakis and Bulyk [9] consider 1000 negative examples for a dataset containing 27 CRMs (see Section on Comparison with other methods).

(ii) It is expected that there are more CRMs than the ones collected in the training dataset. For this reason, some "false positive" predictions might be actually true CRMs.

To account for the first problem, Chan and Kibler [27] propose one more measure, *positive predictive value* (PPV), which is the ratio of true positives among all predictions. They also note that, because of (ii) the values obtained for PPV are underestimates.

We believe that because of (i) any method based solely on the notions of true/false positives/negatives leads to misjudging the performance of CRM prediction. Specifically, if the number of negative examples is high compared to the positive ones, a method returning many predictions is scored better than the one giving fewer results. We consider such methods impractical. Therefore, we use a very simple but useful measure of prediction quality based on the position of the correctly predicted CRM in the ranking of predictions returned by our program. Given a threshold k, we say that the CRM is *found*, if the correct prediction is among the top k predictions (we call a prediction correct if it overlaps with the CRM and is not longer than 3 times the length of the true CRM). As the *mean quality* of a prediction method for a set of sequences annotated with known CRMs, we consider the ratio of found CRMs to the all known CRMs. The choice of k is arbitrary, but in real applications it should not be smaller than 5 because of (ii) and at the same time it should not be limited by the length of the sequence divided by the expected length of a CRM. We use the value $k = 5$ for mammalian datasets and $k = 10$ for fruit fly dataset (since there are 5 annotated CRMs in one promoter sequence, selecting only 5 top ranked alignments would be too restrictive). Another advantage of our scoring procedure is the fact that, as opposed to the PPV score, it is not sensitive to overlapping predictions. More precisely, generating many overlapping results for the high confidence regions may increase the PPV score of a method, it cannot increase the number of found CRMs. Even though we use our quality measure to optimize the parameters, we report in Section on Comparison with other methods the values of Sensitivity, Specificity and PPV. We also report the actual values of the overlap between our predictions and true CRMs (normalized by the sum of the lengths).

Experimental results

For all experiments with biological data discussed here we chose the window length $W = 100$ and step $J = 50$. Other values of W in the interval 50 through 250, as well as other steps, gave similar results (data not shown).

Organization of this section is as follows: we optimize parameters for our method on a set of muscle specific CRMs which have been experimentally verified. Then we show that our method gives reasonable results for other CRMs with these parameters: liver specific CRMs in human and the CRMs for the *eve* gene in *D. melanogaster*.

Parameter estimation – case study of muscle specific CRMs

We estimated the appropriate parameters on a large set of muscle specific CRMs reported by Wasserman and

Fickett [28]. It consists of 43 CRMs, mainly from human, mouse and rat. After removing from the list of CRMs those coming from other species (chicken, hamster, pig, cow) the remaining CRMs, 37 in total, were manually checked and verified at Ensembl database [29]. This step was essential since the original data was outdated with respect to the current genomes deposited in the database. In some cases we were not able to map a CRM to the corresponding gene. After omitting these doubtful cases we were left with 24 CRMs corresponding to 23 genes (5 genes are from human, 12 from mouse and 6 from rat). One human gene (DESMIN) had two CRMs. For all aforementioned genes and their homologs (in the other two species) we retrieved promoter regions flanking from -10 Kb through +5 Kb relative to the TSS according to Ensembl. We thus have created 48 pairs of promoter regions corresponding to homologous genes. The choice of the sequence length here was a trade-off between covering as many CRMs from [29] and running time of the learning procedure.

Estimation of parameters was performed on a grid of values for β and γ. We first examined the intervals 0 through 2 for both parameters with step 0.2 (in fact, we replaced 0 with $1 \cdot 10^{-5}$ due to the considerations in section on comparing window contents). For each set of parameters we computed the mean quality of CRM prediction (see subsection on evaluating the quality of CRM predictions). After the area with an optimal score was localized, we performed again the estimation on intervals 0 through 0.5 with step 0.05 for both parameters (as above replacing 0 with $1 \cdot 10^{-5}$). A plot of the obtained prediction evaluations is shown in Fig. 2. The best results were obtained for $\beta = 0.2$ and $\gamma = 1 \cdot 10^{-5}$. A set of more detailed numerical results is given in the supporting material.

An important question is whether the introduction of the rarity score improves the performance of the method. In order to verify that, we have repeated the parameter fitting procedure using only the raw alignment scores without computing the rarity. The optimal parameters were in fact slightly different ($\beta = 0.4$, $\gamma = 0.05$), but the overall prediction quality dropped dramatically. We were able to predict only 1/3 of examples and it is worth noting that the rankings of the predictions are in 40 cases out of 50 cases lower. It is important that even though in some cases (6 out of 50) the raw score gives a better ranking, these are only in cases where the rarity score gives a prediction within the top 5 as well. However, in 23 cases the correct prediction is ranked among the top 5 by the rarity score, whereas the raw score is below that cutoff. The comparison of the two rankings is presented in Fig. 3. All results of our method on the training dataset are presented in Additional file 1.

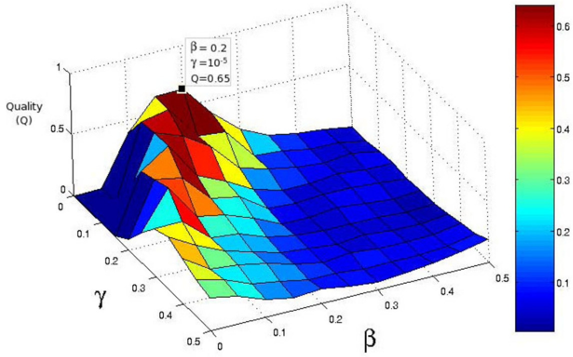

Figure 2
Prediction quality as a function of β and γ parameters. The prediction quality (Q) for the training (muscle) dataset is plotted here as a function of the β and γ parameters. The maximum value is marked by a square. It should be noted, that the prediction quality seems to be close to zero for most values except for the small area around maximum. Our experiments with wider ranges of parameters (data not shown) also support that hypothesis.

Prediction of liver specific CRMs in human

The experiment was performed on the set of liver specific CRMs reported by Krivan and Wasserman [30]. The dataset consists of 16 CRMs: 10 from human and 6 from other species. We have manually selected 7 human CRMs which, according to Ensembl, lie in the regions from -10 Kb through +5 Kb relative to the TSSs. Then, we retrieved the flanking sequences for the selected genes and their homologs in rat and mouse. Both these species have two homologs of human insulin and one homolog for each of the other genes which gives 16 gene pairs altogether.

The algorithm was run on all 16 pairs of homologous promoter regions with the parameters chosen for the muscle specific data in the previous section ($\beta = 0.2$ and $\gamma = 1 \cdot 10^{-5}$). The result of CRM prediction is reported in Table 1 (a more detailed version is included in Additional file 1). Out of 7 CRMs, 4 are clearly predicted (rarity ≤ 0.1, ranking ≤ 5) and the remaining 3 are not found (rarity ≥ 0.8). This result is comparable with the performance of the method on the training dataset. Observe that the well predicted CRMs have a significant overlap.

Predicting even-skipped CRMs in fruit fly

In this experiment we have concentrated on *cis*-regulatory modules for the well studied gene *eve* in *Drosophila melanogaster*. We extracted the relative positions of the following five experimentally verified eve CRMs from REDfly database [31]: eve_stripe1, eve_stripe2,

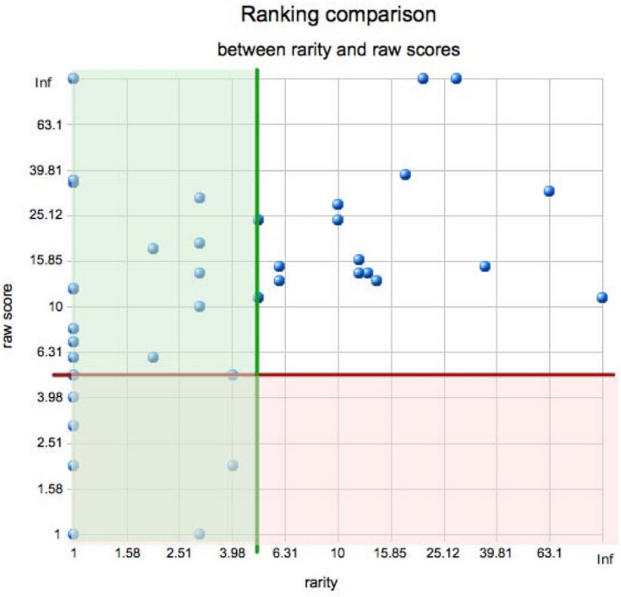

Figure 3
Comparison of the raw score and the rarity score rankings. The plot shows the comparison of rankings obtained for the CRMs from the training set using raw score vs. rarity scores. Each point corresponds to a single CRM. The position of the dot depends on the ranking of this CRM using raw scores (X-axis, log scale) and its ranking according to the rarity score (Y-axis, log scale). The green and red lines are placed at $k = 5$ for rarity and raw scores respectively. The points placed below (red shaded area) and left (green shaded area) of these lines are considered to be found by respective methods. As we can see, there are no points found only when using the raw score but substantial number of them is found only when using the rarity score. It should be noted, that the parameter estimation was done for both rankings separately, i.e. optimal parameters β and γ were used for both methods. The data for this table is available in Additional file 1.

Table 1: Prediction quality of liver specific CRMs in human

human gene	homolog species	prediction
ALDOB	mouse	incorrect
	rat	incorrect
IGF1	mouse	correct
	rat	correct
PAH	mouse	incorrect
	rat	incorrect
PROC	mouse	correct
	rat	correct
CYP7A1	mouse	incorrect
	rat	incorrect
G6PC	mouse	correct
	rat	correct
INS	mouse	incorrect
	mouse	correct
	rat	correct
	rat	correct

For each gene name in column 1, one row shows CRM prediction for one homolog (described in column 2).
If the most significant prediction which overlaps the experimentally verified CRM is in the top 5 in the rarity ranking, we call it correct, and incorrect otherwise. An extended version of this table can be found in Additional file 1.

eve_stripe4_6, eve_stripe5, and eve_stripe_3+7. Their length ranges from 500 to 800 nucleotides. The promoter regions corresponding to the eve gene in (several species) were downloaded from Flybase [32]. The reason for taking the region (-5 Kb, +10 Kb), rather than (-10 Kb, +5 Kb) was in order to make sure that the five CRMs of interest are included in the selected area while keeping the same overall sequence length. These four Drosophila species have diverged from *D. melanogaster* so that *D. erecta* and *D. melanogaster* are the closest relatives, while *D. mojavensis* and *D. melanogaster* are evolutionarily furthest apart among these four. Fig. 4(e) contains a phylogenetic tree for these five flies together with other seven *Drosophila* species.

We ran our method on four pairs of promoter regions with the same parameters as for other datasets ($\beta = 0.2$ and $\gamma = 1 \cdot 10^{-5}$). The quality of prediction of the experimental CRMs as retrieved by our method are shown in Table 2. With the exception of stripe_5, all other CRMs were predicted correctly for the pair *D. melanogaster*/*D. pseudoobscura*. The results for *D. melanogaster*/*D. ananassae* pair were marginally worse than for the previous pair. The results were not satisfactory for the closer relative *D. erecta* nor for the farthest relative *D. mojavensis*. This could be explained by the fact that evolutionary distance between *D.mel.* and *D.pse.*/*D.ana.* is similar to the distance between human and mouse/rat. This suggests that a reference species should be selected so that the evolutionary distance is similar to that in the training dataset. Though we did not investigate this issue because of the lack of proper training dataset, it is possible that we could get better predictions of even skip CRMs for a different set of parameters for different evolutionary distances. Nevertheless, it is remarkable that despite applying parameters that were estimated for different species and for genes with different tissue specificity we obtained predictions of similar quality.

Fig. 4 presents the top ten predictions for four pairs of flies obtained with our method. They are presented by brown stripes (orange stripes are included in this figure

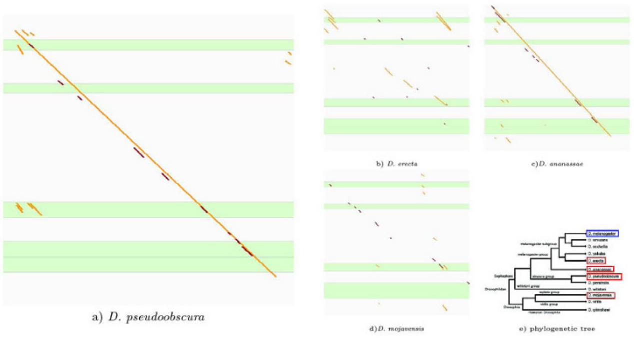

Figure 4
Comparison of CRM predictions in even-skipped gene in different Fruitflies. In parts (a) through (d) the Y-axis represents positions in the promoter region of *eve* gene of *D. melanogaster* with 5' end at the top and 3' end at the bottom, while the X-axis represents positions in the promoter region (with 5' end corresponding to the left and 3' end corresponding to the right end) of *erecta, ananassae* and *mojavensis*, respectively. Brown strips represent the top ten predictions by our method, while orange strips represent the top ten predictions by EEL. Light green horizontal areas represent positions of the experimentally verified even skip CRMs in *D. melanogaster*. Part (e) presents a phylogenetic tree of 12 *Drosophila* species including the ones discussed here.

Table 2: CRMs in the gene eve in fruit fly

CRM homolog	*Drosophila erecta*	*Drosophila ananassae*	*Drosophila pseudoobscura*	*Drosophila mojavensis*
stripe3+7	-	-	+	+
stripe2	-	-	+	+
stripe4 6	-	-	+	+
stripe1	-	+	+	+
stripe5	-	-	-	-

The table reports quality of the most significant prediction of each CRM in the gene eve in *Drosophila melanogaster* obtained by our method with each of other considered *Drosophila* species. The key to values is the same as in Table 1, however since we are looking for 5 CRMs, we have adjusted the value k to 10.

for comparison with another method, see the next section). The reader may notice that in addition to the sought after experimental CRMs there are a number of new putative CRMs which may be false positives, but may turn out to be true *cis*-regulatory modules. Some of the putative CRMs are supported by more than one pair which may be an indication of a true CRM.

The set of motifs used in this study was a broad spectrum of motifs from JASPAR CORE. We have also investigated the impact of choosing more specific motifs for our approach. We have run the experiment for a set of eight specific motifs which constitute the even skip CRMs. However, the results were rather discouraging (data available in Additional file 1), suggesting that the general purpose motifs such as JASPAR CORE are better suited for discovering evolutionarily conserved CRMs. One possible explanation for this observation is that such species-specific motifs are ubiquitously conserved in the fly genome and therefore true CRMs do not stand out in comparison with the background model. We therefore propose that JASPAR CORE (or a similar set of non-species-specific motifs) may be a better choice for predicting conserved CRMs.

Comparison with other methods

Recall that methods available for the task of CRM prediction can be divided into two distinct classes based on the chosen approach:

- tissue-specific methods, tuned for a particular type of CRMs, using either a set of several known specific motifs [9], or learning such motifs from the known tissue-specific CRMs [8],
- general methods based on a universal motif set (e.g. [23, 24]).

The method proposed in the present paper belongs to the second class and our comparison is carried out within this class. We refer the reader to [8] for a thorough comparison of the performance of tissue-specific approaches.

There are two published methods proposing computational prediction of CRMs based on a non-specific motif set. Though we could not compare our results with the BLISS algorithm [24], as it seems to be limited to sequences significantly smaller than 15000 bp. We believe that this is due to the fact, that the computation cost was too high. Actually, the time complexity of BLISS algorithm is $O(L^2m)$, where L is the promoter length and m is the number of considered motifs. On the other hand the complexity of our method is $O(L^2)$ (see Additional file 1). Even though the asymptotic complexity of these methods with respect to L is the same (if we consider m as a constant), the number of atomic operations which have to be performed by our method for the actual values of L and m is approximately 250 thousand times smaller than for the BLISS method. This is in part due to using 'step' by our method. It should be noted that the running time of our method is less than a minute for a pair of sequences of length 15*kb* on a standard workstation PC.

Computation time is also a strong point of EEL software [23], which is available for download. We have rerun the EEL software on the same datasets as our method. For the muscle dataset, it was able to recover altogether 14 of

Table 3: Prediction quality for fruit fly CRMs. The table reports prediction quality for CRMs in *Drosophila melanogaster* obtained with *Drosophila pseudoobscura*.

	our method			EEL		
Top rank	found	SN	PPV	found	SN	PPV
5	2	0.4	0.4	1	0.2	0.2
10	4	0.8	0.4	2	0.4	0.2

the 24 CRMs (detailed results are in Additional file 1). For the liver dataset, the results of the two methods were comparable: EEL was able to recover 5 out of 7 CRMs, while we were able to recover 4 from the same data.

The results for the fruit fly dataset are presented in Table 3. We present here the performance of both methods in recovering the even-skipped gene CRMs using the homology with *D. pseudoobscura*. The values of sensitivity and PPV are presented in two variants: assuming top 5 predictions or top 10. Another view of the same results is presented in Fig. 4. We present there top 10 predictions for both methods for all 4 pairs of homologous sequences. In three cases (a, c, d) our method provides reasonable predictions while the EEL method is often providing wrong predictions. It should be noted that a possible reason for poorer performance of EEL on the Fruifly dataset is that the set of parameters used by EEL for mammalian genomes cannot be used to the analyses on insect genomes, even though the authors of EEL [23] show that it is able to recover the same enhancers of *eve* with a specific set of motifs (and possibly different parameters, however this is not clear from their study). In contrast, the parameters for our method seem to be applicable to Fruifly data, especially when comparing *D. melanogaster* to *D. pseudoobscura*.

Discussion and conclusion

The novel method of predicting *cis*-regulatory modules which is proposed in the present paper is based on the following two salient features:

- implementation of a mixture of sequential and set-theoretic evaluation of similarity measure for groups of motif occurrences;
- introduction of a rarity measure for putative CRMs.

Both of these above features play an important role in the quality of CRM prediction. Introduction of a rarity measure for putative CRMs plays a crucial role in moving true positives up in the ranking. We propose a straightforward method of computing this rarity measure, having mainly computational efficiency in mind. Further research should clarify whether we can improve with respect to the quality of predictions when adopting less naive ways of computing rarity without affecting computational time. The main contribution and power of the present method lays in combining both: sequential and set-theoretic aspects of assessing similarity of motif clusters. As mentioned earlier one way of approaching the problem of predicting CRMs is via discovering conserved non-coding sequences and then finding their subsequences that contain a large number of motif occurrences. This is partly covered by our method, since if two promoter fragments have a similar sequence and are drawn from the same background, then the penalty for symmetric difference of sets of motif occurrences in these fragments is zero, and what really counts is the number of such occurrences. On the other hand, in our method we relax the assumption of sequence conservation since we are working solely with motif occurrences. We also do not assume a strict conservation of the order of motif occurrences by allowing disruptions of this order without any penalty, provided that the occurrences are not too far apart, i.e. they fit into one window. For the same reason we do not penalize differences in relative distances of motifs, providing the differences are within the window length.

Also, even though in our analyses all known CRMs considered are of comparable size (mostly 100–500 bp), the range of possible sizes of CRMs is still debated. Judging by the sizes in databases of CRMs, the regions are often much larger (even up to 5 kb, see [31]), while one of the most well studied enhanceosomes (Iterferon-β) is only 60 bp in length [33]. Given this wide range of possible CRM lengths it may be considered an advantage that our method puts less constraints on the CRM length as other, most notably conservation-based, methods.

An appealing feature of our method is that it is tailored for use with a non-specific set of motifs. The results of experiments presented here show clearly that a non-specific set such as JASPAR CORE works very well for muscle and liver, as well as for the even skip CRMs.

Another important issue which emerges from the results of our method is that the quality of CRM prediction may largely depend on the evolutionary distance of a relative organism against which we compare constellations of CRMs. As we have seen, when the species of interest is *D. melanogaster*, we obtained unsatisfying results when the chosen relative was *D. erecta*. The outcome was better for *D. ananassae*, best for *D. pseudoobscura*, but again worse for *D. mojavensis* (but not as bad as for *D. erecta*).

It should be noted that *D. pseudoobscura* and *D. ananassae* are in a similar evolutionary distance to *D. melanogaster* as rat and mouse are to human [34] It also seems that the set of relatives mouse/rat works well for human.

Parameters

In total our approach uses the following four internal parameters. Remarks on selection of these parameters are given in section on experimental results.

- W – length of the window (set to 100);
- J – length of step (set to 50);
- coefficients in the cost function (as discussed in the text the parameter α can be set to 1 without loss of generality):

(i) β – penalty for difference in motif composition (set to 0.2);
(ii) γ – CRM extension penalty (set to 10^{-5}).

In order to filter the obtained results the user may choose the following two parameters:

- The threshold rarity (we use 0.05);
- the number k of top predictions to be displayed (we used $k = 5$ for liver and muscle data, and $k = 10$ for the gene even-skipped).

It should be noted that the total length of promoter sequences, as well as the position of left and right flank with respect to the start of transcription can be also considered a parameter, since the user may choose to search for CRMs using flanking regions of different size and different relative position.

Availability and requirements

Project name: Billboard

Project home page: http://bioputer.mimuw.edu.pl/papers/crm08

Operating system(s): Platform independent (tested only on Linux)

Programming language: Java (requires JDK6 and the Ant tool)

License: GNU GPL

Any restrictions to use by non-academics: None

Abbreviations

CRM: cis-regulatory module; PPV: positive prediction value.

Authors' contributions

BW, JT and ND conceptualized the study and wrote the paper; All authors read and approved the manuscript; JT proposed the approach based on windows and the score function; BW implemented a prototype, proposed the rarity measure and performed the experiments; ND devised the dynamic algorithm for adaptive motif thresholds; MP implemented the final software.

Additional material

Additional file 1

Supplementary Materials. *Supplementary Materials include description of some technical details (computing thresholds of motif occurences, impact of window size and step on training quality) some result details (tables, figures, ROC curves) and a note on computational complexity.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-82-S1.pdf]

Acknowledgements

We are very grateful to Robert Zinzen for his comments on the paper. This work was supported by grants of Polish Ministry of Science and Education: No. 3 T11F 021 28 and PBZ-MNiI-2/1/2005. BW received support from Foundation for Polish Science.

References

1. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza Spinola MIl, Contreras-Moreira B, Segura-Salazar J, Muñiz Rascado L, Martínez-Flores I, Salgado H, Bonavides-Martínez C, Abreu-Goodger C, Rodríguez-Penagos C, Miranda-Ríos J, Morett E, Merino E, Huerta AMM, Treviño Quintanilla L and Collado-Vides J: **RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation.** *Nucleic Acids Res* 2008, **36:**D120–4.
2. Macisaac KD, Wang T, Gordon BD, Gifford DK, Stormo GD and Fraenkel E: **An improved map of conserved regulatory sites for Saccharomyces cerevisiae.** *BMC Bioinformatics* 2006, **7:**113.
3. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L and Bolouri H: **A genomic regulatory network for development.** *Science* 2002, **295(5560):**1669–1678.
4. Tsonis P: *Anatomy of gene regulation* Garland Publishing; 2003.
5. Davidson EH: *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution* Academic Press; 2006.
6. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM and 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 USA* 2002, **99(2):**757–762.
7. Rajewsky N, Vergassola M, Gaul U and Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:**30.
8. Pierstorff N, Bergman CM and Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA.** *Bioinformatics* 2006, **22(23):**2858–2864.
9. Philippakis AA, He FS and Bulyk ML: **Modulefinder: a tool for computational discovery of cis regulatory modules.** *Pac Symp Biocomput* 2005, 519–530.
10. Berezikov E, Guryev V, Plasterk RHA and Cuppen E: **CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting.** *Genome Research* 2004, **14:**170–178.
11. Sharan R, Ben-Hur A, Loots GG and Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome.** *Nucleic Acids Res* 2004, 32 Web Server: W253–W256.
12. Nazina A and Papatsenko D: **Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency.** *BMC Bioinformatics* 2003, **4:**65.

13. Papatsenko D: **ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors.** *Bioinformatics* 2007, **23(8):**1032–1034.
14. Lifanov AP, Makeev VJ, Nazina AG and Papatsenko DA: **Homotypic Regulatory Clusters in Drosophila.** *Genome Research* 2003, **13 (4):**579.
15. Abnizova I, te Boekhorst R, Walter K and Gilks WR: **Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test.** *BMC Bioinformatics* 2005, **6:**109.
16. Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefèbvre C, Deblois G, Giguère V, Ferretti V, Bergeron D, Coulombe B and Robert F: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16(5):**656–668.
17. Sosinsky A, Honig B, Mann RS and Califano A: **Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting.** *Proc Natl Acad Sci USA* 2007, **104(15):**6305–6310.
18. Sinha S and He X: **MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules.** *PLoS Comput Biol* 2007, **3(11):**e216.
19. Hu J, Hu H and Li X: **MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs.** *Nucleic Acids Res* 2008, **36(13):**4488–4497.
20. Li L, Zhu Q, He X, Sinha S and Halfon MS: **Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.** *Genome Biol* 2007, **8(6):**R101.
21. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD and Eisen MB: **Large-scale turnover of functional transcription factor binding sites in Drosophila.** *PLoS Comput Biol* 2006, **2(10):**e130.
22. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK and Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse.** *Nat Genet* 2007, **39 (6):**730–732.
23. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E and Taipale J: **Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.** *Cell* 2006, **124:**47–59.
24. Meng H, Banerjee A and Zhou L: **BLISS 2.0: a web-based tool for predicting conserved regulatory modules in distantly-related orthologous sequences.** *Bioinformatics* 2007, **23 (23):**3249–3250.
25. Vlieghe D, Sandelin A, De Bleser P, Vleminckx K, Wasserman W, van Roy F and Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34:**D95–97.
26. Rahmann S, Mueller T and Vingron M: **On the power of profiles for transcription factor binding site detection.** *Stat Appl Genet Mol Biol* 2003, **2:**, Article7.
27. Chan BY and Kibler D: **Using hexamers to predict cis-regulatory motifs in Drosophila.** *BMC Bioinformatics* 2005, **6:**262.
28. Wasserman WW and Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167–81.
29. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A and Searle S: **Ensembl 2008.** *Nucleic Acids Res* 2008, 36 Database: D707–14.
30. Krivan W and Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11(9):**1559–66.
31. Halfon MS, Gallo SM and Bergman CM: **REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila.** *Nucleic Acids Res* 2008, 36 Database: D594–8.
32. Wilson RJ, Goodman JL, Strelets VB and Consortium F: **FlyBase: integration and improvements to query tools.** *Nucleic Acids Res* 2008, 36 Database: D588–D593.
33. Panne D, Maniatis T and Harrison S: **An atomic model of the interferon-beta enhanceosome.** *Cell* 2007, **129:**1111–1123.
34. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, curators HF, Project BDG, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM and Kellis M: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.** *Nature* 2007, **450(7167):**219–232.