BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 19 2010, pages 2391–2397
doi:10.1093/bioinformatics/btq453

Sequence analysis

Advance Access publication August 9, 2010

An alignment-free model for comparison of regulatory sequences

Hashem Koohy[1,*], Nigel P. Dyer[1], John E. Reid[2], Georgy Koentges[3] and Sascha Ott[4,*]

[1]MOAC Doctoral Training Centre, Coventry House, University of Warwick, Coventry, CV4 7AL, [2]MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR, [3]Department of Biological Sciences, Gibbet Hill Campus and [4]Warwick Systems Biology Centre, Coventry House, University of Warwick, Coventry, CV4 7AL, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Some recent comparative studies have revealed that regulatory regions can retain function over large evolutionary distances, even though the DNA sequences are divergent and difficult to align. It is also known that such enhancers can drive very similar expression patterns. This poses a challenge for the *in silico* detection of biologically related sequences, as they can only be discovered using alignment-free methods.

Results: Here, we present a new computational framework called Regulatory Region Scoring (RRS) model for the detection of functional conservation of regulatory sequences using predicted occupancy levels of transcription factors of interest. We demonstrate that our model can detect the functional and/or evolutionary links between some non-alignable enhancers with a strong statistical significance. We also identify groups of enhancers that are likely to be similarly regulated. Our model is motivated by previous work on prediction of expression patterns and it can capture similarity by strong binding sites, weak binding sites and even the statistically significant absence of sites. Our results support the hypothesis that weak binding sites contribute to the functional similarity of sequences.

Our model fills a gap between two families of models: detailed, data-intensive models for the prediction of precise spatio-temporal expression patterns on the one side, and crude, generally applicable models on the other side. Our model borrows some of the strengths of each group and addresses their drawbacks.

Availability: The RRS source code is freely available upon publication of this manuscript: http://www2.warwick.ac.uk/fac/sci/systemsbiology/staff/ott/tools_and_software/rrs

Contact: s.ott@warwick.ac.uk; hashem.koohy@warwick.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 20, 2010; revised on July 30, 2010; accepted on August 4, 2010

1 INTRODUCTION

Cis-regulatory modules (CRMs) can drive precise spatio-temporal gene expression patterns. Some recent studies show that CRMs may function similarly in different species despite substantial sequence divergence (Hare *et al.*, 2008; Ludwig *et al.*, 2005). This implies that, first, alignment-based sequence comparison tools are not applicable for further decoding the conserved function of such CRMs and second, some CRMs must share common patterns that drive almost identical regulatory outputs but possibly with different arrangements of binding sites. When different, but functionally related enhancer loci in the same species are considered, then alignment-based tools are not normally suitable for regulatory sequence comparisons as these sequences are not orthologous.

Here, we present an alignment-free regulatory sequence comparison model called Regulatory Region Scoring (RRS) model, which is based on the potential distribution of transcription factors (TFs). Our goals are:

(1) To be able to detect functionally similar enhancer regions even if the enhancer regions do not align.

(2) To find groups of similar enhancers and determine relevant sequence features shared among enhancers within a group.

1.1 Previous work

There has been a great deal of attention over alignment-free methods to further reveal the mechanism of transcription control (see Vinga and Almeida, 2003). Among these methods, two families are of particular interest.

Models in the first family are aimed at predicting spatio-temporal gene expression patterns from the regulatory sequences. A model in this family was recently developed by Segal *et al.* (2008), and then attracted the attention of other researchers (see Gertz *et al.*, 2009; Loo and Marynen, 2009; Segal and Widom, 2009; Zinzen *et al.*, 2009). For previous related work, see Djordjevic *et al.* (2003); Foat *et al.* (2006); Roider *et al.* (2007); Tanay (2006); Vinga and Almeida (2003). The model by Segal *et al.* (2008) is based on a thermodynamic equilibrium assumption. The probability of polymerase occupancy is computed from the intrinsic equilibrium affinities and concentrations of the TFs. The gene expression level is considered to be proportional to the polymerase occupancy. This model takes into account some important aspects of TF–DNA interaction including competition of TFs for TF binding sites, self-cooperativity of TFs, and the effects of weak binding sites. Although this model advances our understanding of how genomic sequences are translated into transcriptional outputs, the complexity and data dependency of the model do not allow for a wide application of this model as a sequence comparison tool. One must deal with a complex model-fitting procedure and provide a combination of data that is rarely found at present: spatial expression patterns of a number of related enhancers, knowledge of what the key regulators are, their binding motifs and their spatial concentration.

*To whom correspondence should be addressed.

The second family of alignment-free methods is based on the rationale that functionally similar sequences must share some common words. Within these methods each sequence is associated with a vector of k-mer counts. A distance function for these vectors is then defined (Aerts *et al.*, 2003; Blaisdell, 1986; Kantorovitz *et al.*, 2007; Leung and Eisen, 2009; van Helden, 2004). From this family, D2z (Kantorovitz *et al.*, 2007) is of particular interest for us because: (i) it was one of the first alignment-free methods for detection and analysis of regulatory sequences; (ii) it is a normalized version of 'D2' Lippert *et al.* (2002), a natural method of comparison of k-mers in two sequences. The background normalization of this model makes it possible to account for sequences from different background distributions. However, some limitations of this method are:

(1) Not all functional motifs in a pair of sequences are in the form of 6-mers. So, by considering only 6-mers as patterns underlying functional similarity of a pair of sequences, some motifs that contribute to the gene expression pattern may be overlooked.

(2) Not all 6-mers are biologically meaningful words, hence using all 6-mers may mean introducing some noise to the model and furthermore, we may want to compare two sequences just based on a subset of meaningful words.

(3) Within the D2z framework, degeneracy of TF binding motifs is not accounted for. So, different 6-mers are treated separately even if they only differ in one base.

(4) The framework does not allow for a sequence and its reverse complement to be combined for the purposes of assessing possible TF binding.

There is a gap between these two families of models. The former is based on a mechanistic understanding of the regulation of gene expression by predicting expression patterns using TF occupancy and interaction and is too dependent on a specific combination of datasets to be generally applicable. The latter family of models is defined very generally and is widely applicable, but some natural principles underlying transcriptional control, such as TF competition, motif degeneracy and effects of weak binding sites, are completely ignored. Consequently, the results are less conclusive. The RRS aims to enhance the conclusiveness of the results and lessen the data dependency of the model by borrowing the key ideas of each family of models so as to get more accurate results on a wider range of data.

1.2 Outline of model and results

The first part of our model uses a modification of the thermodynamic model by Segal *et al.* (2008) to compute the expected number of proteins binding to each of a set of motifs in a sequence. This computation captures some of the strengths of Segal's model, including the competition of TFs and contributions of weak binding sites. Each sequence is summarized by these expectations. The second part of our model compares the expectations of different sequences to compute a similarity score. Our model can be used to detect functionally relevant similarity in unalignable sequences (such as certain promoters and enhancers) and it provides insights into shared regulatory codes. Our main results are:

(1) We have developed the model underlying the RRS. The model comprehensively evaluates the possible configurations of proteins occupying given DNA sequences and provides similarity scores based on shared combinations of motifs.

(2) We show that the RRS can detect functional and evolutionary links between enhancers, which do not have significant alignments.

(3) We find that weak binding sites can make a strong contribution to sequence similarity.

(4) Our model treats statistically significant presence and absence of motifs symmetrically. Similarity of sequences can, therefore, be based on a combination of both for a set of motifs. We show examples of motifs making contributions to sequence similarity through their absence.

(5) We use the RRS to create a network of similarities among a set of 131 known fly enhancers. The network connects 34 enhancers with 43 significant pairwise similarity scores. At least four groups of enhancers show strong statistical evidence to function similarly via a shared regulatory code. One of these groups is strongly supported by the existing experimental data.

2 METHODS

The RRS takes as input a pair of sequences and a set of TF motifs. We call one of the sequences the *template sequence* and the other the *test sequence*. The task is to judge whether the test sequence has the potential to drive similar expression patterns as the template sequence, assuming expression is driven by the given set of motifs. We do not use any cutoff for probabilities of binding of these motifs to the sequences and so allow weak binding events and even absence of motifs to contribute to sequence similarity. The output from RRS is a statistical similarity score and a list of motifs that contribute to that similarity score. The model is built of two main components: one component associates each sequence with a mathematical vector reflecting which proteins with what multiplicity and what specificity have the potential to be bound to the sequence. We call the elements of these vectors *motif occupancy values* or, in short, *o-values*. These vectors give an indication of the potential enhancer function of the given sequences. The second component estimates a probability distribution of motif *o-value* vectors for sequences that function similar to the template sequence. We then compute a Bayes factor to evaluate if the test sequence is more similar to the template sequence or more similar to random background sequences (supplementary Fig. S1 shows a simplified schematic illustration of the RRS concept).

2.1 Occupancy values of proteins binding a sequence (motif *o-values*)

We assume a template sequence T, a test sequence S, and a set of TF motifs $\mathcal{M}=\{M_1,\dots,M_n\}$. We use the term configuration to denote a particular arrangement of protein molecules along the DNA sequence, which is defined by the subsequences at which each molecule is bound to the sequence. Valid configurations are those in which binding subsequences do not overlap. A configuration c with N molecules bound to a sequence is defined as $c=\{(M_i,P_i)|1\leq i\leq N,\ M_i\in\mathcal{M}\ \}$, where M_i is the i-th molecule bound at a subsequence starting at position P_i. Note that N is the number of molecules bound in the configuration while n is the number of TF motifs. The length of the binding subsequence is the size of motif M_i. Further, we denote the subsequence starting at position P_i, where molecule M_i has bound, by B_i. Therefore $p(B_i|M_i)$ means the probability of subsequence B_i using the the corresponding PSSM model and $p(B_i|\bar{M_i})$ means the probability of subsequence B_i given the background model (uniform 0−order

Markov model in our case). We then associate a statistical weight with this configuration:

$$W(c)=\prod_{i=1}^{N}\frac{p(B_i|M_i)}{p(B_i|\bar{M}_i)}=\prod_{i=1}^{N}\frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)}\times\frac{p(\bar{M}_i)}{p(M_i)} \tag{1}$$

in which $\frac{p(B_i|M_i)}{p(B_i|\bar{M}_i)}$ is the contribution of molecule i bound to the sequence at position P_i. This variable is in turn a function of binding affinity i.e., $\frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)}$ and concentration parameter i.e., $\frac{p(\bar{M}_i)}{p(M_i)}$. This definition enables weak binding events to be included in the model. Assume that in a configuration c we have a molecule that has been weakly bound to the sequence many times. If, for the sake of simplicity we assume an equal binding affinity a ($a>1$) in K positions, then the contribution of this factor to the $W(c)$ is equal to a^K. Depending on K, this might be a strong contribution. The probability of each configuration c is then defined as $p(c)=\frac{W(c)}{\sum_{c\in C}W(c)}$ where C is the set of all valid configurations. We use the same dynamic programing technique as in Segal *et al.* (2008) to compute this probability.

There can be more than one expressed protein species that can bind to a given motif. In the absence of information on either the number of protein species capable of binding a motif or the nuclear concentrations of these proteins we assume the total nuclear concentration of such proteins to be equal for each motif and set $\frac{p(\bar{M}_i)}{p(M_i)}$ to a constant value. Where such information is available it can be integrated into the RRS by setting the concentration parameters accordingly. When the concentration parameter is set to a constant value, it determines the average density of proteins bound to DNA within our model. We chose 15 as the setting for the concentration parameter and confirmed that results presented in this work are robust as long as the concentration parameter is set such that the protein density is realistic. Note that the scaling of this parameter depends on the scaling of the binding affinity and, therefore, the absolute value does not have a direct interpretation.

Intuitively, this probability distribution over all possible configurations should reflect a number of aspects of enhancer function in a natural way. Overlapping binding sites will compete with each other, high-affinity binding sites will attract a binding molecule more often, and weak binding sites can exert an effect if they are present in numbers. Proteins are more likely to interact with the polymerase if they occupy the enhancer more often. Therefore, a key quantity relevant to the function of an enhancer is the expected number of copies of a given protein that bind to motifs in the enhancer (T):

$$e^T_{M_i}=\log\sum_{c\in C}p(c)I_{M_i}(c) \tag{2}$$

in which $I_{M_i}(c)$ is the number of occurrences of motif M_i in configuration c. This definition is of particular interest because it captures both the specificity and multiplicity of a binding event of a protein to the sequence in the $p(c)$ and $I_{M_i}(c)$ terms, respectively. A dynamic programming approach is used to compute each occupancy value. Finally, the sequence T is associated with the vector of occupancy values, i.e. $E^T=<e^T_{M_1},\cdots,e^T_{M_n}>$ and similarly sequence S is associated with $E^S=<e^S_{M_1},\cdots,e^S_{M_n}>$. Our results show that these occupancy values are length dependent. We divide them by the length of the sequences to normalize them. Therefore, each of these vectors summarizes the combined specificity and multiplicity that each protein is likely to bind to each of the sequences.

2.2 Similarity scores

Our aim in this section is to define a similarity function over the space of vectors of occupancy values to extract the similarity of a given pair of *o-values*. Having observed *o-values* from the template sequence, E^T, we want to test if the vector of *o-values* from the test sequence, E^S, has been drawn from the same distribution or from a random background distribution. The logarithm of motif *o-values* in randomly picked sequences from the genome of the species of interest approximates a normal distribution. Therefore, the probability of a motif o-vector such as $E^S=<e^S_{M_1},\cdots,e^S_{M_n}>$ can be obtained from a multivariate normal distribution. For the sake of simplicity, we shall consider an independent multivariate normal distribution. This means that the probability of the o-vector E^S under the random model is $p(E^S|R)=\prod_{i=1}^{n}p(e^S_{M_i}|\mu=\mu_{R_i},\sigma=\sigma_{R_i})$, where, μ_{R_i} and σ_{R_i} are the mean and SD of *o-values* for motif i in randomly picked sequences. The probability that E^S has been drawn from the same distribution as the template is $p(E^S|T)=\prod_{i=1}^{n}p(e^S_{M_i}|\mu=e^T_{M_i},\sigma=\sigma_{R_i})$. We define the RRS score as

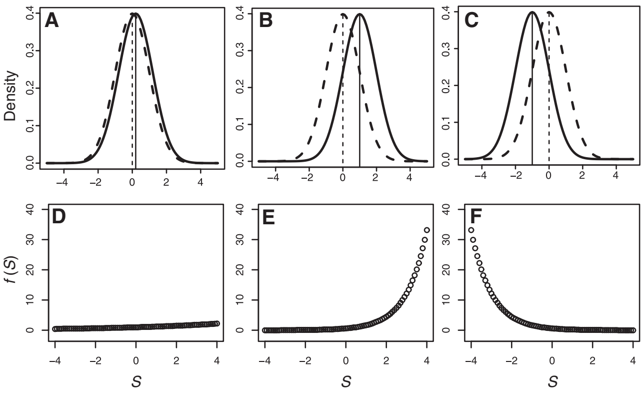

Fig. 1. Illustration of the RRS similarity score for an individual motif. There are three possibilities. (**A**, **D**) The motif is neither significantly present nor absent in the template sequence. The distribution of motif *o-values* in sequences with the same function as the template sequence (solid line) is estimated to be equal to the random background (dashed line). In this case, irrespective of the motif *o-value* in the test sequence, the function $f(e^S_{M_i})$ [see Equation (4)] is constant (D). (**B**, **E**) The motif *o-value* in the template is higher than in random sequences (B), in this case $f(e^S_{M_i})$ is an increasing function (E). (**C**, **F**) The motif *o-value* is lower than in random sequences, indicating significant absence. In this case, $f(e^S_{M_i})$ is decreasing (F).

$$\mathrm{RRS}(S|T)=\frac{p(E^S|T)}{p(E^S|R)} \tag{3}$$

The first point to note about this definition is that it is asymmetric but one may define it as an average to make it symmetric, i.e. $\mathrm{RRS}(S,T)=(\mathrm{RRS}(S|T)+\mathrm{RRS}(T|S))/2$. However, it is sensible to work with the asymmetric version, in particular when comparing two sequences from different species. The second point that makes this definition more realistic and useful is the contribution of the individual motifs, which are the factors making the RRS-score:

$$f(e^S_{M_i}):=\frac{p(e^S_{M_i}|\mu=e^T_{M_i},\sigma=\sigma_{R_i})}{p(e^S_{M_i}|\mu=\mu_{R_i},\sigma=\sigma_{R_i})} \tag{4}$$

for any motif M_i, where $1\leq i\leq n$. For any test sequence S, one can consider the Equation (4) as a function of variable $e^S_{M_i}$ with three extra parameters: $e^T_{M_i}$, μ_{R_i} and σ_{R_i}. The following cases illustrate this definition and its usage in the rest of this article:

(1) If $e^T_{M_i}\approx\mu_{R_i}$ (see Fig. 1A), then $f(e^S_{M_i})$ can be considered as a constant function with value ≈ 1 (Fig. 1D). This means that if the expected number of occurrences of this motif in the template sequence is very close to the average of its expected number of occurrences in the random sequences, then the overall RRS score for the test sequence will be largely independent of number of occurrences of this motif in the test sequence. In biological terms, if the test sequence shares a regulatory code with the template sequence, but also contains additional binding sites, then these additional sites do not reduce the sequence similarity.

(2) If $e^T_{M_i}>\mu_{R_i}$, then $f(e^S_{M_i})$ is an increasing function. More accurately, if we assume that $e^T_{M_i}>A>\mu_{R_i}$, where A is the intersection point of the two distribution curves (Fig. 1), then $f(e^S_{M_i})\leq 1$ if $e^S_{M_i}\leq A$ else it is

greater than one. This case occurs when the motif is strongly present in the template sequence. Accordingly, the greater the motif *o-value* in the test sequence, the greater the contribution of the motif (Fig. 1B and E). Note that a strongly negative RRS score in this case implies poor presence of the motif in the test sequence.

(3) Similarly, if $e^{T}_{M_i} < \mu_{R_i}$, then $f(e^{S}_{M_i})$ is a decreasing function. In other words, $f(e^{S}_{M_i}) > 1$, if $e^{S}_{M_i} < A$ (where $e^{T}_{M_i} < A < \mu_{R_i}$ is the intersection point of two curves) then the motif will be assigned a contribution greater than one, otherwise $f(e^{S}_{M_i})$ has a value less than one, contributing negatively to sequence similarity (Fig. 1C and F).

3 DISCUSSION AND RESULTS

Our goal in this section is to evaluate if the RRS can distinguish functionally/evolutionarily related sequence pairs (positive sets) from the sequence pairs randomly picked from the genome (negative sets). For this, we apply it to the same fly datasets as used in Kantorovitz *et al.* (2007). We first demonstrate that the distribution of alignment significance levels, or *e*-values in short, of positive sets is not significantly different from the distribution of alignment *e*-values of negative sets. Using RRS, however, there are about 40 pairs of sequences (edges in Graph 3) whose scores are significantly greater than the scores obtained using random pairs. The statistical significance of some of these scores are highlighted. We show that according to the RRS results, a subset of these 40 enhancers are regulated by the regulator Bicoid (BCD) (subgraph highlighted by rectangles in Fig. 3). This finding is of particular significance as it has been experimentally confirmed by Ochoa-Espinosa *et al.* (2005). Finally, we do some analysis, first, to show the contribution of strongly absent motifs to the similarity of a pair of sequences, and second to highlight the substantial contribution of weak binding sites in our model scheme.

3.1 Datasets

This study uses the same fly datasets as are used in Kantorovitz *et al.* (2007), which are four positive and four negative sets. The positive sets consist of 82 FLY_BLASTODERM, 23 FLY_PNS, 9 FLY_TRACHEAL and 17 FLY_EYE enhancers. The negative sets consist of 82 BLASTODERM, 23 PNS, 9 TRACHEAL and 17 EYES counterparts all of which were randomly picked from non-coding parts of the genome (each sequence has the same length as its corresponding real enhancer). RRS scores of each pair in a positive set as well as negative set is obtained. It is then assessed if pairs in a positive set score higher than pairs in the counterpart negative set. This is done by sorting all scores and then looking at top $K = \frac{k(k-1)}{2}$ pairs, where k is the number of enhancers in that set. For the set of TF motifs, we used 67 insect-specific PSSMs available in the TRANSFAC database (Matys *et al.*, 2003). The full list of the motif-IDs is included in the Supplementary Material.

3.2 Statistical links between sequences

We first used a local sequence alignment tool from the NCBI (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi; ‘Blast 2 Sequences’) as well as an implementation of the Smith–Waterman algorithm (the water tool from the EBI; http://www.ebi.ac.uk/Tools/emboss/align/index.html) to show that these sequences are not alignable. The best hit found over all of these sets for BLAST had an *e*-value of 1e-08 corresponding to a stretch of 23 bp from a pair in the negative BLASTODERM set (Supplementary Table S1). Supplementary Figure S2 shows the results for both algorithms in BLASTODERM positive and negative sets. Therefore, by looking at only the alignment scores, one cannot say if a particular pair is likely to be from the positive set or negative set.

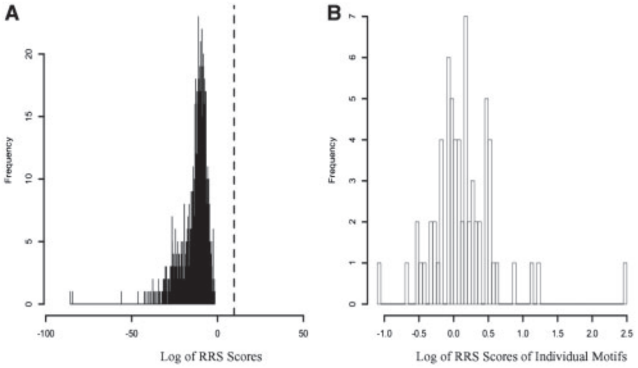

Fig. 2. (**A**) Illustrating the the statistical significance of the RRS score of *eve_stripe*1 vs *oc_otd-186*. The dashed vertical line shows the log of RRS score from this pair, which is 9.64. The black histogram shows the distribution of log of RRS scores of *eve_stripe*1 versus 1000 randomly picked sequences from *D.melanogaster* longest chromosome. (**B**) Depiction of the contribution of individual motifs in the RRS scheme. Shown here, is the distribution of the individual motif scores in comparison of *eve_stripe*1 versus *oc_otd-186*. Three strongly positively contributed factors that are obtaining scores above 1, in descending order, are: BCD, KR and FTZ. The factor that is negatively contributing to this scheme, i.e. obtaining a score less than −1 is SRY_β.

The functional conservation of these sequences presents a very different picture. To examine this, we looked at the RRS scores of all pairs of sequences in any of both positive and negative sets. For instance, in BLASTODERM enhancers, 43 out of 50 top scores belong to pairs from the positive set. The best (log of) RRS score was 9.64 corresponding to the comparison of *eve_stripe*1 (length 801 bp) with *oc_otd-186* (length 187 bp). To check the statistical significance of the RRS score, we compared *eve_stripe*1 with 1000 sequences randomly picked from the longest chromosome of the *Drosophila melanogaster* genome, with length ranging from 100 bp up to 3000 bp. Interestingly, when comparing *eve_stripe*1 with these random sequences, no pairs gave an RRS score with log greater than 0. The result of this analysis is illustrated in Figure 2A, in which the vertical dashed line is a reference line to show the position of the RRS score from *eve_stripe*1 versus *oc_otd-186* and the black histogram is the distribution of the RRS scores of *eve_stripe*1 versus 1000 randomly picked sequences.

We went on to consider what motifs contribute to the functional conservation that is seen. If the log of the score for a specific motif is greater than 1 (see section 2.2), this indicates a significant similarity between the presence of the motif in the template and test sequence either by multiplicity or by specificity. An RRS score around zero is expected for a random DNA sequence and scores of less than −1 indicates a significant dissimilarity between the presence of the motif in the two sequences. RRS scores of all 67 insect motifs individually computed. Figure 2B depicts the distribution of these scores. As we can see, there are three factors that are assigned scores greater than one. These factors are (in descending order): BCD, Krüppel (KR) and fushi tarazu (FTZ). This means that according to our model these three factors are main functional similarity-makers of this pair of enhancers. In comparison to the background sequences, all of these three factors are strongly presented in both of these sequences (see Section 3.4). This finding is of particular significance as it is supported by Ochoa-Espinosa *et al.* (2005), where they show both computationally and experimentally that the regulation of the eve1 plus 10 other CRMs are strongly dependent to the regulator BCD. This suggests that the BCD is a regulator for *oc_otd-186*, too. We will come back to this point in more detail in Section 3.3.

3.3 Identification of enhancers with similar function

To make a more global analysis of these enhancers rather than analysing each individual set of enhancers, we put all 131 enhancers into one set (referred to as G_Positive set). Similarly, all 131 randomly picked counterpart sequences were placed into another set called G_Negative set. The RRS scores were computed and a directed graph was generated, in which each node is an enhancer from the G_Positive set and each edge represents a high RRS score for two corresponding nodes. The threshold for inclusion of edges was set above the maximum score within the G_Negative set (equal to three). Therefore, only enhancer pairs that are scored above any pair from the G_Negative set are shown. The resulting graph (see Fig. 3) shows the RRS prediction of the functional and/or evolutionary relationship of the enhancers associated to the top 43 scores from the G_Positive set. From this graph, we can see that only 34 enhancers (nodes) are associated to these 43 scores (edges). This means that some of these enhancers are in a close relationship with others in a sense that they paired up more often than just by random chance. For instance, HLHg* is paired up with six other enhancers ($P <$1e-04, P-value of binomial test for one node out of 131 to be part of six or more edges). The presence of a large number of high-scoring edges and the dense connectivity of the graph confirm that the RRS uncovers statistically significant structure in this dataset.

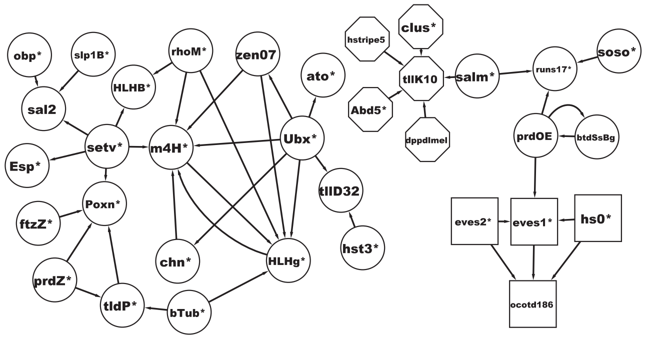

Fig. 3. This graph represents the functional relationship of some of the top-scored enhancers from the G_Positive set. Each node represents an enhancer and $Enhancer_1 \rightarrow Enhancer_2$ means that the $\log(RRS(Enhancer_1, Enhancer_2)) \geq 3$. The threshold 3 is to filter out other scores that are less than a score from the G_Negative set. Enhancers with star signs are abbreviated names. For full names of these enhancers see Supplementary Table S3.

We might want to think of the subgraph highlighted by rectangular nodes as a core subgraph because: first, all of the four nodes are from BLASTODERM enhancers; second, it contains a pair that gets the highest score in BLASTODERM enhancers; and third, it satisfies a transitivity property. Focusing more deeply on this subgraph reveals that, according to our analysis, the factor BCD is the most strongly contributing factor in the functional similarity of any pair in this subgraph. This significant finding is experimentally supported by Ochoa-Espinosa *et al.* (2005), where the regulation of the *eve_stripe*1, *eve_stripe*2 and *hstripe*0 and eight more CRMs are reported to be strongly dependent to the activator BCD. They also showed that many of the BCD-dependent CRM contain a cluster of the gap protein Krüppel, which is again in a high agreement with ours (see Table 1) in that in all of these five comparisons KR is either the second or third strongly contributed factor. We must recall that according to our model, a motif can obtain a high score either by its strong presence (because of multiplicity or specificity) or by strong absence in both sequences. It is also important to note that the five enhancers in this subgraph are regulated by a set of common factors (as colour-coded in Table 1), and this might be the reason that RRS can almost distinguish it as a subgraph. Supplementary Table S2 provides similar results for the subgraph with octagon-shaped nodes distinguished by the RRS and a set of common motifs that we predict to regulate that subgraph.

Overall, these findings reveal that our model indeed captures some of the core principles governing functional conservation of modules, and hence performs much better than random expectation.

3.4 Contributions of motif absence and weak binding sites

We are interested to see whether the strong absence of a motif in a pair of sequences can underly the statistically significant similarities we observed. We looked for motifs that are associated with a relatively high RRS score but whose associated *o-values* are lower than the *o-values* of the motif in random sequences. In Section 2.1 and Figure 1, we considered two situations where a motif is assigned a high RRS score because the motif is strongly present or it is strongly missing in both sequences. The strong presence may be more intuitive and it is illustrated in Figure 4A1 and A2, where we can see both RRS scores for any of the 67 used motifs in comparison of the *eve_stripe*1 versus *oc_otd-186* (Fig. 4A1) and

Table 1. Top five factors that are strongly contributing to the functional similarities of each pair in the subgraph highlighted by rectangles in Figure 3

Pair of enhancers	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
eve_stripe1 versus oc_otd-186	I$*BCD*_01	I$*KR*_01	I$*FTZ*_01	I$*HSF*_03	I$*HSF*_04
eve_stripe2 versus oc_otd-186	I$*BCD*_01	I$*KR*_01	I$*FTZ*_01	I$*HSF*_03	I$*MAD*_*Q6*
hstripe0 versus oc_otd-186	I$*BCD*_01	I$*KR*_01	I$*MAD*_*Q6*	I$*HAIRY*_01	I$*HSF*_04
hstripe0 versus eve_stripe1	I$*BCD*_01	I$*STAT*_01	I$*KR*_01	I$*GCM*_01	I$*EVE*_*Q6*
eve_stripe2 versus eve_stripe1	I$*BCD*_01	I$*FTZ*_01	I$*KR*_01	I$*GCM*_01	I$*TTK69*_01

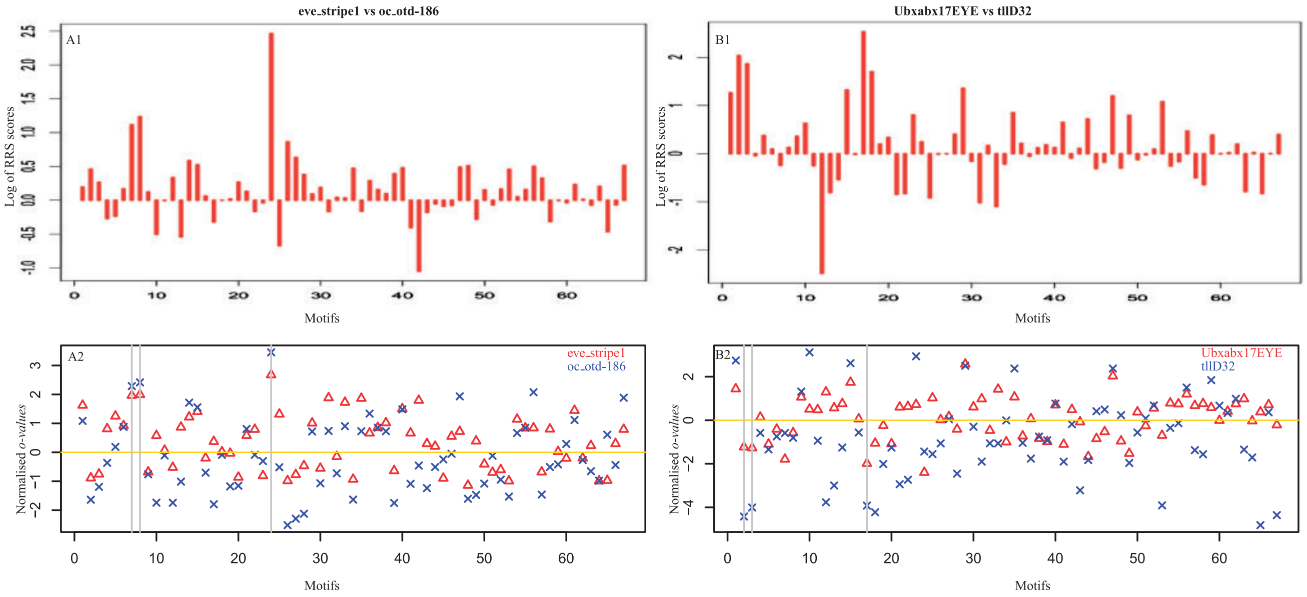

Fig. 4. (**A1**) shows the log of RRS scores for each of the 67 insect motifs that were used for the comparison of *eve_stripe*1 versus *oc_otd-186*. Motifs 24, 8 and 7 (in descending order) are the three top contributors to this comparison. (**A2**) illustrates the *o-values* of these motifs from *eve_stripe*1 (red) and from *oc_otd-186* in blue. The *y*-axis is the number of SDs that an o-value deviates from the mean. The yellow base line shows the background *o-values*. The vertical lines highlight the positions of the top three motifs by RRS score. The main feature of (A1) and (A2) is that motifs with high RRS scores (A1) have *o-values* considerably higher than background level (A2), indicating strong presence of the motifs. (**B1**) and (**B2**) show an example where strong absence of motifs contributes to the statistical link between the sequences. (B1) shows the individual contributions of each of the motifs in the comparison of *Ubxabx*17*EYE* with *tllD*32. In (B2), the *o-values* of the motifs from *Ubxabx*17*EYE* are shown in red and those from *tllD*32 in blue. The three motifs that contribute strongly to the RRS scores (motifs 17, 2 and 3 in descending order) all have *o-values* less than background. This is referred to as a strongly absent motif.

also the normalized vector of *o-values* for *eve_stripe*1 in red and *oc_otd-186* in blue (Fig. 4A2). The golden base line is to show the *o-values* from the background (random sequences). Motifs 24, 8 and 7 associated with the top three RRS scores (in order) in *eve_stripe*1 versus *oc_otd-186* comparison. The reader can see from *A*2 that for all of these three motifs, the motif *o-values* are considerably higher than background. This is called strong presence of motifs in both sequences. But the interesting part is shown in Figure 4B1 and B2, where first we can see again in Figure 4B1 the contribution of the individual motifs to the RRS scores of Ubxabx17EYE versus tllD32 and in *B*2 the *o-values* from Ubxabx17EYE in red, tllD32 in blue and motifs that are obtaining the top three RRS scores. We see that all three motifs are associated with *o-values* lower than background (strong absence) but these contribute to the RRS score and, therefore, to the recognition of functional conservation.

The contribution of weak binding sites to the RRS scores can be seen in Figure 2B and C. The log of RRS score for *eve_stripe*1 versus *oc_otd-186* is 9.64. This is the sum of scores of each motif. The four motifs making the strongest contribution only contribute about half of this score (Fig. 2C), while any RRS score above 0 is still significantly different from noise as none of the random sequences evaluated in Figure 2B had a score above 0. Therefore, the similarity of these two enhancers cannot be solely attributed to strong binding sites. This is consistent with previous findings in Segal *et al.* (2008).

4 CONCLUSION

We have presented an alignment-free method for detection of functional conservation of the regulatory sequences based only on occupancy level of some TFs of interest. It has been designed such that it is less data-dependent with a wider range of applications and more conclusive results. We have demonstrated that this model can be used for comparison of regulatory sequences, where sequences are functionally related but are not orthologous. The RRS can also be used for comparison of regulatory sequences from different species, where they have undergone a substantial evolutionary divergence. For statistical validation of the RRS scores, the sequences that obtained top scores were compared with 1000 randomly picked sequences and showed that it is not possible to get such a high RRS scores just by chance. We have shown that the RRS can significantly detect the functional and/or evolutionary similarities of the regulatory sequences. In particular, RRS can categorize some enhancers that are regulated by a set of common factors, a result that was in high agreement with experimentally validated reports. Based on predictions of our model, we have proposed the hypothesis that strong absence of a motif in pair of sequences might be a feature for functional conservation. Finally, we would like to close this article by listing some finer points and shortcomings of our model, where further development may lead to a more accurate model.

- In the current version of the RRS, we use a set of known TF motifs, focusing the sequence analysis on validated motifs. However, there may be yet unknown binding motifs relevant to the function of the sequences analysed. We could introduce some complementary sequence patterns into the analysis to test for a possible contribution to sequence similarity.
- There are further sources of prior knowledge that could be fed into the analysis in principle. For example, we are assuming

equal concentrations of all regulators even though these will vary in different cell types. Some motifs belong to particular pathways that may be of particular interest in some cases. It would be possible to define a weight for such subsets of motifs.

- Within the current version, the synergy between pair of motifs is ignored, but there are some reports that regulation of some fly enhancers requires synergy between pairs of motifs (Simpson-Brose *et al.*, 1994).
- Rather than using a single template sequence, it would be possible to use multiple template sequences with similar expression pattern. This should help to define a more accurate distribution of motif *o-value* vectors.

ACKNOWLEDGEMENTS

The authors are thankful to the reviewers for their valuable comments which helped improving the manuscript. We would also like to acknowledge discussions with Dr Miguel A. Juárez, Dr Liqun Luo, Dr Maria Spletter and people from both SO group and GK group that led to further improvements of the model.

Funding: H.K. was funded by the Human Frontier Science Program Organization (HFSPO grant RGP0029/2007-C awarded to G.K.). N.D. acknowledges funding from UK Engineering and Physical Sciences Research Council (EPRSRC). S.O. acknowledges funding from the Research Councils United Kingdom (RCUK) with whom he holds an Academic Fellowship. The computing facilities were provided by the Centre for Scientific Computing of the University of Warwick with support from the Science Research Investment Fund.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2003) Computational detection of cis-regulatory modules. *Bioinformatics*, **19** (Suppl. 2), ii5–ii14.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl Acad. Sci. USA*, **83**, 5155–5159.

Djordjevic,M. *et al.* (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.

Foat,B.C. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, **22**, e141–e149.

Gertz,J. *et al.* (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.

Hare,E.E. *et al.* (2008) Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet.*, **4**, e1000106.

Kantorovitz,M.R. *et al.* (2007) A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, **23**, i249–i255.

Leung,G. and Eisen,M.B. (2009) Identifying cis-regulatory sequences by word profile similarity. *PLoS One*, **4**, e6901.

Lippert,R.A. *et al.* (2002) Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl Acad. Sci. USA*, **99**, 13980–13989.

Loo,P.V. and Marynen,P. (2009) Computational methods for the detection of cis-regulatory modules. *Brief. Bioinform.*, **10**, 509–524.

Ludwig,M.Z. *et al.* (2005) Functional evolution of a cis-regulatory module. *PLoS Biol*., **3**, e93.

Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Ochoa-Espinosa,A. *et al.* (2005) The role of binding site cluster strength in Bicoid-dependent patterning in drosophila. *Proc. Natl Acad. Sci. USA*, **102**, 4960–4965.

Roider,H.G. *et al.* (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

Segal,E. and Widom,J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.*, **10**, 443–456.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, **451**, 535–540.

Simpson-Brose,M. *et al.* (1994) Synergy between the hunchback and bicoid morphogens is required for anterior patterning in drosophila. *Cell*, **78**, 855–865.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*., **16**, 962–972.

van Helden,J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, **20**, 399–406.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

Zinzen,R.P. *et al.* (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.