BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 17 2010, pages 2109–2115
doi:10.1093/bioinformatics/btq358

Sequence analysis Advance Access publication July 11, 2010

An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes

Manonmani Arunachalam[1,2], Karthik Jayasurya[2], Pavel Tomancak[1,*] and Uwe Ohler[2,*]

[1]Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany and [2]Institute for Genome Sciences and Policy, Duke University, Durham, NC, USA

Associate Editor: David Posada

ABSTRACT

Motivation: Evolutionarily conserved non-coding genomic sequences represent a potentially rich source for the discovery of gene regulatory region such as transcriptional enhancers. However, detecting orthologous enhancers using alignment-based methods in higher eukaryotic genomes is particularly challenging, as regulatory regions can undergo considerable sequence changes while maintaining their functionality.

Results: We have developed an alignment-free method which identifies conserved enhancers in multiple diverged species. Our method is based on similarity metrics between two sequences based on the co-occurrence of sequence patterns regardless of their order and orientation, thus tolerating sequence changes observed in non-coding evolution. We show that our method is highly successful in detecting orthologous enhancers in distantly related species without requiring additional information such as knowledge about transcription factors involved, or predicted binding sites. By estimating the significance of similarity scores, we are able to discriminate experimentally validated functional enhancers from seemingly equally conserved candidates without function. We demonstrate the effectiveness of this approach on a wide range of enhancers in *Drosophila*, and also present encouraging results to detect conserved functional regions across large evolutionary distances. Our work provides encouraging steps on the way to *ab initio* unbiased enhancer prediction to complement ongoing experimental efforts.

Availability: The software, data and the results used in this article are available at http://www.genome.duke.edu/labs/ohler/research/transcription/fly_enhancer/

Contact: tomancak@mpi-cbg.de; uwe.ohler@duke.edu

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

Received on April 19, 2010; revised on June 25, 2010; accepted on June 29, 2010

1 INTRODUCTION

The identification of functional regulatory elements present in non-coding DNA sequences is complicated by the fact that such elements are highly variable and the sequence features are not sufficiently understood. However, functional elements tend to evolve at slower rates than non-functional regions because they are subject to selection. Due to this slower rate of evolution, comparisons among evolutionarily distant genomes make it possible to use sequence conservation to identify functional regions in the sea of non-coding DNA (Blanchette and Tompa, 2002; Wang and Stormo, 2003). The performance of this 'phylogenetic footprinting' strategy depends on the evolutionary distance between given species and on the conservation level of individual regulatory regions. A number of methods have used cross-species alignments to identify regulatory regions or elements (Cliften *et al.*, 2001; Corcoran *et al.*, 2005; Hardison, 2000; Kellis *et al.*, 2003; Loots *et al.*, 2002); in addition, standard tools such as BLAST (Altschul *et al.*, 1990) are regularly used to identify conserved (segments of) enhancers as well (Erives and Levine, 2004; Hare *et al.*, 2008). The availability of complete genomes also allows for applying general approaches to detect regions under selection in a multiple alignment, such as PhastCons which is based on a phylogenetic hidden Markov model (Siepel and Haussler, 2004).

Despite many individually successful applications, conservation based on sequence alignments can be a poor predictor of functional enhancers (Cliften *et al.*, 2001). In contrast to protein-coding sequences, functional non-coding sequences are frequently not constrained in the ordering and number of functional elements within them (Ludwig *et al.*, 1998; Markstein and Levine, 2002). Comparative sequence analysis have identified significant small-scale insertions, deletions and rearrangements of transcription factor binding sites (TFBSs) within functional modules (Ludwig *et al.*, 2000; Ludwig, 2002). Tracking the evolutionary path of such non-coding elements is proving difficult with current alignment-based methods. A recent study on the conservation of the best characterized eukaryotic enhancers of the even-skipped gene, between *Drosophila* and the highly diverged sepsid flies shows that BLAST finds some short stretches of sequence similarity but scores too low to confirm the evolutionary relationship between those species (Hare *et al.*, 2008). It has also been shown that, especially in distantly related species, the enhancer sequences are simply not alignable (Wolff *et al.*, 1999).

The basic problem we address here thus poses itself as follows: given a regulatory region such as an enhancer without any prior knowledge of functional sites within, can we devise a way to identify its location in a (distantly) related species? And, given a candidate region, can we predict its functionality as an enhancer based on its conservation? With the availability of enormous sets of uncharacterized putative regulatory regions (e.g. based on assays detecting regions of open chromatin), methods tailored to these problems would be immensely useful for providing complementary

*To whom correspondence should be addressed.

evidence, and for studying the evolution of gene regulation at a larger scale. We address these two questions based on so-called alignment-free methods. These methods work under the assumption that similar sequences will share their word (k-mer) composition to some extent, comparing two sequences based on co-occurring words regardless of their order and orientation, thus making no prior assumptions on the presence of particular TFBSs. Vinga and Almeida (2003) performed an early systematic review on a number of different word frequency measures to identify similar regions between two given sequences. Recent studies have proposed and applied alignment-free methods to regulatory sequence problems: Kantorovitz *et al*. (2007) introduced the 'D2Z' score and used it to identify enhancers *ab initio* for genes with similarly annotated function in benchmark *Drosophila melanogaster* data (Ivan *et al*., 2008). Sosinsky *et al*. (2007) combined motif detection in a set of enhancers with a score allowing for local permutations of motifs to identify conserved enhancers. Our work is based on van Helden *et al*. (2004) who introduced a Poisson-based metric which relies on probability theory for comparing sequences on the basis of pattern counts. In the original study, sets of yeast promoters of co-expressed genes were used to identify overrepresented words. The metric, restricted to this subset of words, was then successfully applied to cluster genes based on promoter similarity.

In summary, previous approaches use alignment-free metrics in the same genome to either compare and cluster promoters (van Helden *et al*., 2004), or to identify enhancers for sets of genes known to share regulatory patterns (Chan and Kibler, 2005; Ivan *et al*., 2008; Nazina and Papatsenko, 2003); some also use sets of enhancers with similar function to first determine a subset of words on which similarity is calculated (Sosinsky *et al*., 2007; van Helden *et al*., 2004). Different from these approaches, we here apply alignment-free approach on phylogenetic footprinting problems. Each candidate enhancer is regarded separately, eliminating the need of any additional information on known or predicted motifs, or on similarly regulated genes. We extend the Poisson-based metric to work with multiple related genomes and show that it successfully identifies the location of enhancer orthologs in distantly related species. We also show that appropriate significance values allow for separating functional from non-functional candidates, a task on which other alignment-based methods fail. We evaluate our approach on different sets of validated enhancers from the *D.melanogaster* genome, a well-studied model system for gene regulation.

2 METHODS AND DATA

The basic outline of our approach is as follows: given a candidate or known enhancer sequence in one genome, we scan the corresponding intergenic region in a second genome with a moving window and calculate a similarity score between the enhancer and the current window. The maximum in the resulting similarity profile along the intergenic region is then identified, and the significance of the maximum score is calculated. Below, we explain the individual steps in detail.

2.1 Identification of orthologous intergenic regions

To locate enhancers in related species, we need to first identify the putative orthologous control region. It is well known that orthologous regulatory sequences show a significantly lower level of similarity than their corresponding coding sequences. In order to define the search space, we therefore make the assumption that enhancers will be present in syntenic regions flanked by orthologous genes, and use best reciprocal BLAST hits to identify control regions in a related (non-melanogaster) genome. Repeats and low-complexity regions in the control sequence were masked using DUST (Altschul *et al*., 1990); we observed that the identified intergenic regions for the considered enhancer sets typically ranged from 5 to 50 kb.

2.2 Background word probabilities

To score the occurrence of different words, we need to define their baseline occurrence frequency in the genomes under consideration. We considered the whole genome from each species as background sequence to compute the background frequencies. With frequencies (f_i) for k-mers from a given sequence, the expected number of occurrences m_i of the pattern i in species j in a sequence of length L is obtained by

$$m_i^j = f_i T = f_i(L-w+1), \tag{1}$$

where w the length of the pattern and T the number of possible positions. We fixed the pattern length as 6 and stored the estimated pattern frequencies in precalculated tables for efficiency reasons.

2.3 Poisson-based metric

Following the successful example of van Helden *et al*. (2004), we adapted Poisson-based metrics to compare sequences from two different species. The metric considers word counts (in our case, all k-mers) together with their expected probabilities. It relies on the assumption that pattern occurrences are Poisson distributed and independent of each other. For two given sequences, a Poisson-based similarity score S is calculated based on the probability of common occurrences. Sequence divergence is reflected in a Poisson-based dissimilarity score D which reflects the difference between word occurrences in the two sequences. Finally, a mixed metric M is a score that is defined as weighted combination of the similarity and dissimilarity score.

In our context, we evaluate a given enhancer sequence e from species 1 against a candidate orthologous regulatory control region c from species 2. The Poisson-based similarity score S_i^{ec} for a single word i from the two sequences is calculated as follows:

$$S_i^{ec} = [1 - P_1(x \geq C_i^{ec})] * [1 - P_2(x \geq C_i^{ec})] \tag{2}$$

where C_i^{ec} is the common count for word i in the sequence e and c:

$$C_i^{ec} = \min(N_i^e, N_i^c) \tag{3}$$

N_i^e, N_i^c are the number of times the word i appears in e and c and $P_j(x \geq C_i^{ec})$ is the probability of observing at least C_i^{ec} in e and c from species 1 and 2, respectively.

$$P_j(x \geq C_i^{ec}) = \begin{cases} [1 - F(C_i^{ec} - 1, m_i^j)] \text{ if } C_i^{ec} > 0 \\ 1, \text{ otherwise} \end{cases} \tag{4}$$

where the Poisson distribution function $F(C_i^{ec} - 1, m_i^j)$ gives the probability to observe $\leq C_i^{ec}$ occurrences when the expected value is m_i^j. The multivariate similarity S^{ec} between e and c, over p words, is then obtained by

$$S^{ec} = (1/p) \sum S_i^{ec} \tag{5}$$

The dissimilarity D_i^{ec} is calculated as follows:

$$D_i^{ec} = |F(N_i^e - 1, m_i^1) - F(N_i^c - 1, m_i^2)| \tag{6}$$

and the multivariate dissimilarity D^{ec} across all p words is thus obtained by

$$D^{ec} = \frac{1}{p} \sum D_i^{ec} \tag{7}$$

Finally, the mixed metric M^{ec} is defined as

$$M^{ec} = S^{ec} - aD^{ec} + b, \tag{8}$$

where S^{ec} is the similarity from Equation (5) and D^{ec} is the dissimilarity between e and c from Equation (7), a is a positive weighting parameter which

can be tuned to give more emphasis on the common or distinct occurrences between two sequences, and b is an offset which ensures that the metric is always positive. When two sequences have exactly the same counts for all the patterns, their dissimilarity is 0. This score increases with the number of distinct counts. For any two sequences, if the pattern is found in both sequences, then the mixed metric is positive, which implies that the given sequences are similar to each other. The mixed metric contains a negative contribution for the dissimilar sequences.

2.4 Obtaining a background score distribution

To assess the significance of a mixed metric score, we computed an empirical background distribution by sampling random intergenic sequences from the *D.melanogaster* and a related genome. Starting from *D.melanogaster* sequences >10 kb in length, we identified the orthologous sequences in the distantly related species (*D.ananassae, D.pseudoobscura, D.willistoni, D.mojavensis, D.virilis and D.grimshawi*). We first picked a random orthologous pair, and then sampled windows of equal length at random locations in each region, and computed the mixed metric between the windows [using Equation (8)]. The first and last 500 bp were excluded to remove effects of potential proximal and core promoter sequences which show generally elevated conservation levels. Repeating this procedure 10 000 times established a background mixed metric score distribution of two arbitrary sequence windows from orthologous intergenic regions. We used this distribution to determine score threshold values corresponding to specific significance P-values (typically, $P \leq 0.05$ or 0.01). This step is important to distinguish background sequence conservation, which can be high for compact genomes such as *D.melanogaster*, from regions which show significantly high scores based on the Poisson distance metric.

2.5 Scanning a known enhancer against a control region

To locate an enhancer in an orthologous genome, a known enhancer is scanned against the regulatory control region from a related species in a sliding window fashion. As we do not know the size of the orthologous enhancer, we select a fixed window size in the orthologous region and shift the window by 50-bp steps. Using the background word frequencies from the two species, the mixed metric score is computed for each window using Equation (8). The result of this scanning process is a profile along the region of the related species. We applied the same procedure on the reverse complement sequence of the control region. The window with global maximum mixed metric score on either strand is considered as candidate orthologous enhancer region in the related species. As the true enhancer length may exceed the chosen window size, we merge windows consecutive to the window with the best score if their similarity score exceeds the given threshold value defined by the background score distribution. Orthologous control regions differ in length and typically exceed the window size, and we correct for multiple hypothesis testing in the P-value calculation to take the sequence length into account. A Bonferroni-style corrected P-value is obtained by

$$\text{Corrected } P\text{-value} = [\ P\text{-value} * (\text{Sequence length} / \text{window size})\ ] \quad (9)$$

where P-value is the significance threshold value obtained from the background, and the sequence length is the size of the control orthologous sequence. Note that we approximate the correction factor by the number of non-overlapping windows, as the scores of neighboring overlapping windows will be highly correlated and do not constitute independent tests.

2.6 Combining scores from multiple pairwise scans

For each known enhancer region in *D.melanogaster*, we can identify similar sequences in multiple diverged species by comparing each one individually to *D.melanogaster* in a pairwise fashion. To gain confidence in predicting functionality based on good similarity scores, we combine these pairwise scores. In order to compare multiple species, we need to take the evolutionary distance and relationship into account. For this, we considered available evolutionary distance information and normalized all branch lengths used in a particular combination of species to sum up to 1. We then compute a phylogenetic-tree-reweighting score (PRS) based on the normalized path length between the two compared species (the phylogenetic tree and examples for reweighting are shown in the Supplementary Material). Pairwise PRSs are used as weights to compute a combined P-value across all species j as

$$\text{Weighted } P\text{-value} = \sum_j (\text{PRS} * \text{Corrected } P\text{-value}) \quad (10)$$

This weighted P-value thus indicates the significance of the predicted orthologous enhancer region across multiple non-melanogaster genomes. We limited our analyses to six species outside the melanogaster subgroup, as the branch lengths within the subgroup are comparably short and would not contribute much to a combined score.

2.7 Datasets

To evaluate our approach, we used several datasets from *Drosophila* and non-*Drosophila* insect species, with varying degrees of information about functionality of enhancers outside *D.melanogaster*.

(a) A set of 37 regions from *D.melanogaster* predicted to act as enhancers for anterior/posterior (A/P) embryonic patterning were compiled from Berman *et al*. (2002). These regions exhibited unusually high densities of predicted binding sites for the early-active transcription factors Bicoid, Hunchback, Kruppel, Knirps and Caudal. In a follow-up study, the same group carried out an evolutionary analysis of these enhancers based on comparisons of the *D.melanogaster* and *D.pseudoobscura* genomes (Berman *et al*., 2004). All 37 candidates were conserved based on the alignment analysis in *D.pseudoobscura*, and 33 of them were experimentally tested; yet, only 15 enhancers drove expression along A/P axis whereas the other 18 candidates did not. This data is therefore an ideal dataset to evaluate our approach regarding its ability to predict functionality of enhancer candidates; those experimentally validated enhancer regions that drive expression patterns form a positive set, and the enhancers that do not drive expression are considered as negative set.

(b) As a large-scale dataset, we used a well-annotated collection of known *cis*-regulatory modules (CRMs) from the *Drosophila* genome that are deposited in the REDfly database (Gallo *et al*., 2006). Release 5 contained 728 CRMs recorded along with information on the gene expression pattern driven by each CRM. We considered 421 CRMs with a sequence length between 250 and 2000 bp for our analysis (this excludes short modules for which our approach mostly is not applicable). This dataset contains regulatory sequences from nearly all stages of embryogenesis and some post-embryonic stages. For instance, it includes CRMs that drive ectoderm expression (CNS, PNS, SNS, trachea and epidermis), or enhancers that drives part ectodermal (foregut, hindgut, salivary glands and Malpighian tubules) and endodermal (midgut) expression, and it contains enhancers for mesoderm and its major derivatives (visceral musculature, fat body, dorsal vessel and somatic musculature). There is no curated information about enhancer locations outside of *D.melanogaster*, and is thus used to evaluate whether our approach makes significant predictions across a wider range of expression patterns exceeding the well-annotated segmentation network enhancers above.

(c) Finally, in a recent study on the conservation of well-characterized *eve* enhancers between *Drosophila* and Sepsid genomes, it was shown that enhancers from distantly related genomes (>100 Mio years) can still produce identical expression patterns despite almost completely rearranged binding sites (Hare *et al*., 2008). Six sepsid species from three families which have well-characterized phylogenies were

considered for this analysis. Predicted orthologous enhancers in sepsid genomes were based on alignment methods such as BLASTZ and contained a small number of short (20–30 bp) sequences conserved between sepsids and *Drosophilids*. The expression patterns of *eve* genes in sepsid genomes were experimentally validated using *in situ* hybridization and confirmed the characteristic expression pattern of seven transverse stripes. We used this dataset to test whether our method is able to detect orthologous enhancers in sepsid genomes which is significantly more diverged from *D.melanogaster* than any *Drosophila* species, and for which standard alignment methods break down due to the significant divergence of linear sequence similarity.

With sufficiently large set of known enhancers, the parameters (i) word length, (ii) window size and (iii) mixed metric weighting can be automatically trained to achieve optimal results. In our case, the datasets with known orthologous fly enhancers are relatively small, and the larger sets generally lack information about orthologs. For all experiments, we therefore fixed the word length as 6 and the window size as 500, reflecting commonly made choices in the literature, and set the weighting parameter $a = 0.1$ to reflect our stronger interest in shared rather than differing word occurrences.

3 RESULTS

3.1 The alignment-free method correctly predicts enhancers that drive A/P pattern during *Drosophila* embryogenesis

To test its effectiveness, we first applied our method on a set of 33 annotated enhancer candidates for A/P patterning during *Drosophila* embryogenesis. According to alignment-based methods, all candidates were conserved in *D.pseudoobscura*; yet, experiments showed that only 15 of these candidates were indeed enhancers driving an A/P expression pattern, while 18 were non-functional. Our task was therefore first to identify the best location of these candidates in the related genomes, and then to computationally distinguish the real candidates from false ones using our alignment-free method. We first identified orthologous intergenic regions (cf. Section 2.1) from non-melanogaster genomes (*D.ananassae, D.pseudoobscura, D.willisoni* and *D.mojavensis, D.virilis, D.grimshawi*), in which the orthologous enhancers were expected to be located. We scanned a known enhancer from *D.melanogaster* against a moving window of fixed size (500 bp) along the corresponding orthologous intergenic region (with the mixed metric weight a set to 0.1; cf. Section 2.5). Since functional enhancers can be located on the forward or the reverse strand, we computed the mixed metric along both strands of the entire orthologous intergenic region. The window with the global maximum mixed metric score provided us with the information of the location of orthologous enhancer regions. As an example, Figure 1 shows the result of scanning the well-known enhancer Eve_stripe_3/7 against the orthologous intergenic region from *D.pseudoobscura*. It is evident that there is one clear maximum, at the location of the annotated enhancer in *D.pseudoobscura*, and that this maximum exceeds the significance threshold.

After all pair-wise comparisons between the reference species *D.melanogaster* and the 6 non-melanogaster genomes, we computed a weighted *P*-value from the corrected *P*-values from the pairwise scans, taking the evolutionary tree into account (Section 2.6). Our method was successful in separating functional from non-functional enhancer candidates: 14 of the 15 positive candidates received weighted *P*-values <0.05, with the remaining receiving a borderline

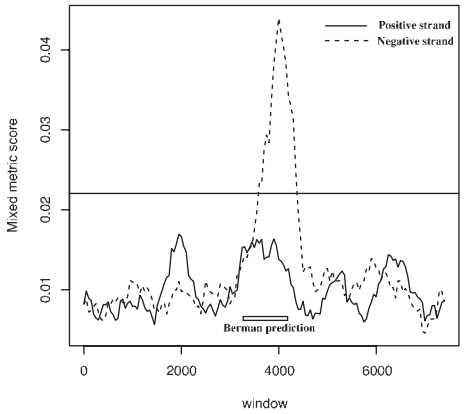

Fig. 1. Scanning the Eve_stripe_3/7 enhancer against the orthologous intergenic region from *D.pseudoobscura*. The *X*-axis indicates the window position, and the corresponding mixed metric score at each position is shown. Forward and reverse strand scanning results are shown with solid and dotted lines, respectively. The horizontal line represents an adjusted *P*-value threshold of 0.05 as obtained from the background distribution (Section 2.4). The orthologous enhancer region in *D.pseudoobscura* based on Berman's analysis is indicated as a bar.

value just above the threshold (Fig. 2a). In turn, 16 out of 18 negative candidates clearly exceeded significant *P*-values (all >0.1). For all positive candidates, the location of enhancers predicted by our approach overlapped with the locations as annotated in Berman *et al*. (2004).

The weighted *P*-values of most of the negative candidates indicates that those are in fact not conserved based on the Poisson similarity metric, which implies those candidates are non-functional regions. However, there were two candidates (CE8021 and CE8023) in which the orthologous regions were identified with significant weighted *P*-values in all six non-melanogaster species (cf. Fig. 2a). These candidates are as well conserved in distantly related species as the functional enhancers, and given the clear separation we observe for the remaining negative set, may therefore be functional in *Drosophila* embryo development, despite the fact that they could previously not be experimentally validated (Berman *et al*., 2004). As a first step toward validation, at least one of the genes flanking the candidate should be expressed in an A/P-like pattern similar to other validated enhancers. The candidate CE8023 is located in between the 'Dfd' and 'Ama' genes. The Atlas of Patterns of Gene Expression during *Drosophila* Embryogenesis (APOGEE; Tomancak *et al*., 2007) has *in situ* hybridization image data on *Dfd*, and its expression indeed shows patterning along the A/P axis (Fig. 3). While this does not confirm that the precise region we predicted is the functional one, it indicates that the *Dfd* genomic locus must contain an enhancer such as the one we predicted. The other candidate, CE8021, is located 7 kb upstream of the 'reaper' gene. Unfortunately, the expression pattern of *reaper* is not yet available in APOGEE.

3.2 Performance evaluation compared with alignment-based approaches

To put the performance of our alignment-free method in context, we compared it to general purpose alignment-based methods that equally do not require knowledge of binding site information. We performed the sequence comparison for all 33 predicted enhancers

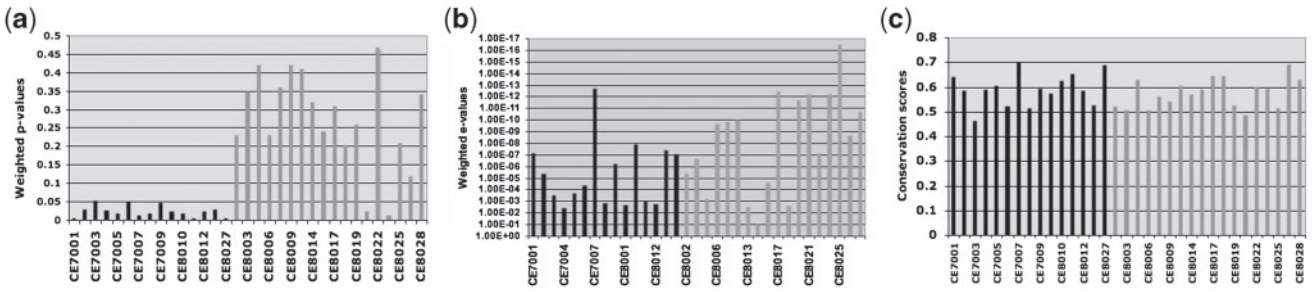

Fig. 2. Performance of different methods in detecting enhancers based on conservation across multiple species. (**a**) Alignment-free approach; (**b**) BLAST; (**c**) phastCons. The positive candidates are indicated in black bars and negatives are represented in gray.

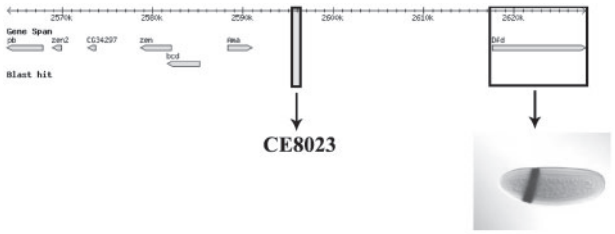

Fig. 3. Genomic region containing the putative enhancer CE8023 and the *Dfd* gene on chromosome 3R [image extracted from Flybase (Ashburner and Drysdale, 1994)]. Box in the right side and middle indicate the location of *Dfd* and CE8023. An image showing the expression pattern of *Dfd* taken from the *in situ* database is shown.

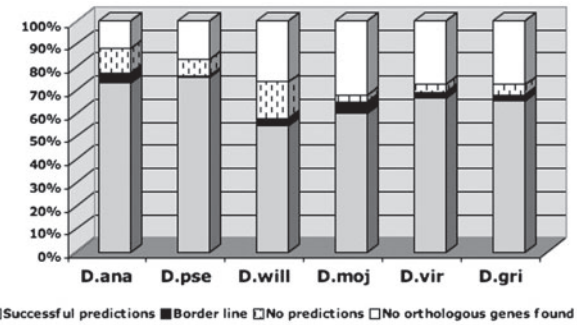

Fig. 4. Result of orthologous enhancer identification of REDfly candidates in pairwise scans. The successful predictions are shown in gray ($P<0.05$), borderline candidates ($0.05 \leq P < 0.1$) are indicated in black, dashed vertical line indicates regions without significant predictions and white shows candidates for which we could not assign orthologous flanking genes.

from *D.melanogaster* using BLAST (combined pairwise scores weighted as in our approach) or the average phastCons value from its track in the UCSC genome browser (Siepel and Haussler, 2004); the results are shown in Figure 2b and c, respectively. Compared with Figure 2a, it is clear that our alignment-free method performed significantly better than BLAST and phastCons, which largely failed to discriminate the positive and the negative enhancers. It therefore appears that the alignment-free metrics reflect the particular patterns of enhancers (shared word occurrences, but possibly in different order and/or orientation) more adequately than approaches to detect conserved regions in general, and which assume linearity over longer stretches of sequence.

3.3 Evaluation on additional datasets

To demonstrate that the applicability extended beyond the well-studied A/P patterning, we considered two additional datasets. We first evaluated a well-studied and experimentally validated set of conserved enhancers that control dorsal/ventral (D/V) patterning of *Drosophila* embryogenesis (Papatsenko and Levine, 2005). Results were very similar to the A/P set (see Supplementary Material for details), and our method successfully identified the location of nearly all enhancers in divergent genomes, including secondary or 'shadow enhancers' (Hong *et al*., 2008). Both A/P and D/V datasets concerned regulatory regions which define expression patterns in the early *Drosophila* blastoderm. To evaluate our approach on a larger set of regions with a broader range of expression patterns, we examined enhancers from a large collection of experimentally verified enhancers in *D.melanogaster* taken from the REDfly database. In total, 421 enhancers were considered; for some of the enhancers we could not identify orthologs of the flanking genes in a related species, and thus no syntenic intergenic region to analyze.

Using the alignment-free method, we first performed a systematic pairwise comparison between the reference genome (*D.melanogaster*) and the other non-melanogaster species (Fig. 4). For instance, for 73% of candidates we were able to identify a candidate orthologous enhancer with significant *P*-value in *D.ananassae*; 4% had borderline significance values between 0.05 and 0.1; for 12% of enhancers we were not able to identify the orthologous flanking genes of the corresponding enhancers; and we did not make any significant prediction for 11%. Overall, we reached the highest fraction of significant predictions in *D.pseudoobscura* (76%), and in more distantly related species (*D.mojavensis, D.virilis* and *D.grimshawi*) we made predictions for at least 60% of enhancers.

While the fraction of non-orthologous flanking genes scaled with the phylogenetic distance, the fraction of enhancers without significant prediction showed interesting trends; more distantly related genomes had smaller fractions of enhancers without predictions, indicating that the smaller fraction of identified syntenic intergenic regions may also be highly reliable. Not fitting into this overall pattern is *D.willistoni*, which shows a remarkably higher fraction of enhancers for which we could not predict orthologs. Combining all six species, we could make predictions with significant weighted *P*-value for 58% of enhancers, which suggests that our alignment-free method performs well in detecting

Table 1. Evaluation on the REDfly database

Annotation term	REDfly	Analyzed	Predicted
mapping1.blastoderm	77	44	36 (82%)
mapping1.cns	34	17	14 (82%)
mapping1.ectoderm	37	20	15 (75%)
mappling1.imaginal disc	47	28	21 (75%)
mapping1.pns	24	12	10 (83%)
mapping2.ectoderm	51	27	19 (70%)
mapping2.eye	18	10	5 (50%)
mapping2.imaginal disc	12	10	9 (90%)
mapping2.mesoderm	45	16	7 (44%)
mapping2.neuronal	54	26	21 (81%)
mapping2.wing	33	21	17 (81%)
mapping3.larva	69	34	27 (79%)

The first column shows the gene expression pattern, the second the number of CRMs in REDfly. The third column indicates the number of sequences that have identifiable orthologous search regions; the fourth shows the number of predictions with significant similarity (out of the sequences in Column 3). We show results for patterns with at least 10 analyzed CRMs.

candidate orthologous enhancers with a wide range of expression patterns. In a recent study on CRM prediction in *Drosophila* based on REDfly (Ivan *et al*., 2008), experimentally validated CRMs were grouped based on common gene expression annotation at different levels of abstraction, such as blastoderm (which contains both the A/P and the D/V datasets discussed above), mesoderm and neuronal. Breaking down our results by these different groups, Table 1 shows the fraction of enhancers for which we obtained significant weighted *P*-values. Significant predictions are made across a wide variety of patterns, with rates ranging from about 50% to >80%, which indicates that the Poisson metric is applicable to not just a particular group of enhancers. However, as information on the location of these regions is not available in other genomes, we cannot validate these predictions as we could above for the better studied smaller sets.

Finally, we evaluate a recent dataset of *eve* enhancers mapped in genomes of scavenger flies (sepsids) that are significantly more diverged from *D.melanogaster* than the other fly genomes used in our analyses so far. Although these enhancers share only short stretches of block sequence similarity with *D.melanogaster*, it has been experimentally shown that the sequences can recapitulate expression patterns similar to *Drosophila*—i.e. despite strong sequence divergence, the function appears to be conserved (Hare *et al*., 2008). We applied the scanning method on six sepsid genomes for which the orthologous intergenic regions, and the locations of the enhancers in them, are available. The genome sequences of sepsid species are currently not available and so we used background scores computed from *D.grimshawi* as approximation, which is more distantly related to the reference species.

Interestingly, for the enhancers eve_stripe_3+7 and eve_stripe_4+6, our method predicted the candidate orthologous enhancers in all six sepsid genomes with significant *P*-value and at the annotated locations (Table 2). This provides strong evidence that our method can identify conserved orthologous regulatory enhancers in genomes highly diverged from the reference genome, where alignment methods fail to detect any similarity. However, we failed to identify the orthologous enhancers for eve_stripe_2 and eve_MHE, with a sequence size in *D.melanogaster* of 392 and 318 bp, respectively; yet as the analysis above showed, we can identify the eve_stripe_2 in all *Drosophila* genomes (results are shown in the Supplementary Material). Sepsid intergenic regions, and the annotated enhancers in them, are larger compared with fruit flies, and the failure may be related to sequence rearrangements, which exceed our window size of 500 bp.

Table 2. Evolutionary conservation analysis on *eve* enhancers in Sepsid genomes

Sepsid genomes	Eve_stripe_3+7	Eve_stripe_4+6
S.punctum	0.009	0.0079
S.cynipsea	0.0086	0.007
D.sp.	0.0035	0.0094
T.superba	0.009	0.0107
T.minor	0.0073	0.0064
T.putris	0.006	0.0097

Similarity *P*-values of enhancers Eve_stripe_3+7 and Eve_stripe_4+6 are shown.

4 DISCUSSION

We presented an alignment-free method based on a Poisson metric to identify orthologous enhancers in distantly related species for a given known enhancer, without any knowledge of specific binding sites in them. Based on earlier metrics applied to enriched words in single species, we extended the metric to pairs of related genomes, and developed methods for significance estimation and combination of more than two genomes. A characteristic feature of enhancers is that they convey their function regardless of orientation relative to the gene they regulate. As described, we therefore calculate the similarity across both strands of an intergenic region. Alternatively, we also evaluated collapsing counts for words and their reverse complements, and we did not observe a significant difference. We validated the enhancer detection analysis on a set of enhancer candidates initially predicted by clustering of known TFBSs that act at very early stages of *Drosophila* development to define the A/P axis of the embryo. We were able to discriminate between positive candidates from negative candidates without enhancer function, and provided evidence for functionality for one of two cases which had previously not been validated.

Applying our approach to different *Drosophila* datasets demonstrated the flexibility of our alignment-free method. In cases in which we failed, the enhancers were shorter than 250 bp, suggesting that our alignment-free approach relies on a certain size to obtain discriminative evidence. We were furthermore able to make predictions for at least 58% of enhancers in all six species on a large and diverse enhancer set with functions beyond just early *Drosophila* development. Finally, the method was at least partially successful in identifying enhancers of the *eve* gene in non-*Drosophila* species (sepsids), supporting the notion that our approach can identify CRMs in alignment-free manner across a wide range of sequence divergence. A recent, concurrently developed approach was mostly applied to enhancers within the same species, but interestingly also detected the CRMs for eve_stripe_3+7 and eve_stripe_4+6 and, like us, failed to detect CRMs for eve_stripe_2 and eve_MHE (Leung and Eisen, 2009).

Different from some other non-alignment methods, our approach identifies conserved CRMs without prior knowledge of known TFBSs or enriched words, and all *k*-mers are equally involved in

scoring a candidate region. However, to reduce noise, it is easily possible to restrict the scoring to a subset of *k*-mers with known or predicted regulatory roles, such as TFBS consensus strings. This may enhance the detection success in cases where binding site density is relatively scarce, or the sequences considered relatively short. It may also help to increase the window size and deal with cases where rearrangements happen over large distances. There are several directions in which our approach can be extended. We plan to next study its application for *de novo* enhancer discovery, i.e. when the CRM information is totally unknown, by scoring all possible window pairs of two intergenic regions against each other. This would provide an unbiased computational approach for genome-wide regulatory region detection and would offer a complementary approach to ongoing efforts, such as the modENCODE project. This approach would be particularly interesting when analyzing more distantly related species, where orthologous enhancers are difficult to identify using available alignment-based methods.

5 CONCLUSION

We show that candidate orthologous enhancers in multiple *Drosophila* and non-*Drosophila* genomes can be identified by an unbiased alignment-free method which uses no information of predicted or known functional motifs. Our results show promise in the effort to characterize non-coding sequence conservation on a higher functional rather than the nucleotide level. Since our method showed encouraging results in a wide range of *Drosophila* enhancers with different functions and active at different life stages, it will be interesting to explore its application for the discovery of regulatory modules in other organisms with larger and more complex non-coding sequence spaces.

ACKNOWLEDGEMENTS

We thank David Corcoran for helpful comments, and Aaron Wise for evaluation of third-party software.

Funding: National Institutes of Health (R01 HG004065 to U.O.); the Human Frontier Science Program (RGY0084/2008 to P.T., U.O.).

Conflict of Interest: none declared.

REFERENCES

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

Ashburner,M. and Drysdale,R. (1994) FlyBase - the Drosophila genetic database. *Development*, **120**, 2077–2079.

Berman,B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Berman,B.P. *et al.* (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol*., **5**, R61.

Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.,* **12**, 739–749.

Chan,B.Y. and Kibler,D. (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics*, **6**, 262.

Cliften,P.F. *et al.* (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.

Corcoran,D.L. *et al.* (2005) Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. *Genome Res.*, **15**, 840–847.

Erives,A. and Levine,M. (2004) Coordinate enhancers share common organizational features in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **101**, 3851–3856.

Gallo,S.M. *et al.* (2006) REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*, **22**, 381–383.

Hardison,R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, **16**, 369–372.

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

Hong,J.W. *et al.* (2008) Shadow enhancers as a source of evolutionary novelty. *Science*, **321**, 1314.

Ivan,A. *et al.* (2008) Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol.,* **9**, R22.

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

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

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

Loots,G.G. *et al.* (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.

Ludwig,M.Z. *et al.* (1998) Functional analysis of eve strip 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, **125**, 949–958.

Ludwig,M.Z. *et al.* (2002) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, **403**, 564–567.

Ludwig,M.Z. (2002) Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.*, **12**, 634–639.

Markstein,M. and Levine,M. (2002) Decoding cis-regulatory DNAs in the Drosophila genome. *Curr. Opin. Genet. Dev.*, **12**, 601–606.

Nazina,A.G. and Papatsenko,D.A. (2003) Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics*, **4**, 65.

Papatsenko,D. and Levine,M. (2005) Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proc. Natl Acad. Sci. USA*, **102**, 4966–5971.

Park,P.J. *et al.* (2002) Comparing gene expression profiles in genes with similar promoter regions. *Bioinformatics*, **18**, 1576–1584.

Siepel,A. and Haussler,D. (2004) Combining phylogenetic and hidden markov models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.

Sosinsky,A. *et al.* (2007) Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. *Proc. Natl Acad. Sci. USA*, **104**, 6305–6310.

Tomancak,P. *et al.* (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **8**, R145.

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.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380

Wolff,C. *et al.* (1999) Structure and evolution of a pair-rule interaction element: runt regulatory sequences in D. melanogaster and D. virilis. *Mech. Dev.*, **80**, 87–99.