BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 16 2007, pages 2031–2037
doi:10.1093/bioinformatics/btm299

Sequence analysis

Finding *cis*-regulatory modules in *Drosophila* using phylogenetic hidden Markov models

Wendy S.W. Wong[1,3,*] and Rasmus Nielsen[2]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA, [2]Department of Biology and Center for Comparative Genomics, University of Copenhagen, Universitetsparken 15, 2100 Kbh Ø, Denmark and [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received on November 1, 2006; revised on May 28, 2007; accepted on May 29, 2007

Advance Access publication June 5, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Finding the regulatory modules for transcription factors binding is an important step in elucidating the complex molecular mechanisms underlying regulation of gene expression. There are numerous methods available for solving this problem, however, very few of them take advantage of the increasing availability of comparative genomic data.

Results: We develop a method for finding regulatory modules in Eukaryotic species using phylogenetic data. Using computer simulations and analysis of real data, we show that the use of phylogenetic hidden Markov model can lead to an increase in accuracy of prediction over methods that do not take advantage of the data from multiple species.

Availability: The new method is made accessible under GPL in a new publicly available JAVA program: EvoPromoter. It can be downloaded at http://sourceforge.net/projects/evopromoter/

Contact: sww8@cornell.edu

1 INTRODUCTION

The regulation of gene expression is one of the molecular processes of greatest scientific interest. For example, understanding gene regulation is fundamental for understanding developmental processes (e.g. (Berman *et al*., 2002; Crickmore and Mann, 2006; Kassis *et al*., 1989; Potter *et al*., 2000; Schroeder *et al*., 2004)). Much research has been conducted with *Drosophila* due to its relatively well-understood development process and the conservation of developmental genes between *Drosophila* and humans.

The initial step of gene expression is controlled by transcription regulation; it starts when transcription factors (TFs) bind to their corresponding transcription factor binding sites (TFBSs). TFBSs are typically 5–15 base pairs long and they tend to have degenerate sequences. In bioinformatics research, they are typically represented by position weight matrices (PWMs). These matrices are generated from aligned experimentally verified sequences. Due to their functional importance, detecting TFBSs in upstream regulatory regions has been an important research focus for decades (Berman, *et al*. 2002).

TFBSs tend to cluster together to form *cis*-regulatory modules (CRMs). Finding CRMs is an important first step into studying how gene expression regulation works. The existing approaches for identifying CRMs can be roughly classified into three approaches: (1) identification of regions in a genome with a significant number of TFBSs. (2) Identification of putative regulatory regions that are significantly more conserved than the nearby non-coding regions (phylogenetic footprinting). (3) Identification of the regulatory regions of genes that are regulated by the same set of TFs (Blanchette *et al*., 2006). Note that these approaches do not have to be carried out separately; they can be combined to enhance the accuracy and power of the algorithms.

The first approach can be further classified into two distinct methods. The first type is knowledge-based methods (Duret and Bucher, 1997), based on some prior knowledge of the characteristics of the TFBSs (e.g. the positional frequencies of each position in the TFBS, and/or which TFs work together, etc.). Most of the methods in this category make use of the previously known PWMs. For example, Frith *et al*. (2001) derived a hidden Markov Model (HMM) to specifically model the intra- and inter-CRM regions, as well as the CRM regions in a single sequence (Frith *et al*., 2001).

The second kind of methods in this category are the *ab initio* methods (Duret and Bucher, 1997). This kind of methods assume no prior knowledge of the TPBSs and only makes use of the content differences in the input sequences. A simple and elegant method was presented by Chan and Kibler (Chan and Kibler, 2005), where they train their program on the hexamer frequencies in CRMs and no-CRMs. They found that their method performs very well in finding CRMs in gap and pair-rule genes in *Drosophila* compared with other CRM predictive tools.

The phylogenetic footprinting method is also popular in finding regulatory elements that are conserved among species (Duret and Bucher, 1997; Loots *et al*., 2000; McGuire *et al*., 2000; Wang and Stormo, 2005). Recently, Pierstorff *et al*. (2006) developed a method that finds CRMs by identifying

*To whom correspondence should be addressed.

regions with an excess of specific local short ungapped aligned sequences (Pierstorff *et al*., 2006). The authors demonstrated that the method finds more CRMs than any other methods they compared to.

A few groups attempted to combine both approaches for better prediction accuracies. Sinha *et al*. (2003) developed a method that combines the benefits from both approaches (Sinha *et al*., 2003). Their method, *Stubb* uses an HMM to model the correlations between TFBSs, and at the same time uses conserved sequence information from multiple sequences to enhance the prediction. However, the method handles two sequences at a time and does not take the phylogenetic tree topology into account. Grad *et al*. (2004) proposed a method that first identifies potential CRMs using phylogenetic footprinting, and then use these potential hits to generate a model for searching CRMs in the genome. The biggest advantage of the method is that no previous knowledge is needed.

Here we extend the HMM to comparative data using a phylogenetic HMM (phylo-HMM) (Siepel and Haussler, 2004). By allowing the HMM to emit phylogenetic evolutionary models, instead of nucleotide frequencies, the HMM approach is extended for the use on comparative data. This method, therefore, rigorously combines the ('vertical') information used in phylogenetic footprinting with the ('horizontal') information used in HMMs for knowledge-based CRM prediction. The idea of combining phylogenetic information and HMM has been around for more than a decade (Felsenstein and Churchill, 1996; Yang, 1995). However, it was not until recently that the phylo-HMMs drew significant research attention, likely to be contributed by the large amount completed genome sequences. For instance, as of 21 September 2006, the Genomes OnLine Database has over 420 completed genome available online, and over 40 of them are eukaryotic genomes (http://www.genomesonline.org/, 2006). A few recent interesting applications include genome-wide search for evolutionary conserved elements in vertebrate, insect, worm and yeast (Siepel *et al*., 2005); annotation of viral genomes (McCauley and Hein, 2006); identification of alternatively spliced sites (Allen and Salzberg, 2006) and of homologous proteins (Qian and Goldstein, 2004).

2 METHODS

2.1 The hidden Markov model

The model used here (Fig. 1) is among the simplest imaginable models for TFBS and CRM prediction. There are two types of states in our model: the silent states and the emitting states. The silent states do not emit symbols while the emitting states emit one or more columns of symbols at a single event. The silent states are included for a clean representation of the HMM even though we could have constructed an equivalent model without them. The state space of the model consists of three silent states, namely, start, intra and end state. The emitting states include the background state and the TFBS states. The chain has to start with the start state and end with the end state. The emitting states are not accessible from each other and have to go through the non-emitting intra state. The background state emits a single column of nucleotides according to the background substitution matrix and the phylogenetic tree. Each TFBS state emits multiple columns of nucleotides at each instance. In this model, the length of each TFBS state equals the number of sites in the TFBS as indels are not allowed.

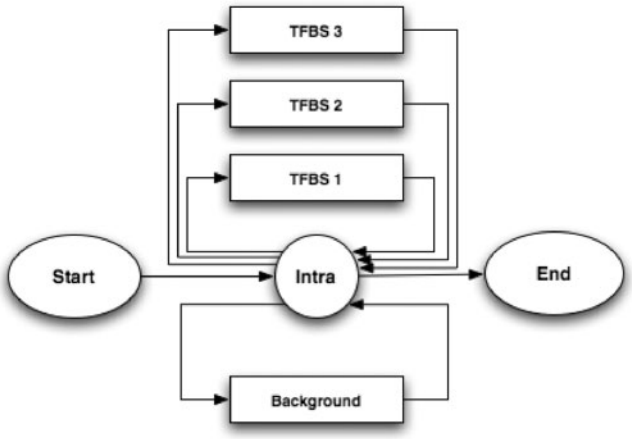

Fig. 1. The HMM. Rectangular boxes represents emitting states and the round shapes represent silent states. A solid line with an arrow represents a positive probability of going from one state to another.

More formally, our phylogenetic HMM is associated with a phylogenetic tree T, its state space consists of S_b, S_1, $S_2 \ldots S_n$, where S_b is the background state and the rest are the TFBS states, and $n =$ number of TFBSs of interest. It is also associated with background substitution matrix M_b, substitution matrices for each site in each TFBS state M_{11}, $M_{12} \ldots M_{1m}$, $M_{21} \ldots M_{2m} \ldots M_{k1} \ldots M_{km} \ldots M_{nm}$, where $k_m =$ number of sites in TFBS k, and the transition probabilities P_{ij}'s between the states i and j. Each site in each TFBS emits a column of nucleotides according to its own substitution matrix M_{k1} and the phylogenetic tree T. In practice, each site is a 'state' in the traditional HMM sense and the transition probabilities are set to 1 for consecutive sites in a TFBS.

2.2 Substitution models

Given the substitution matrix for state i, and the phylogenetic tree T, the likelihood of an observed column $P(X|\, S_i, T)$ can be calculated using Felsenstein's algorithm (Felsenstein, 1991). When there are one or more gaps in the column of the alignment, the likelihood is set to 1 for each of the letters at the gap.

The background substitution matrix can be represented by any nucleotide substitution model. Here, we use the HKY85 model (Hasegawa *et al*., 1985) as it is simple and yet captures several of the most important features on nucleotide evolution: uneven base composition and different rates of transition and transversion. The model parameters (base frequencies and the transition/transversion rate ratio) can be estimated from the data using various phylogenetic software [e.g. PAML (Yang, 1997), Phylip (Felsenstein, 2005), PAUP*(Swofford, 1998), etc.). To obtain the substitution matrix of each column in the motif of interest, we have adopted the approach used in Moses *et al*. (2004), where the Halpern and Bruno (HB) model (Halpern and Bruno, 1998) is used to convert the position-specific nucleotide frequencies in each column of the motif into its corresponding substitution matrix. Using the notation from (Moses *et al*., 2004), the substitution rate going from nucleotide a to nucleotide b in the ith column of the motif is given by:

$$R(i)_{ab} = \begin{cases} Q_{ab} \frac{\log\left(\frac{f_{ib}Q_{ba}}{f_{ia}Q_{ab}}\right)}{1-\frac{f_{ia}Q_{ab}}{f_{ib}Q_{ba}}} & \text{if } f_{ia}Q_{ab} \neq f_{ib}Q_{ba} \\ Q_{ab} & \text{if } f_{ia}Q_{ab} = f_{ib}Q_{ba} \end{cases} \quad (1)$$

Where Q is the background substitution matrix, i is position in the TFBS (runs from 1 to the length of the TFBS) and f_{ia} is the probability of observing nucleotide a in the ith column of the TFBS. This can be obtained from the TFBS PWM. Note that the new transition matrix R is also time reversible.

2.3 EvoPromoter

These evolutionary models were implemented in the program EvoPromoter. The program was implemented in Java™ 1.5 and uses BioJava libraries (http://www.biojava.org, 2006). It reads in the model specification in an XML file and is, therefore, very flexible allowing exploration of multiple different HMMs.

The workflow of EvoPromoter is shown in Figure 2. The user inputs two files: alignment file for the sequences of interest, and the XML file of parameters and model specification. The user specifies the phylogenetic tree, parameter values for the transition/transversion bias κ, the background base frequencies, the PWMs for each TFBS, the sliding window size and the step size. EvoPromoter is constructed to work as automated as possible; while it is possible to manually set the transition probabilities between each state in EvoPromoter, or train it using real data; the default option of EvoPromoter is to use the Baum–Welch algorithm (Baum, 1972) to self-train the transition probabilities in the HMM model. The self-training procedure terminates when the difference of log-likelihoods in two consecutive iterations is less than the threshold 10^{-5}. EvoPromoter then finds the Viterbi paths on the forward and backwards strand using the trained parameters. A sliding window is then scanned through the Viterbi paths with pre-specified window size and step size. The mean and SD of number of predicted TFBSs is calculated and those windows that have number of TFBSs greater than one SD from the mean are considered being within a CRM. Finally, the overlapping 'CRM' windows are concatenated and the result is returned.

2.4 Closely Related Models

Our model presented here is quite similar to the MCAST model (Bailey and Noble, 2003), except that we only have one silent state and one background state. On the other hand, MCAST has two silent states and two background states. It distinguishes background states that are between CRMs and within CRMs. The two models also differ in the calculation of the scores for each state. In our model, the scores are the log-likelihoods of the observed data given the evolutionary model of the state. In MCAST, the scores are calculated in two steps: first, the log-odds ratio of the TFBS state against the background state is calculated, denoted by s; second, the P-score is calculated by taking the negative value of the log-odds ratio of the P-value of s and a P-value threshold defined by the user. The P-value of s is calculated as the probability of obtaining a random sequence with a score as high, or higher, than the observed score. The calculation procedure is described in detail in Bailey and Gribskov (1998). The P-value threshold is fixed for all TFBS states. When the p-score is negative the state transition is not allowed. For comparison, we also ran EvoPromoter with the simplified MCAST HMM model, i.e. transitions between states are controlled by the transition probabilities only but not the P-scores.

MSCAN (Alkema *et al.*, 2004; Johansson *et al.*, 2003) employs a sliding window approach. The P-value of a particular TFBS hit is calculated using the PWM of the TFBS and also the background distribution in the window. It then calculates k-hits score in the window, which is the combined significance of all the hits in the window. The advantage of this approach is that the algorithm does not have to be trained.

MONKEY (Moses *et al.*, 2004) was the first that used the Halpern and Bruno (HB) model (Halpern and Bruno, 1998) to model the TFBSs. They calculated P-values of each potential TFBS being a true TFBS and outputs the ones with significant P-values. On the other hand, EvoPromoter finds the transition probabilities for each TFBS by self-training. With the simple HMM model we use, EvoPromoter predicts the exact same set of TFBSs with the set of transition probabilities when matching the significance level used in MONKEY.

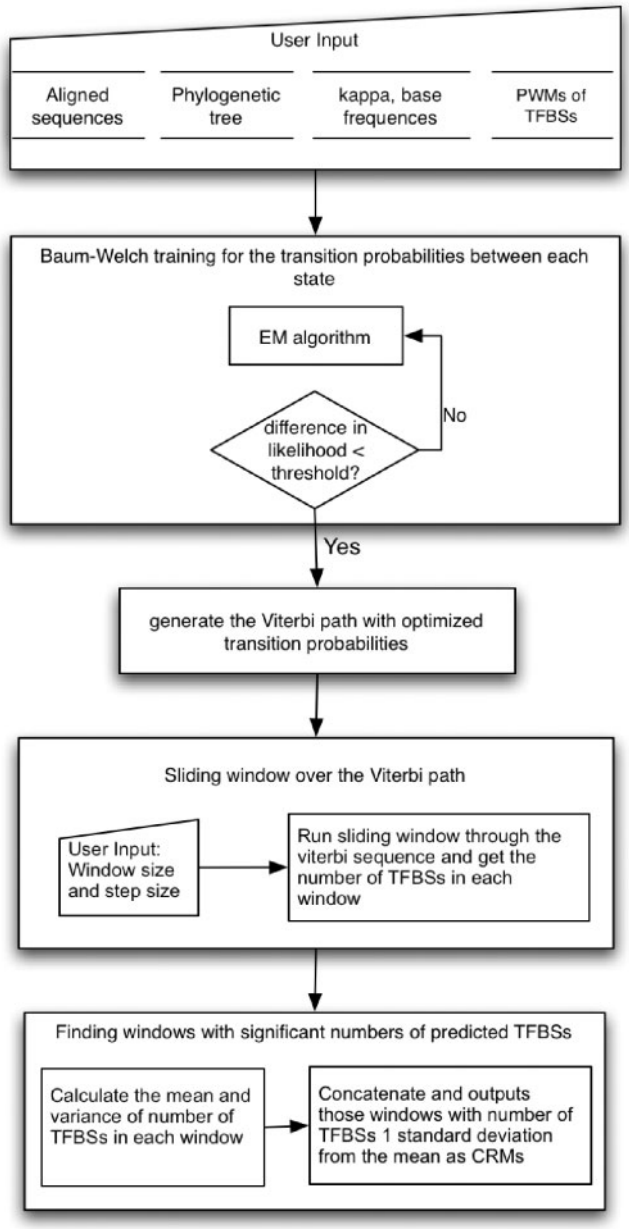

Fig. 2. The workflow of EvoPromoter illustrated. Parallelograms indicate user inputs and rectangles indicate a process in the program.

2.5 Simulation

We carried out a simulation study to verify our method. All simulations were carried out using CisEvolver (Pollard *et al.*, 2006b). The simulation procedure is described below. Table 1 shows the actual values used.

Table 1. Parameters used in the simulation study

Parameter(s)	Description	Value(s)
$\pi_a\ \pi_c\ \pi_G\ \pi_T$	Background frequencies of the nucleotides	$\pi_a = \pi_T = 0.3$
		$\pi_c\ \pi_G = 0.2$
κ	transition/transversion bias	2.0
T	Phylogenetic tree	(((dmel:0.61410, dana:0.6622):0.3237, dpse:0.59585):0.35745, (dvir:0.47455, dgri:0.51015):0.3434);
W	Size of CRM	500 bp
R	Indel rates	See (Pollard *et al.*, 2006b)
L	Total sequence length	10 kb
C	Number of CRMs in a sequence	3
d	Density of TFBSs	0.003

(1) Overall 10 sets of aligned nucleotide sequences, with ancestor sequences being L bp long were simulated according to the HKY85 model (Hasegawa *et al.*, 1985) with background frequencies $(\pi_a\ \pi_c\ \pi_G\ \pi_T)$, transition/transversion bias κ, indel rates R and the phylogenetic tree T obtained from (Pollard *et al.*, 2006a).

(2) For each set of the sequences, C aligned CRMs (ancestor sequences W bp long) were used to substitute part of the original sequences. The start positions of substitution were randomly selected according to a uniform distribution. If the difference of two consecutive start positions was shorter than the upstream CRM being inserted, another set was drawn until there are no overlapping CRMs.

(3) For each set of CRM, the nine TFBSs were simulated with the same density d and evolved with the Halpurn Bruno 1998 (HB98) model (Halpern and Bruno, 1998). Outside of the TFBS regions, sequences were simulated with the HKY85 model with background frequencies $(\pi_a\ \pi_c\ \pi_G\ \pi_T)$, transition/transversion bias κ, indel rates R and the phylogenetic tree T.

The data was analyzed with the simple EvoPromoter model with their true alignment, the alignment by DIALIGN + CHAOS, and with the MCAST model, with outputs refers to the locations in *Drosophila melanogaster* The *D.melanogaster* sequences were also analyzed by MSCAN and MCAST.

2.6 Performance measure

We measured the performance using the following two measures, CRM level and nucleotide level, defined in Pierstorff *et al.*, (2006).

(1) *Sensitivity and positive predictive values (PPVs) of the locations of the predicted CRMs at the CRM level*. As in Chan and Kibler (2005), a match is declared if a predicted CRM and a known CRM overlap by 50 bp. Following (Chan and Kibler, 2005), we calculated the number of CRMs recovered by each algorithm by counting the number of known CRMs that overlap with the predicted CRMs. The number of true positives (TP) is defined as the number of CRMs predicted that overlap with the known CRMs. The number of false negatives (FN) is the number of known CRMs with no overlap with predicted CRMs. Sensitivity is defined as the proportion of CRMs recovered over the total number of known CRMs [TP/(TP + FN)]. To calculate specificity, we needed the number of CRMs that are true negatives, which were not applicable in this case and hence not used. PPV is defined as TP/(TP + FP); which can be interpreted as the probability that the predicted CRM is a true one. Notice that we were using the known CRMs compiled by Schroeder *et al.*, (2004) to represent the true positive set. As shown in Pierstorff *et al.* (2006), the known set by Chan and Kibler (2005) does not cover all the true CRMs and hence the true PPV values were probably underestimated.

(2) *Sensitivity, specificity and PPVs of the locations of the predicted CRMs at the nucleotide level*. Here we considered all nucleotide positions in the sequence, and a match was declared if a nucleotide site was predicted to be in a CRM and it was indeed within a known or simulated CRM. TP was this case the number of bases predicted to be CRM that overlap with known CRMs. TN was the number of bases that were predicted to be not in a CRM which were truly not in a known CRM. FP was False Positives, i.e. the number of bases that are predicted to be in a CRM but do not overlap with any known CRM. The number of FN was the number of bases that were predicted to not be in a CRM but overlap with a known CRM. Sensitivity and PPV were defined earlier, while specificity was defined as TN/(TN + FP).

3 RESULTS

We compared the performance of EvoPromoter with MCAST (Bailey and Noble, 2003) and MSCAN (Alkema *et al.*, 2004), which both use the PWM information from known TFBSs. These two programs use more sophisticated models horizontally (along the sequence) but do not use the phylogenetic information.

3.1 Biological data

We chose to test our method using early developmental genes in *Drosophila* because the system is relatively well studied and understood. We looked at the same 16 genes that were used in (Chan and Kibler, 2005) because they have known CRMs. The upstream 10 kb region of *D.melanogaster* from these genes were downloaded using GBrowse in FlyBase (http://flybase.org/cgi-bin/gbrowse/, 2006). The corresponding regions in the other four species (*D.ananassae, D.pseudoobscura, D.virilis and D.grimshawi*) were obtained by performing a BLAT online search at the UCSC

Genome Bioinformatics Site (Kent, 2002). The sequences were then aligned by the CHAOS + DIALIGN web service (Brudno *et al.*, 2003). The phylogenetic relationship were obtained from Pollard *et al.* (2006a) and the phylogenetic tree of the five *Drosophila* species used is shown in Figure 3. Using the HKY85 model in the program *baseml* in the PAML package (Yang, 1997), we obtained the maximum likelihood estimates for tree lengths and the transition/transversion rate ratio for each set of aligned sequences. Unfortunately, for eight of the genes, the pairwise distances among the sequences were out of the PAML predefined range for its likelihood analysis. The level of divergence was hence too high for our analysis and these data were discarded. We were then left with eight genes to study; namely, Empty spiracles (ems), Even-skipped (eve),

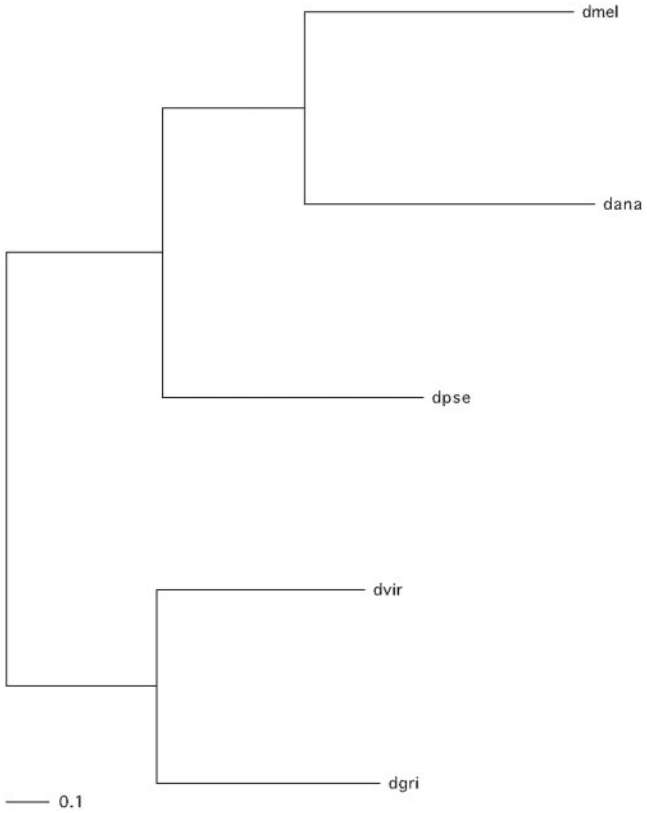

Fig. 3. The phylogenetic tree of the five *Drosophila* species used in the data analysis.

Fushi-Tarazu (ftz), Hairy (h), Hunchback (hb), Knirps (Kni), Paired (prd) and Runt (runt). Three of these genes do not have known CRMs in the upstream 10 kb region (hb, kni and prd). We included these three genes for calculating the sensitivity and PPVs of the three algorithms in finding approximate locations for the known CRMs, but they were not used in calculating these measures in terms of exact locations. The position frequency matrices of the nine TFBSs used were the same ones used in Schroeder *et al*. (2004), namely, Bicoid (Bcd), Hunchback (Hb), Caudal(Cad), the Torso-response element (TorRE), Stat92E (D-Stat), Kruppel (Kr), Knirps (Kni), Giant (Gt) and Tailless (Tll).

The performance of EvoPromoter, MCAST and MSCAN are summarized in Tables 2 and 3. Table 2 shows the sensitivity and PPVs of the three algorithms in discovering locations of the known CRMs at the CRM level. The same comparison was also done in Chan and Kibler (2005). The performances of the three algorithms are quite similar although EvoPromoter has a slightly higher sensitivity in discovering known CRMs and has a slightly higher PPV. In terms of finding the locations of the CRMs at the nucleotide level, EvoPromoter's performance varied in the five genes but in general is a bit lower than the other two methods. It indicates that EvoPromoter is perhaps not as good in picking out the exact boundary of the CRMs and it could be due to errors in alignment of the genes.

Table 2. Comparison of the predictive power of EvoPromoter, MCAST and MSCAN on finding CRM locations at the CRM level in upstream region of *Drosophila* development genes, measured by sensitivity and PPV

	CRMs recovered	Number of CRMs	Sensitivity	TPs	Number of CRMs predicted	PPV
Evopromoter	11	15	73.33%	9	34	26.47%
MCAST	10	15	66.67%	8	66	12.12%
MSCAN	8	15	53.33%	7	30	23.33%

Table 3. Comparison of the predictive power of EvoPromoter, MCAST and MSCAN for finding CRM locations at the nucleotide level in upstream region of *Drosophila* developmental genes, as measured by sensitivity, specificity and PPV

Gene	CRMs	Measure	EvoPromoter	MCAST	MSCAN
ems	2	Sensitivity	37.53%	35.47%	30.81%
		Specificity	45.11%	68.27%	93.46%
		PPV	11.13%	16.99%	46.31%
eve	2	Sensitivity	21.93%	81.60%	58.49%
		Specificity	65.39%	73.72%	93.17%
		PPV	8.45%	31.15%	55.52%
ftz	3	Sensitivity	36.58%	9.75%	0.00%
		Specificity	69.76%	91.19%	97.31%
		PPV	21.27%	19.82%	0.00%
h	4	Sensitivity	85.09%	92.15%	55.53%
		Specificity	75.66%	72.20%	83.20%
		PPV	60.79%	59.51%	59.44%
run	3	Sensitivity	19.30%	39.00%	33.42%
		Specificity	68.18%	55.76%	69.56%
		PPV	50.25%	59.48%	64.64%

Table 4. Comparison of the predictive power of EvoPromoter, MCAST and MSCAN on finding CRM locations at the CRM level in simulated data, measured by sensitivity and PPV

Algorithm	CRMs recovered	CRMs	Sensitivity	TP	CRMs predicted	PPV
EvoPromoter[1]	28	30	93.33%	19	20	95.00%
EvoPromoter[2]	20	30	66.67%	17	26	65.39%
EvoPromoter[3]	29	30	96.67%	22	28	78.57%
MCAST	17	30	56.67%	15	102	14.71%
MSCAN	23	30	76.67%	21	31	67.74%

[1]The correct simulated alignment was used.
[2]The CHAOS + DIALIGN alignments on the simulated data were used.
[3]EvoPromoter with the MCAST model was used.

3.2 Simulated data

The analysis of the 10 simulated data sets is summarized in Tables 4 and 5. When the correct alignment was used, at the CRM level, EvoPromoter was able to detect 28 out of

Table 5. Comparison of the predictive power of EvoPromoter, MCAST and MSCAN on finding CRM locations at the nucleotide level in simulated data, measured by sensitivity, specificity and PPV

Algorithm	TP	FP	TN	FN	Sensitivity	Specificity	PPV
EvoPromoter[1]	15992	10229	76945	1779	89.99%	88.27%	60.99%
EvoPromoter[2]	10654	17736	69438	7117	60.00%	79.65%	37.53%
EvoPromoter[3]	17273	10038	77136	498	97.20%	88.49%	63.25%
MCAST	11663	12717	74457	6108	65.63%	85.41%	47.84%
MSCAN	9241	3609	83565	8530	52.00%	95.86%	71.91%

[1]The correct simulated alignment was used.
[2]The CHAOS + DIALIGN alignments on the simulated data were used.
[3]EvoPromoter with the MCAST model was used.

the 30 implanted CRMs (high sensitivity) and only has one falsely identified CRM (high specificity). The number of CRMs predicted by EvoPromoter is less than the number of true CRMs. At a closer look, it was due to some of the simulated CRMs being right next to each other and EvoPromoter combining the two as one big CRM. At the nucleotide level, EvoPromoter also achieved much higher sensitivity than the MCAST and MSCAN, and comparable specificity and PPV, showing good predictive power when comparative data were used. We also showed that the performance of EvoPromoter is enhanced when a more complicated HMM model was used.

On the other hand, when we used the DIALIGN-CHAOS (Brudno *et al*., 2003) alignments we obtained using the default options of the simulated data, the performance for EvoPromoter significantly decreased, indicating that alignments play an important role on how well the software performs.

4 DISCUSSION

Our study showed that phylogenetic information could be used to improve the prediction of CRMs. As a proof of concept, we implemented a simple HMM across the aligned sequences, which is far simpler than the ones used in MCAST(Bailey and Noble, 2003), yet it still achieved a better performance. Note that the horizontal HMM (across the sequence) can be easily modified by altering the Extensible Markup Language (XML, http://www.w3.org/XML/) specification file. More complex models can, therefore, quite easily be implemented, hereby increasing the advantage of the phylo-HMM even further. The specifics of the implementation will depend on the species analyzed and are not pursued further here.

Another advantage of EvoPromoter is that the system is self-trained. Hence it does not depend on training data. This approach offers two advantages: first, the system is not biased towards the structure of the training data; second, it is directly applicable to species where no previous data on CRMs is available.

A disadvantage of the method is that it relies on reliable alignments. In cases where sequence divergence is so high that alignments are unreliable, the method is inapplicable. Not all data can be analyzed in the context of phylogenetic models, but in the presence of the rapid genomic sequencing, many interesting closely related species (e.g. *Drosophila* species and mammalian species) will show levels of divergence where phylo-HMM analysis is relevant.

As we can see from the analysis of the simulated data, EvoPromoter's performance is extremely good when the underlining assumptions are satisfied and the alignment is perfect. In real life, the alignment in the non-coding region may often be problematic. Recently Pollard *et al*. (2006b) performed a simulation study using the same HKY85 model as the evolutionary model for the background genomic sequence and the Halpern Bruno 1998 (HB98) model for the TFBSs. They also incorporated indels in their simulation program. Their results imply that, with standard alignment programs, at the evolutionary distance similar to *D.melanogaster* and *D.pseudoobscura*, only ~40% of true conserved binding sites are overlapping in alignments. Therefore, when real data are used, it is no surprise that the phylo-HMM method did not perform much better than the methods based on a single sequence. However, much recent progress has been done on the alignment of the non-coding regions. A recent study by Dewey *et al*. developed an algorithm for more reliable alignments for genomic data (Dewey *et al*., 2006). Using the alignments produced by their new algorithm, they found that the conservation of the *cis*-regulatory elements between the two Drosophila species (*D.melanogaster and D.pseudoobscura*) is greater than previous thought. We believe that our algorithm would benefit significantly from higher quality genomic data and more reliable alignment algorithms. Methods that incorporate alignment uncertainty, and evolutionary change of TFBSs/CRMs into CRM prediction in a phylogenetic framework would be ideal. However, such methods are currently not computationally tractable.

ACKNOWLEDGEMENTS

We like to thank Alan Moses for the discussion on using the Halpern-Bruno model. We are grateful to the two anonymous reviewers who helped in significantly improving the manuscript. This research was conducted using the resources of Coherent Logic Ltd, London, United Kingdom (http://www.coherentlogic.com/) and the resources at The Wellcome Trust Sanger Institute. This work was supported by funding from the Wellcome Trust and grants to R.N. from the Danish FNU.

Conflict of Interest: none declared.

REFERENCES

Alkema,W.B. *et al*. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res*., **32**, W195–W198.

Allen,J.E. and Salzberg,S.L. (2006) A phylogenetic generalized hidden Markov model for predicting alternatively spliced exons. *Algorithms Mol. Biol*., **1**, 14.

Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19** (Suppl. 2), II16–II25.

Baum,L.E. (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

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.

Blanchette,M. *et al*. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*., **16**, 656–668.

Brudno,M. *et al*. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.

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

Crickmore,M.A. and Mann,R.S. (2006) Hox control of organ size by regulation of morphogen production and mobility. *Science*, **313**, 63–68.

Dewey,C.N. *et al*. (2006) Parametric alignment of Drosophila genomes. *PLoS Comput. Biol*., **2**, e73.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol*., **7**, 399–406.

Felsenstein,J. (1991) Counting phylogenetic invariants in some simple cases. *J. Theor. Biol*., **152**, 357–376.

Felsenstein,J. (2005) PHYLIP (Phylogeny Inference Package). Distributed by the author, Seattle, Department of Genome Sciences, University of Washington, Seattle.

Felsenstein,J. and Churchill,G.A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol*., **13**, 93–104.

Frith,M.C. *et al*. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.

Grad,Y.H. *et al*. (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics*, **20**, 2738–2750.

Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol*., **15**, 910–917.

Hasegawa,M. *et al*. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol*., **22**, 160–174.

http://flybase.org/cgi-bin/gbrowse/ (2006) FlyBase Genome Browser. http://flybase.org/cgi-bin/gbrowse/.

http://www.biojava.org (2006) BioJava. http://www.biojava.org.

http://www.genomesonline.org/ (2006) Genomes OnLine Database. http://www.genomesonline.org/.

Johansson,O. *et al*. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19** (Suppl. 1), i169—i176.

Kassis,J.A. *et al*. (1989) Evolutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the Drosophila segmentation gene engrailed. *Mol. Cell. Biol*., **9**, 4304–4311.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.

Loots,G.G. *et al*. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.

McCauley,S. and Hein,J. (2006) Using hidden Markov models and observed evolution to annotate viral genomes. *Bioinformatics*, **22**, 1308–1316.

McGuire,A.M. *et al*. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*., **10**, 744–757.

Moses,A.M. *et al*. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*., **5**, R98.

Pierstorff,N. *et al*. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, **22**, 2858–2864.

Pollard,D.A. *et al*. (2006) Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genet*., **2**, e173.

Pollard,D.A. *et al*. (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**, 376.

Potter,C.J. *et al*. (2000) Drosophila in cancer research. An expanding role. *Trends Genet*, **16**, 33–39.

Qian,B. and Goldstein,R.A. (2004) Performance of an iterated T-HMM for homology detection. *Bioinformatics*, **20**, 2175–2180.

Schroeder,M.D. *et al*. (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*., **2**, E271.

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

Siepel,A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.

Sinha,S. *et al*. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19** (Suppl. 1), i292–i301.

Swofford,D. (1998) PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods). Sinauer Associates, Sunderland.

Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*., **102**, 17400–17405.

Yang,Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics*, **139**, 993–1005.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.