An Empirical Comparison of Tools for Phylogenetic Footprinting

Mathieu Blanchette*
Samson Kwong
Martin Tompa

Department of Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195-2350 U.S.A.
{blanchem,skk,tompa}@cs.washington.edu

Abstract

Phylogenetic footprinting is an increasingly popular comparative genomics method for detecting regulatory elements in DNA sequences. With the profusion of possible methods to use for phylogenetic footprinting, the biologist needs some guidance to choose the most appropriate tool. We present methods for comparing tools on phylogenetic footprinting data. More specifically, we discuss two different classes of comparative experiments: those on simulated data and those on real orthologous promoter regions. We then report the results of a series of such empirical comparisons. The tools compared are the alignment-based methods using ClustalW and Dialign, and the motif-finding programs MEME and FootPrinter. Our results show that methods taking the species' phylogenetic relationships into consideration obtain better accuracy.

1 Phylogenetic Footprinting

Phylogenetic footprinting [27] is a method for identifying regulatory elements, given orthologous regulatory regions from multiple species. It is based on the simple premise that functional elements tend to evolve at a slower rate than nonfunctional elements, due to selective pressure. If the given orthologous regulatory regions contain unusually well conserved subsequences, it is a reasonable conjecture that these conserved subsequences have some regulatory function. With the wide variety of prokaryotic and eukaryotic genomes being partially or fully sequenced, phylogenetic footprinting is becoming the approach of choice for computational detection of regulatory elements. This paper presents a methodology to compare the accuracy of different tools for phylogenetic footprinting, and reports the results of such a comparison.

There have been a number of phylogenetic footprinting studies, using an almost equal number of different computational tools and methodologies. The most commonly used methods are based on some global multiple alignment tool such as ClustalW [12], rVISTA [17], MultiPipMaker [10], or Dialign [22]. Such methods begin by computing a global multiple alignment of the orthologous sequences. If the alignment is correct, corresponding regulatory elements from different sequences will be aligned. It is then possible to parse the alignment from left to right and identify short substrings that appear unusually well conserved. Criteria for what should be considered a "conserved substring" vary from study to study. Phylogenetic footprinting studies based on these ideas include Cliften *et al.* [5], Dubchak *et al.* [6], Gumucio *et al.* [9], Jiao *et al.* [14], Krawczak *et al.* [15], Loots *et al.* [18], Manen *et al.* [19] and Tagle *et al.* [27].

Stojanovic *et al.* [25] empirically compared the accuracy of some criteria for assessing conserved substrings in a multiple alignment. They used a multiple alignment program called yama2 and compared five methods for identifying conserved blocks in the alignments produced. They showed that criteria based on relative entropy and on phylogenetic parsimony obtained slightly better results on four sets of orthologous sequences with known regulatory sites. These two approaches are investigated further in this study.

To understand why methods based on global multiple alignment may not always succeed, consider the typical lengths of the sequences involved. Regulatory

*Corresponding author. Current address: Baskin Engineering Center for Biomolecular Science & Engineering, University of California, Santa Cruz, CA 95064, blanchem@soe.ucsc.edu

elements tend to be quite short (5-20 bp long) relative to the entire regulatory region in which we search for them (a 1000 bp promoter region being typical). Given these relative lengths, if the species are somewhat diverged it is possible that the noise of the diverged nonfunctional background will overcome the short conserved signal. The result is that the optimal global alignment may well not align the short regulatory elements together, in which case they will go undetected.

To avoid this problem, a second approach to phylogenetic footprinting has been proposed more recently. This approach uses tools developed for the discovery of general sequence motifs such as MEME [1], Gibbs sampling [16], Consensus [11], and AlignACE [13]. Such tools do not attempt to align the complete input sequences, but instead search directly for short conserved subsequences. These methods assume the input sequences to be independent, ignoring the phylogenetic relationships among them. This can be problematic, for example in data sets containing a mixture of some closely related species and some distant ones. If the phylogeny underlying the data is ignored, similar sequences from the set of closely related species will have an unduly high weight in the choice of motifs reported. Phylogenetic footprinting studies based on such motif finders include Cliften *et al.* [5], McCue *et al.* [20], and McGuire *et al.* [21].

To address the drawbacks of global multiple alignment schemes and general motif discovery schemes, we introduced a tool called FootPrinter designed specifically for finding regulatory elements by phylogenetic footprinting (Blanchette *et al.* [2], Blanchette and Tompa [4]). FootPrinter is a motif-finding program making use of available phylogenetic information to evaluate motif conservation more accurately. Specifically, FootPrinter reports all sets of subsequences of the input sequences that have a small parsimony score with respect to a given phylogenetic tree relating the sequences.

With the profusion of possible methods to use for phylogenetic footprinting, the biologist needs some guidance to choose the most appropriate tool. The purpose of this paper is to present a method for comparing tools on phylogenetic footprinting data, and to report the results of a series of such empirical comparisons that we performed. (See Sinha and Tompa [24] for an analogous performance comparison among motif discovery programs based on statistical overrepresentation.) We start by describing in more detail the methods compared. The accuracy of these approaches is then analyzed using both simulated data (Section 3) and real orthologous promoter regions (Section 4).

Table 1: Summary of the strategy used by each method evaluated in this paper.

	Parsimony	**Rel. Entropy**	**Other**
Alignment	ClustalPars DialignPars	ClustalEntr DialignEntr	Dialign
Motif	FootPrinter	MEME	

2 Phylogenetic Footprinting Tools Compared

We compare the accuracy of several approaches for phylogenetic footprinting. These approaches fall into two categories: alignment-based and motif-finding-based approaches. Each category is subdivided into phylogenetic and non-phylogenetic approaches. Table 1 summarizes the strategy adopted by each of the seven methods evaluated.

We considered two global multiple alignment programs: ClustalW and Dialign. ClustalW is a true global alignment program, trying to optimize the overall alignment conservation, while Dialign starts by finding highly conserved segments that are eventually pieced together into a global alignment. The pairwise alignment programs VISTA and PipMaker have also been used for phylogenetic footprinting but were left out of this study because they do not output multiple alignments. Once a multiple alignment is produced by ClustalW or Dialign, one needs to identify the best conserved regions of the alignment. There are several ways to do so. Dialign outputs, together with the multiple alignment, a conservation profile indicating how conserved each column of the alignment is, so we decided that the prediction made by Dialign would be based on this profile. For example, if we are looking for one motif of ten nucleotides, the region of ten consecutive columns of the alignment having the highest average conservation profile will be reported. ClustalW does not output such a conservation profile, so we implemented our own conservation measures. First, we compute the parsimony score on the phylogenetic tree T for each column. (The parsimony score [8] of a set of orthologous sequences is the least number of substitutions, performed along the branches of T, needed to explain these sequences.) The gapless region of ten consecutive columns with the least parsimony score is reported. We call this method ClustalParsimony. We also used this parsimony criterion on the alignment produced by Dialign, yielding a third prediction method we call DialignParsimony. The use of parsi-

mony scores relies on the knowledge of the topology of T, which we will assume is given to us with the input sequences.

Alternatively, we considered another conservation measure popular in sequence analysis, the relative entropy. The relative entropy [11] of a column of the alignment measures the difference between the distribution of the four types of nucleotides observed in the column and the background nucleotide distribution. This method ignores the phylogenetic relationships among species, but has been shown very effective for regulatory element detection in contexts other than phylogenetic footprinting. Again, the region of ten consecutive gapless column with the highest total relative entropy is reported, yielding methods we call ClustalEntropy and DialignEntropy.

We studied two motif-finding programs: MEME, which searches for motifs with high relative entropy, and FootPrinter, which identifies motifs with low parsimony score. For all experiments reported in this paper, we asked each program to report its best motif of length ten. Two other motif-finding programs, AlignACE and Consensus, were also initially considered, but obtained poor results because their input parameters do not allow the user to ask for at most one motif per sequence.

Notice that the parsimony-based methods (FootPrinter, ClustalParsimony, and DialignParsimony) are given a little more information than the others, namely the topology of the tree T that generated the input sequences. One of the goals of this paper is to measure how much accuracy can be gained by using this extra information.

3 Comparisons on Simulated Data

3.1 Methodology

A key feature of the data used for phylogenetic footprinting is that the sequences considered are not independent from each other but rather are related through a phylogenetic tree. Our simulated data sets also have this property. We generated simulated data sets and compared the phylogenetic footprinting programs by the following procedure:

1. Choose a tree topology T, together with the length $\lambda(e)$ of each branch e of the tree.
2. Choose the motif relative mutation rate $R(e) < 1$ for each branch e.
3. Repeat the following steps for 200 trials:
   (a) Choose a random DNA sequence S of length 1000 for the root of T (each nucleotide chosen uniformly and independently) .
   (b) Choose a random substring of length 10 of S to be the *planted motif.*
   (c) Simulate evolution on T to produce DNA sequences at each of its nodes, keeping track of the planted motif position in each sequence. For a branch e of length $\lambda(e)$, the expected number of substitutions per site outside the planted motif is $\lambda(e)/100$, while inside the planted motif it is $R(e)\cdot\lambda(e)/100$. Insertions and deletions also occur outside, but not inside, the planted motif, at 40% of the rate of substitutions. This sequence evolution simulation is carried out using the program Rose [26].
   (d) Discard the sequences at the internal nodes. Run each of the programs on the leaf sequences, comparing the predicted motif positions to the known planted motif positions. The exact parameters for each program are given on our web site [3].

We have attempted to choose the simulation parameters to reflect the types of data sets encountered in practice on real biological sequences.

For each of the n leaf sequences S_i, each program predicts a set X_i of ten consecutive positions. Let P_i be the set of ten positions containing the planted motif. The *success rate* of a prediction is

$$\frac{\sum_{i=1}^{n} |X_i \cap P_i|}{\sum_{i=1}^{n} |X_i \cup P_i|}.$$

This measure was proposed by Pevzner and Sze [23] in a similar context. A perfect prediction obtains a score of one, while a completely incorrect one obtains a score of zero.

3.2 Results and Discussion

We ran simulations for a wide choice of tree topologies, number of leaves n, branch lengths $\lambda(e)$, and relative mutation rates $R(e)$. Only a subset of those results are reported here. Varying the tree topology and number of leaves did not qualitatively affect the results, so we report only results for simulations where T is a perfectly balanced binary tree with 8 leaves (with the exception of Figure 2 and Figure 4(b) below).

Failure to identify the planted motif can have two causes. First, it is possible that some region other than the planted motif has been better conserved by chance than the planted motif. This may arise if the sequences considered are very closely related and regions outside the planted motif have not had time to diverge sufficiently, or if the sequences considered were

very long. Such motifs will be very difficult to discern with accuracy. The second cause of erroneous prediction, the one in which we are the most interested, is that although the planted motif was in theory detectable, the algorithm simply failed to determine the correct region. This may be due to the fact that the scoring method used to evaluate candidate motifs is inaccurate, or because the program failed to identify the optimal solution to the problem it is solving. We will see examples of both errors in this section.

Figure 1(a) gives the average success rates obtained from 200 simulations on a tree where all branches have length $\lambda = 8$. As expected, the larger the planted motif's mutation rate R becomes, the harder it is to distinguish it from the surrounding sequence, and thus the accuracy of any method decreases. However, even when $R = 0$ (i.e. the planted motif is perfectly conserved), it sometimes happens that some other region is chosen because it is equally conserved. It is interesting to note that all three parsimony-based methods have a slight but quite consistent advantage over the entropy-based methods, and in particular over MEME. (Error bars on estimates of the average success rates are omitted for clarity, but the standard deviation for each estimate is always less than 0.03.) For branch lengths $\lambda = 8$, sibling sequences are approximately 80% identical, while the most remotely related pairs of sequences are about 50% identical (slightly more diverged than non-functional regions of human and mouse), which makes finding a correct multiple alignment relatively easy. This explains why ClustalParsimony and DialignParsimony appear to have a slight advantage over FootPrinter for large values of R. Indeed, although all three methods use the same motif evaluation criterion, FootPrinter faces the risk of choosing substrings that are not truly orthologous, resulting in incorrect predictions, because it is not aided by alignment outside its motif. As will be the case in all our experiments on simulated data, predictions based on the original Dialign conservation profiles are much less accurate than those of the four other methods. It should be noted though that the conservation profiles output by Dialign were not designed for detecting regulatory elements.

Figure 1(b) reports results for the same tree topology, but with the 7 branches in the left subtree having length 8 and the 7 branches of the right subtree having length 0.8. This is a situation where methods that ignore the phylogenetic relationships among the input sequences are at a greater disadvantage. Indeed, for these parameters, MEME's success rate is quite consistently 20% less than that of the three other methods making use of phylogenetic information.

To ensure that the trends observed in Figure 1 apply to other tree topologies and branch lengths, we repeated the experiment using trees inferred from some of the biological data sets used in Section 4. As an example, Figure 2 depicts the tree inferred from the beta-globin data sets. It contains 12 species, with groups of closely related species (e.g. the four old-world primates) and other very distant species (e.g. chicken and fishes). Figure 2b shows the success rates obtained when this tree is used to simulate sequences, varying the motif's mutation rate. It is interesting to notice that all methods (except Dialign) maintain a very high accuracy even for large relative mutation rates $R(e)$. This is mostly due to the fact that this simulation uses more sequences than those for Figure 1, and also because the sequences are more diverged, both of which help to distinguish the planted motif from the background. For low rates, FootPrinter has a small advantage over alignment-based methods, probably because in a few cases there are errors in the alignment produced by ClustalW or Dialign. However, the accuracy of FootPrinter decreases quickly as $R(e)$ becomes large. This phenomenon is also observed in Figure 3 and we will defer its explanation to that discussion. Finally, we observe that for large $R(e)$, alignment-based methods clearly outperform both motif-finding approaches, despite the occasional alignment errors, which indicates that the context of the alignment is very useful to identify subtle motifs.

Figure 3 displays success rates as a function of the length of the branches of the tree using equal length for all branches, with a relative mutation rate $R(e) = 0.1$ for all branches. When the branches are very short ($\lambda \leq 3$), no program is able to identify the planted motifs because the background sequences have not had sufficient time to diverge. As the branch lengths are slightly increased ($4 \leq \lambda \leq 12$), motif detection becomes increasingly accurate. Sequences are still relatively closely related and thus easy to align, yielding good scores for all alignment methods. Relative entropy methods are at a slight disadvantage compared to phylogenetic methods.

When the branch length is further increased ($\lambda \geq 16$), the sequences become more difficult to align, and errors in alignment cause the accuracy of alignment-based methods to drop sharply. We are now entering MEME's territory, where the input sequences are nearly unrelated, and where any region other than the planted motif is so poorly conserved that the planted motif, once identified, should clearly stand out, no

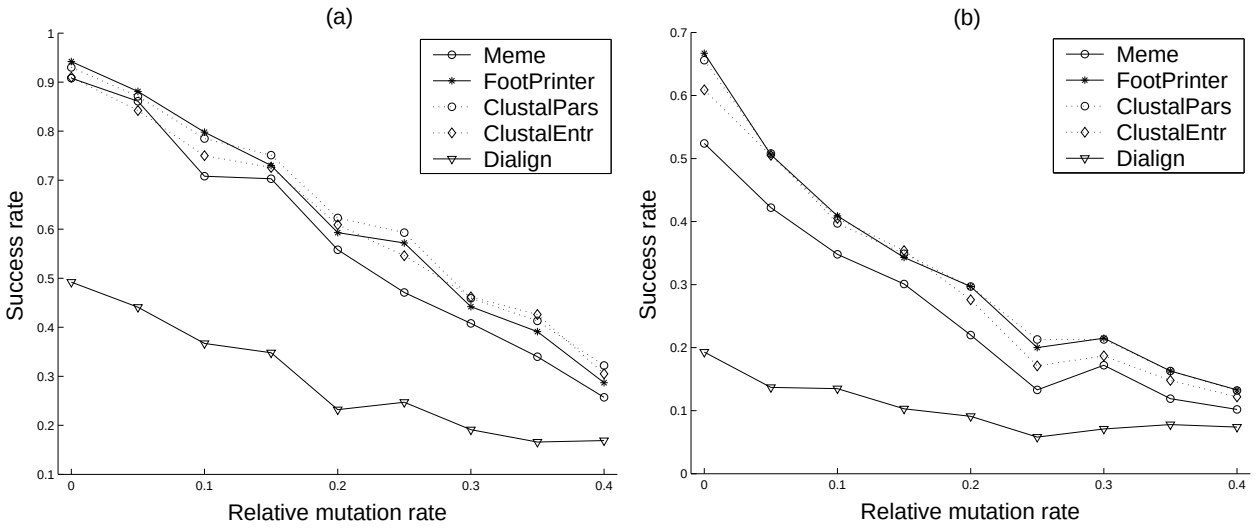

Figure 1: Success rates on the balanced binary tree with 8 leaves. In (a), all branches have length $\lambda(e) = 8$, whereas in (b) all branches of the left subtree have length 8 and those of the right subtree have length 0.8. For all figures in this paper, each point represents the average success rate over 200 trials. Error bars on estimates of the average success rates are omitted for clarity, but the standard deviation for each estimate is always less than 0.03. The results for DialignParsimony and DialignEntropy are not shown as they are very similar to those of ClustalParsimony and ClustalEntropy, respectively.

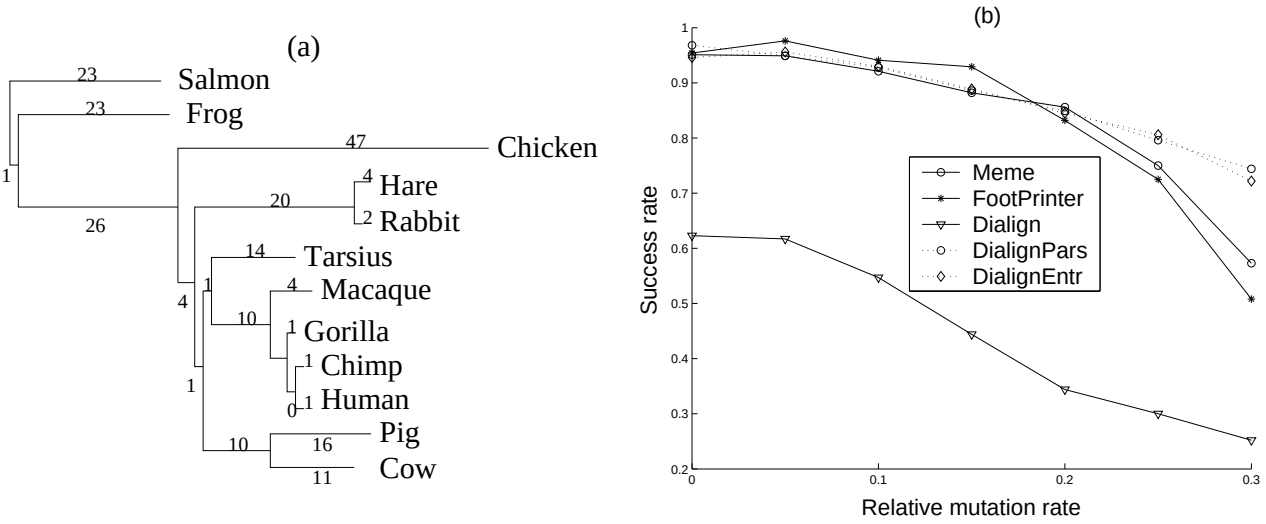

Figure 2: On the left is the tree inferred from the beta-globin data set of Section 4 using pairwise alignment scores and the fitch program from the phylip package [7]. The success rates reported on the right are for sequences simulated using this tree. The results for ClustalParsimony and ClustalEntropy are not shown as they are very similar to those of DialignParsimony and DialignEntropy, respectively.

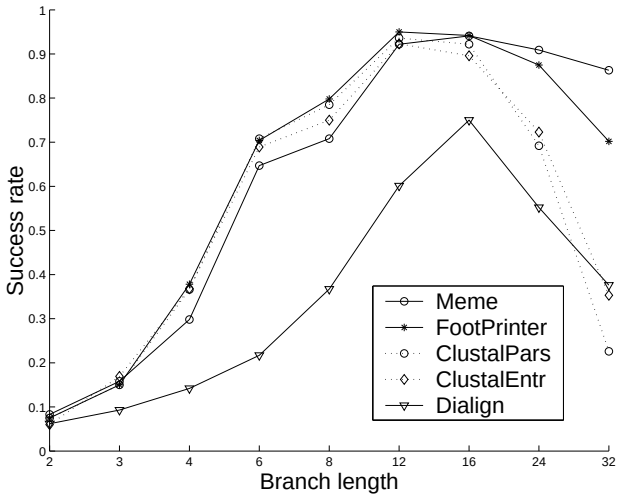

Figure 3: Success rates on the balanced binary tree with 8 leaves, as a function of the length $\lambda(e)$ of the branches of the tree. The motif relative mutation rate R is 0.1. Results for DialignParsimony and DialignEntropy are omitted for clarity but are very similar to their Clustal counterparts.

matter what scoring method is used. The accuracy of FootPrinter also starts decreasing. This is because, to achieve reasonable FootPrinter running times, we constrained the search to motifs with at most two mutations per branch of the tree and at most five mutations in total. As λ increases, it becomes likely that the planted motif will violate this condition, and Footprinter is thus not be able to detect it.

3.3 Planted motifs absent from some species

In real biological sequences, some regulatory elements are conserved in only a subset of the sequences considered (usually forming a phylum). To assess how accurately each program identifies such motifs, we modified our experimental design so that the mutation rate inside the planted motif is $R(e) < 1$ along each branch e of the left subtree, but $R(e) = 1$ (i.e. the same as outside the motif) along the branches of the right subtree. In this case, the correct solution is to identify the planted motif in the four species of the left subtree, and to predict nothing in the right subtree (i.e. $P_i = \emptyset$ for the species derived from a branch e with $R(e) = 1$). To allow each of the seven methods to make predictions that cover only a subset of the species, we modified them as follows. FootPrinter has an option that allows the detection of such motifs. When this option in used, FootPrinter reports motifs that have low parsimony score but are present in a subset of species that span a large amount of evolution [2]. We used that option for this study. The same type of parsimony score vs. evolutionary span tradeoff was used to report motifs for ClustalParsimony and DialignParsimony. This was achieved by running FootPrinter on the set of *aligned* sequences, and by considering as admissible motifs only those whose substrings are aligned. In the case of relative entropy-based methods, we report the region of the alignment with the highest relative entropy, ignoring any sequence with at least one gap in the 10-nucleotide region scored. Regions in which fewer than 50% of the species are gapless are ignored altogether. Finally, in the case of MEME, the "zero or one occurrence per sequence" option was used. The exact parameters for each program are listed on our web site [3].

Figure 4(a) shows that on such data sets, the phylogenetic information used by parsimony-based methods makes a big difference. This is because the topology of the tree loosely determines which subsets of species are more likely to contain the same motif. Figure 4(b) shows the results of a similar experiment, where the tree T has been re-rooted so that it has six leaves in the left subtree and two in the right subtree, and where the mutation rate is $R(e) < 1$ in the left subtree and $R(e) = 1$ in the right subtree. Here again, phylogenetic information provides a clear advantage.

Running times have rarely been an issue in this study. In general, MEME and ClustalW have similar running times, each taking a few seconds to run. Dialign is five to 10 times slower. The running time of FootPrinter depends on the conservation of the motif sought. It is similar to that of MEME and ClustalW for relatively well conserved motifs, and up to 10 to 50 times slower for motifs with high parsimony scores.

4 Comparisons on Orthologous Promoters

4.1 Methodology

To evaluate the accuracy of these programs on real biological data, we assembled sets of orthologous sequences for which at least one binding site has been experimentally verified. More precisely, we considered all metazoan genes for which known binding sites are listed in Transfac 5.0 [28]. There are more than 1000 such genes. We extracted orthologous upstream sequences for these genes, keeping only data sets for which we found at least 250 bp of upstream sequence in at least three species. This process was carried out automatically using keyword searches in GenBank, and is thus likely to have missed a few orthologous sequences. Nonetheless, after manual refining to ensure

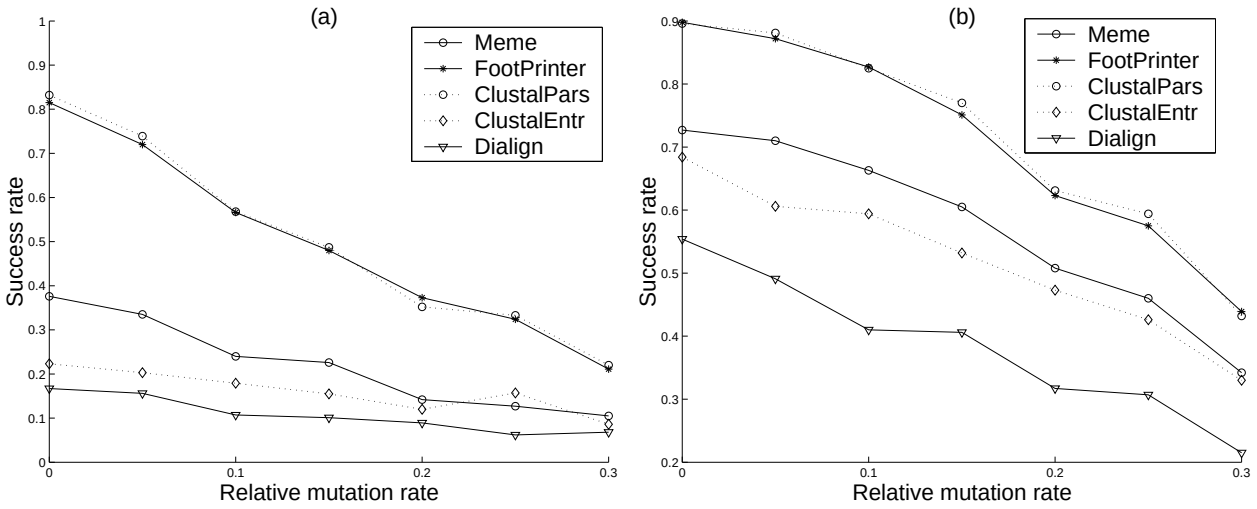

Figure 4: Success rates on the balanced binary tree with 8 leaves, with all branches of the tree of length $\lambda(e) = 16$. In (a), the motif evolves at the given relative mutation rate along the 7 branches of the left subtree, and evolves at the background rate along the branches of the right subtree. In (b), the tree has been re-rooted so that it has 6 leaves in the left subtree and 2 in the right subtree; the motif evolves at the given relative mutation rate along all branches of the left subtree, and evolves at the background rate along the 3 branches of the right subtree. Results for DialignParsimony and DialignEntropy are omitted for clarity but are very similar to those for ClustalParsimony and ClustalEntropy, respectively.

that each dataset contains truly orthologous promoters, we obtained the set of 27 sets of orthologous upstream sequences listed in Table 2. Each set contains at least one sequence in which at least one experimentally verified binding site is known. Although most data sets contain sequences from at least four species, it is quite common that only one or two of the orthologous input sequences contain any experimentally verified sites. If more than one does, we evaluate the correctness of the predictions only on the species that contains the greatest number of positions in annotated sites. The length of the available upstream sequence varies from gene to gene, between 250 bp and 2000 bp.

We compared the performance of the same programs used in Section 3.3. Each program was directed to report the five best motifs of length 10. It is easy to do so with MEME, by setting the parameter `nmotifs` = 5. For Dialign, DialignEntropy and ClustalEntropy, we greedily chose the 5 nonoverlapping regions with the highest average scores. FootPrinter was run allowing for motif losses and, for each sequence, the 50 nucleotides belonging to the motifs with greatest statistical significance were output. The same method produced predictions for DialignParsimony and ClustalParsimony. The 27 sets of sequences and the predictions of each program are available on our web site [3].

4.2 Results and Discussion

Table 2 lists, for each program, the number of positions correctly predicted to belong to a regulatory element (out of 50 positions predicted by each program). For each gene, we declare a program a winner (in bold in the table) if its number of correct predictions is within 25% of the number of correct predictions made by the best program for that gene, and if the number of correct predictions is at least 50% more than what would be expected from a random predicter (which is 50 times the fraction of the input sequence that belongs to the known binding site). We declare a program a loser (in italic) if its number of correct predictions is less than 50% of that of the best program, or if its number of correct predictions is less than would be expected from a random predicter. These choices regarding what constitutes a winner or a loser are arbitrary, but variations in this scheme yield similar qualitative results.

The results are for the most part in agreement with those obtained on simulated data. Parsimony-based approaches uniformly outperform all other methods, with FootPrinter obtaining the best performance. Compared to the results on simulated data, MEME does much worse than expected. Further investigation will be needed to understand the reason. ClustalEntropy and DialignEntropy do fairly poorly compared

Table 2: Number of nucleotides correctly predicted to belong to a binding site, for a random predictor (rand), MEME (M), FootPrinter (FP), ClustalParsimony (CP), ClustalEntropy (CE), Dialign (D), Dialign Parsimony (DP), and DialignEntropy(DE). The second column indicates the fraction of the input sequence that belongs to a known binding site, for the given species. See the text for an explanation of wins (in bold) and losses (in italic).

Gene (Species)	#pos/length	rand	M	FP	CP	CE	D	DP	DE
acetylcholinesterase (Mouse)	48/2000	1.2	*0*	*0*	*0*	*0*	10	*0*	**16**
albumin (Human)	38/250	7.6	*6*	**17**	**17**	*0*	**22**	**18**	14
α-lactalbumin (Rat)	16/250	3.2	*0*	**6**	**6**	4	*0*	*0*	*0*
apolipoprotein E (Human)	132/1500	4.4	*0*	**9**	**8**	*0*	6	**8**	*0*
β-actin (Human)	19/500	1.9	**14**	**14**	10	**14**	**17**	10	10
β-globin (Chicken)	22/450	2.4	**10**	**10**	**10**	**11**	6	8	*5*
c-fos (Human)	87/500	8.7	21	**32**	**32**	**30**	**32**	**31**	**24**
c-myc (Human)	178/2000	4.5	*0*	**20**	*0*	12	*0*	*0*	*0*
dihydrofolate reductase (Mouse)	72/750	4.8	*13*	**33**	20	*5*	*0*	20	*8*
fibroin (Bombyx)	194/750	12.9	*7*	28	**35**	**31**	**38**	**35**	*4*
fibronectin (Rat)	48/350	6.9	*2*	10	*0*	**19**	**19**	*0*	*9*
growth hormone (Rat)	137/250	27.4	**45**	31	31	30	**45**	37	34
insulin (Human)	71/500	7.1	8	8	**13**	*3*	9	8	*5*
interferon γ (Human)	13/300	2.2	*0*	**12**	**12**	*4*	*0*	**12**	**10**
interleukin-2 (Human)	156/1000	7.8	**18**	*2*	*2*	**20**	14	*2*	11
lipoprotein lipase (Human)	64/1000	3.2	*0*	*0*	*0*	*0*	*0*	*0*	*0*
lysosyme (Chicken)	79/450	8.8	*0*	12	*7*	9	**15**	12	12
myogenin (Chicken)	117/750	7.8	36	**46**	**46**	**42**	**49**	**46**	**40**
myoglobin (Human)	25/300	4.2	*0*	**8**	**8**	*0*	*0*	**8**	*0*
olfactory marker (Rat)	108/750	7.2	*3*	*0*	*8*	*8*	**18**	*8*	*0*
prolactin (Human)	85/250	17.0	32	**45**	**45**	*10*	30	30	30
SRY (Human)	26/250	5.2	*0*	*0*	*0*	*0*	*4*	*0*	*0*
thymidine kinase (Human)	25/250	5.0	*5*	**11**	*0*	*5*	*1*	**11**	*0*
thyroglobulin (Rat)	111/250	22.2	26	31	**38**	*20*	25	**38**	*10*
triosephosphate isomerase (Human)	38/250	7.6	*5*	*0*	*0*	8	8	*0*	*7*
tumor necrosis factor α (Human)	34/1000	1.7	*0*	**10**	**10**	*0*	*0*	**10**	*0*
tyrosine hydroxylase (Rat)	28/250	5.6	*6*	*0*	*0*	**16**	*0*	*0*	**16**
	# WINS		**4**	**14**	**13**	**8**	**9**	**10**	**5**
	# LOSSES		*18*	*7*	*11*	*14*	*10*	*10*	*16*

to their Parsimony counterparts.

Using a slightly different measure of performance relative to the random predictor, all the methods perform reasonably: FootPrinter obtains on average 51% more correct predictions than the random predictor, and MEME obtains 29% more. However, we also observe that, for several genes, the number of correct predictions made by even the best program is quite low. These low scores should not be seen as too discouraging, as it is possible that many of the predictions made by the programs are actual binding sites that have not yet been documented. Given the small number of data sets available, it is difficult to draw conclusions regarding correlation between a program's performance and the type of data used (e.g. number and diversity of species, length of the sequences, etc.). Still, it is interesting to notice that the few data sets where alignment methods beat motif-finding approaches contain relatively closely related species (fibroin: four diptera species; thyroglobulin: four mammals; tyrosine hydroxylase: three mammals). Conversely, motif-finding techniques seem to have the advantage when the species considered are more highly diverged (c-myc: fish, frog, chicken, mammals; dihydrofolate reductase: mammals, drosophila).

5 Conclusion

We have evaluated the accuracy of programs for phylogenetic footprinting on both synthetic and biological data sets. On synthetic data, we observe that phylogenetic methods generally outperform other approaches. When the sequences are highly diverged, motif-finding approaches are the only ones yielding good accuracy. When the sequences are closely related, alignment-based approaches obtain slightly better results than motif-finding methods. However, this advantage is quite small, and in general, FootPrinter's predictions are almost always among the best ones. The only exception is when one is searching for highly diverged motifs, which are very expensive to find using FootPrinter, whose accuracy is then decreased by heuristics used to speed up computation. In that case, MEME appears preferable. Results on biological sequences confirm these results, favoring phylogenetic approaches, but yielding surprisingly low accuracy for MEME.

The results reported in this paper suggest the following guidelines for the use and development of programs for phylogenetic footprinting. When the orthologous sequences can be aligned reliably (e.g. if we only have mammalian sequences), using the alignment provides a small increase in accuracy, and also greatly reduces the computational complexity of motif detection. However, in most cases, it is difficult to know in advance whether the multiple alignment is trustworthy. In such cases, motif-finding programs provide a safe alternative, usually without losing too much accuracy. Finally, for sets of closely related sequences, or for data sets with some sequences not containing the regulatory elements sought, using phylogenetic information helps evaluate a motif's conservation more accurately and yields much better predictions.

Acknowledgments

This material is based upon work supported in part by a fellowship from the Fonds Québécois de la Recherche sur la Nature et les Technologies, in part by the National Science Foundation under grants DBI-9974498 and DBI-0218798, in part by the National Human Genome Research Institute under grant R01 HG02602-01, and in part by the Howard Hughes Medical Institute.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80, Oct. 1995.

[2] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *Journal of Computational Biology*, 9(2):211–223, 2002.

[3] M. Blanchette and M. Tompa. Footprinter web site. `bio.cs.washington.edu/FootPrinterResults`.

[4] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12:739–748, 2002.

[5] P. Cliften, L. Hillier, L. Fulton, T. Graves, T. Miner, W. Gish, R. Waterston, and M. Johnston. Surveying saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Research*, 11:1175–1186, 2001.

[6] I. Dubchak, M. Brudon, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Research*, 10:1304–1306, 2000.

[7] J. Felsenstein. Phylip - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[8] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[9] D. Gumucio, D. Shelton, W. Zhu, D. Millinoff, T. Gray, J. Bock, J. Slightom, and M. Goodman. Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the β-line globin genes. *Molecular Phylogenetics and Evolution*, 5(1):18–32, 1996.

[10] R. C. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Research*, 8:959–966, 1997.

[11] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7/8):563–577, July/August 1999.

[12] D. Higgins, J. Thompson, and T. Gibson. Using CLUSTAL for multiple sequence alignments. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 383–401. Academic Press, New York, 1996.

[13] J. D. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, 296(5):1205–1214, 2000.

[14] K. Jiao, J. Nau, M. Cool, W. Gray, J. Fassler, and R. Malone. Phylogenetic footprinting reveals multiple regulatory elements involved in control of the meiotic recombination gene, rec102. *Yeast*, 30;19(2):99–114, 2002.

[15] M. Krawczak, N. Chuzhanova, and D. Cooper. Evolution of the proximal promoter region of the mammalian growth hormone gene. *Gene*, 237:143–151, 1999.

[16] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[17] G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, and E. Rubin. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research*, 12(5):832–839, 2002.

[18] G. G. Loots, R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288:136–140, 2000.

[19] J. F. Manen, V. Savolainen, and P. Simon. The *atpB* and *rbcL* promoters in plastid DNAs of a wide dicot range. *Journal of Molecular Evolution*, 38(6):577–582, 1994.

[20] L. McCue, W. Thompson, C. Carmack, M. Ryan, J. Liu, V. Derbyshire, and C. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Research*, 29(3):774–782, 2001.

[21] A. M. McGuire, J. D. Hughes, and G. M. Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research*, 10:744–757, 2000.

[22] B. Morgenstern. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, 1999.

[23] P. Pevzner and S. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278. AAAI Press, Aug. 2000.

[24] S. Sinha and M. Tompa. Performance comparison of algorithms for finding transcription factor binding sites. In *3rd IEEE Symposium on Bioinformatics and Bioengineering*. IEEE Computer Society, Mar. 2003.

[25] N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman, W. Miller, and R. Hardison. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research*, 27(19):3899–3910, 1999.

[26] J. Stoye, D. Evers, and F. Meyer. Rose: generating sequence families. *Bioinformatics*, 14:2:157–163, 1998.

[27] D. Tagle, B. Koop, M. Goodman, J. Slightom, D. Hess, and R. Jones. Embryonic ϵ and γ globin genes of a prosimian primate (Galago crassicaudatus) nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, 203:439–455, 1988.

[28] E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, 1996. `transfac.gbf-braunschweig.de/TRANSFAC/`.