The statistical power of phylogenetic motif models

John Hawkins and Timothy L. Bailey

Institute for Molecular Bioscience, QLD 4072,
The University of Queensland, Australia
j.hawkins@imb.uq.edu.au,t.bailey@imb.uq.edu.au,
WWW home page: http://www.imb.uq.edu.au

Abstract. One component of the genomic program controlling the transcriptional regulation of genes are the locations and arrangement of transcription factors bound to the promoter and enhancer regions of a gene. Because the genomic locations of the functional binding sites of most transcription factors is not yet known, predicting them is of great importance. Unfortunately, it is well known that the low specificity of the binding of transcription factors to DNA makes such prediction, using position-specific probability matrices (motifs) alone, subject to huge numbers of false positives. One approach to alleviating this problem has been to use phylogenetic "shadowing" or "footprinting" to remove unconserved regions of the genome from consideration. Another approach has been to combine a phylogenetic model and the site-specificity model into a single, predictive model of conserved binding sites. Both of these approaches are based on alignments of orthologous genomic regions from two or more species. In this work, we use a simplified, theoretical model to study the statistical power of the later approach to the prediction of features such as transcription factor binding sites. We investigate the question of the number of genomes required at varying evolutionary distances to achieve specified levels of accuracy (false positive and false negative prediction rates). We show that this depends strongly on the information content of the position-specific probability matrix and on the evolutionary model. In particular, we show that the use of site-specific equilibrium distributions for calculating the substitution rates significantly reduces the number of genomes that are required for a given accuracy. We also quantify the loss in accuracy caused by the presence of non-motif regions evolving more slowly than the assumed background substitution rate. Finally, we explore the accuracy of the theoretical model by applying it to a transcription factor binding site prediction task in yeast, and show that it provides a reasonable estimate of the potential accuracy of phylogenetic motif search.

1 Introduction

Phylogenetic motif models are probabilistic models of sequence features. They are a natural extension of the probabilistic motif models used in computational biology to represent and identify sequence features such as transcription factor binding sites (TFBSs), splice junctions and binding domains in DNA, RNA and protein molecules, respectively [GuhaThakurta, 2006, Stormo, 2000]. Phylogenetic motif models extend the usefulness of standard motifs by leveraging the knowledge that important features in biological sequences tend to evolve more slowly than the neutral rate, a standard assumption of comparative genomics. Phylogenetic motif models are a refinement of the idea of phylogenetic footprinting [Gumucio et al., 1992] and shadowing [Boffelli et al., 2003], key tools in the arsenal of comparative genomics. This study examines the statistical power provided by phylogenetic motif models for identifying sequence features as a function of the number of comparative genomes, their average evolutionary distance and the information content of the motif.

Standard probabilistic motif models assume that sequence features have a fixed length, and that the frequencies of the letters (e.g., base or residue) that occur at each position in an occurrence of the feature are independent. This allows the motif model to be completely described by a single position-specific probability matrix (PSPM), M, where $M_{a,i}$ gives the probability of observing letter a at position i in the motif. Thus, the motif model defines the probability of any sequence, x, of the correct length, as the product of the corresponding terms in M, written here as $Pr(x|M)$.

Phylogenetic motif models extend standard motif models to allow them to define the probability of a *multiple alignment*, rather than of a sequence. In addition to the motif model, M, they incorporate a model of evolution (substitution model, e.g., Jukes-Cantor or Hasegawa-Kishino-Yano (HKY) [Felsenstein, 1981]), E, and a phylogenetic tree, T. Each sequence in the alignment is associated with one leaf in the tree, as they are assumed to be orthologous (descended from a common, ancestral sequence.) In essence, the model treats each column of the multiple alignment as though it were a "letter", and defines the probability of the alignment (with the same width as the motif) as the product of the probabilities of the individual columns. Under the model the probability of an alignment *column* is the probability of observing the letters in the column assuming the evolutionary substitution model and assuming that the sequences (and their ancestors) have been under purifying selection to maintain the frequencies given in the corresponding column of the motif, M. Thus, the model is a direct generalization of the standard probabilistic motif model and it defines the probability of a multiple alignment column, σ, here written $Pr(\sigma|M, E, T)$. (Since our model assumes that alignment columns are independent, this is easily generalized to the probability of an alignment of width w by taking the product of the column probabilities.)

The focus of this paper is on the theoretical limits on the utility of phylogenetic motif models for *identifying* genomic features when the motif is known, here referred to as "motif search". In the past few years, algorithms have been developed that use phylogenetic motifs for motif search, notably the Monkey algorithm [Moses et al., 2004b, 2006] and Motiph (unpublished, available as part of the Meta-MEME software `http://metameme.sdsc.edu`). These tools make more sophisticated use of the information implicit in an alignment of orthologous sequences than tools such as the UCSC Genome Browser [Kent et al., 2002], the ECR Browser [Loots and Ovcharenko, 2004], and ConSite [Sandelin et al., 2004b], because they explicitly use a model of substitution and the evolutionary relationships and distances specified by a phylogenetic tree.

Despite the demonstrated utility of such tools [Moses et al., 2004b, 2006], little is known about the limits of their ability to detect genomic features. This is mainly due to the difficulty and expense of obtaining "gold standard" sets of all known, functional instances of a feature in a genome. Lacking such a gold standard, it is difficult to validate the "false positive" (FP) rate of a model, defined as the number of false positive predictions divided by the total number of positive predictions made. This is difficult to validate since one doesn't know how many of the supposed false positive predictions may be real. Similarly, if true instances of a feature are missing from the validation set, one cannot accurately estimate the "false negative" (FN) rate of a model, defined as the false negatives divided by all negative predictions.

An important biological application where this problem is particularly acute is in the identification of transcription factor binding sites (TFBSs), where it is well known that standard probabilistic motif search suffers from overwhelming numbers of false positive predictions (the so-called "Futility Theorem" [Wasserman and Sandelin, 2004].)

The evolutionary motif search algorithms already mentioned were developed in large part specifically to overcome this problem, but little or no data is available as to the extent to which they succeed.

In this paper we develop a theoretical framework for analyzing the statistical power of an evolutionary motif model used in motif search in a "target" genome. We assume that the search uses the standard approach for scoring putative sites in the multiple alignment–the log-odds score–the logarithm of the ratio of the probability of the site given the evolutionary motif model or given the neutral ("background") model, respectively. Our framework allows us to compute, for any specified motif and evolutionary model, the number of comparative genomes required in order to achieve given FP and FN rates. Conversely, we can compute a theoretical ROC-like curve for a motif, plotting FN rate as a function of FP rate for a given number of genomes at a given evolutionary distance from the target genome.

To compute the theoretically achievable FN and FP rates of a motif, we must estimate the distributions of the log-odds scores under the motif and background models, respectively. To make this computation feasible, we make the simplifying assumption that each of the comparative genomes are on a phylogenetic star with equal branch lengths. This assumption makes the probabilities of the letters in each of the genomes independent, given the letter at the root of the tree. It also makes the contributions to the log-odds score from each genome additive, and allows us to parameterize a problem with a single distance, D, the length of each of the branches in the star tree.

This is the same approach as taken by Eddy [Eddy, 2005], who studied the simpler problem of determining if a column or set of columns in a multiple alignment was *conserved*, as opposed to our goal of identifying if a set of columns is a *conserved instance of the particular feature type* defined by the motif model. In contrast to the Eddy study, however, we do not require that the target genome be placed at the center of the star topology, but consider the case where an unknown ancestral genome is at the center. This allows our results to be directly applicable to existing phylogenetic motif search algorithms such as Monkey. In what follows, we will refer to Eddy's goal as estimating the statistical power of *phylogenetic footprinting*.

Our theoretical analysis quantifies the maximum sensitivity and specificity of phylogenetic motifs during motif search under ideal conditions. We assume that we have a correct alignment of orthologous sequences. We presume that we know the substitution rate of the motif, R_M, in reference to the neutral substitution rate, which is our metric of evolutionary distance. In some cases, we add an additional assumption that the background substitution rate R_B, varies from the neutral rate. This variation allows us to investigate searching for motifs within sequence regions that are more conserved than neutral sequence. We assume that the evolutionary substitution model is correct. In these assumptions, we mirror the analysis of Eddy [Eddy, 2005]. We further assume that the feature of interest is accurately represented by the probabilistic motif, M, and the non-site positions are accurately modeled by a 0-order Markov model with parameters B. Of course, we also assume that the underlying premise of phylogenetic motif search is correct–that motif sites (features) are under identical purifying selection in each of the comparative organisms and their common ancestor. We summarize our model parameters as $\theta_M = \{M, E, T, R_M\}$ for the motif model, and $\theta_B = \{B, E, T, R_B\}$ for the background model.

Our framework allows us to explore a number of factors affecting the statistical power of phylogenetic motif models. We demonstrate that identifying conserved motifs across phylogenies requires fewer genomes than the number predicted for phylogenetic footprinting. We show that this is due almost entirely to the site specific probability distributions used by the motif models. We demonstrate that the information content of a motif has an inverse relationship with the number of required genomes, up until about 17 bits. We provide estimates of the number of genomes needed when the motif is less conserved, and we explore the difficulty encountered when searching for motifs in genomes with large regions evolving more slowly than the neutral rate. Finally, we test the accuracy of the model by comparing the theoretical and actual accuracy of a transcription factor binding site prediction task in yeast.

2 Methods

2.1 Phylogenetic motif model

Our phylogenetic motif model involves computing a log-odds ratio of an alignment column σ of N sequences given a model of both the motif θ_M and the background θ_B. (Since log-odds scores are additive, this generalizes easily to the score for an alignment of width w by summing the scores of the

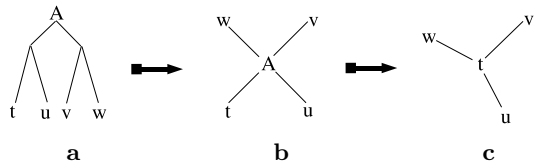

Fig. 1. Depiction of the relationship between the different phylogenetic tree topologies.

individual columns.) This log-odds score is written as

$$S(\sigma) = \log \frac{Pr(\sigma|\theta_M)}{Pr(\sigma|\theta_B)}.$$

The two models θ_M and θ_B incorporate the frequencies in the position-specific probability matrix (PSPM) of the motif, M, the background frequencies of the residues, B, different substitution rates for the two models R_M and R_B respectively, and an evolutionary model for calculating the substitution probabilities. In both cases, we use a phylogenetic star tree with equal branch lengths to describe the relationship between the genomes. However, we have two alternative implementations, one that places the target genome in the center of the star (As Eddy does [Eddy, 2005]), Fig. 1**c**, and the other in which we place an unknown ancestor in the center (as has been done in other work [Siddharthan et al., 2005, Sinha et al., 2004, Moses et al., 2004b]), Fig. 1**b**. In order to be able to compare results using the two different types of model, the branches in the ancestor-in-the-center model have length D, and those in the target-in-the-center have length $D(N-1)/N$. Thus, our D parameter is smaller than Eddy's, and is an estimate of the average independent branch length in the actual species tree (Fig. 1**a**).

When we place an unknown ancestor in the center, the probabilities of each of the letters in the alignment column are independent, given the letter in the ancestor. So, the score of the alignment column, $S(\sigma)$, is the sum of the scores of the individual letters in the alignment,

$$S(\sigma) = \sum_{i=1}^{N} \log \frac{Pr(\sigma_i|\theta_M)}{Pr(\sigma_i|\theta_B)}.$$

However, when we place the target genome in the center of the star, we must replace the first component of this sum,

$$S(\sigma) = \log \frac{Pr(\sigma_1|M)}{Pr(\sigma_1|B)} + \sum_{i=2}^{N} \log \frac{Pr(\sigma_i|\theta_M)}{Pr(\sigma_i|\theta_B)}.$$

Note that, when the target genome is in the center of the star, the probability of the site in the target (first) genome is defined completely by the motif model, M, and background model, B, and does not involve the evolutionary substitution model. We compute both of these scores using the "pruning algorithm" of Felsenstein [Felsenstein, 1981].

In this study, we use the HKY [Hasegawa et al., 1985] substitution model to calculate the substitution probabilities for both the background and the motif evolutionary models. (Our analysis allows any of the standard substitution models, and our implementation incorporates the Jukes-Cantor, Kimura 2-parameter, F81, F84, HKY and Tamura-Nei models). For the motif evolutionary model, we apply the Halpern-Bruno modification [Halpern and Bruno, 1998], using the appropriate column of the motif PSPM as the equilibrium frequencies. The default values for the substitution rates (unless specified otherwise) are identical to Eddy's values of $R_B = 1.0$ and $R_M = 0.2$.

We use the parameter settings of the HKY model employed in MONKEY, so that the transition-transversion ratio is set to 3.8, and the background distribution, B, is set to $B_A = B_T = 0.3$ and $B_C = B_G = 0.2$. These values are very similar to the ones employed by Eddy in his numerical verification of his phylogenetic footprinting study using an HKY-generated sample [Eddy, 2005]. When the unknown ancestor is in the center of the phylogenetic star tree, the scoring function we use in this study is identical to that of MONKEY [Moses et al., 2004b].

The probability of an alignment column σ, given a model θ has two different formulas, depending on the phylogenetic tree we use. With the unknown ancestor, A, in the center (refer to Fig. 1**b**), it is a sum over the probabilities of seeing each letter in the unknown ancestor, multiplied by the probability of that ancestor generating the observed alignment.

$$Pr(\sigma|\theta) = \sum_{a \in \mathcal{A}} Pr(A = a|\theta) \prod_{i=1}^{N} Pr(\sigma_i|\theta, A = a),$$

where σ_i is the letter in the alignment column from genome i. However, when the target genome, t, is placed in the center (refer to Fig. 1**c**), the probability of the alignment becomes conditional on the letter in the target genome sequence, t. That is,

$$Pr(\sigma|\theta) = Pr(\sigma_1|M) \times \prod_{i=2}^{N} Pr(\sigma_i|\theta, t = \sigma_1),$$

where σ_1 is the target genome's letter.

There are two assumptions of independence that simplify the process of calculating these probability distributions. Firstly, the assumption of a phylogenetic star means that each genome evolves from the target (or unknown ancestor) independently, hence the probability of N genomes, is the probability of the first $N-1$ genomes times the probability of seeing the N^{th} genome. Secondly, the fact that we assume independence between the positions within the motif, means that the probability distribution for the score considering only the first m columns in the multiple alignment is the probability of seeing the first $m-1$ columns times the probability of the m^{th} column.

These assumptions allow us to apply dynamic programming to calculate a discretized approximation to the probability distribution of log-odds (S) scores [Staden, 1990]. We calculate the distribution under both the assumption that we are dealing with a conserved motif, and under the assumption that we are dealing with a neutral sequence. We are then able to generate the cumulative distributions under each model and determine if, for the given number of genomes, there is an S score threshold that satisfies the false positive and false negative criteria.

The computational complexity of our algorithm is linear in the length of the motif, l, the maximum number of genomes to be tested, g, and the size, s, of the discretized distribution, so it has "big-O" complexity $O(lgs)$. However, we found that to obtain reliable cumulative distributions we needed to use a discretization size, s, of $2 \cdot 10^4$. Hence, the computation time for long motifs or large numbers of genomes can be long. For example to produce the two cumulative probability distributions for four genomes at evolutionary distances of 0.19 and 0.31 the algorithm takes on the order of 30 minutes on a 2 gigahertz workstation. However, to calculate the number of required genomes over 100 different evolutionary distances with a low information content motif requires 80 hours of computing time on the same workstation.

We validated the cumulative distributions by generating Q-Q plots, in which we sample 10,000 alignments generated under the two models, and plotted the p-values of each log-odds score predicted by our model against those suggested by the random sampling. The Q-Q plots show that the estimated p-values are very accurate (data not shown).

2.2 Motifs and information content

In this paper, we use motifs from the JASPAR [Sandelin et al., 2004a] and SCPD [Zhu and Zhang, 1999] databases. These databases contain "count" matrices, computed by aligning known TFBSs and counting the number of occurrences of each nucleotide in each position in the known sites. We convert these counts to a probability matrix, M, by normalizing each column to sum to one. To account for small-sample errors, we add a "pseudocount", equal to 0.375, to each count before normalizing. (This value was determined to be optimal for normalizing TFBS motifs by Frith *et al.* [Frith et al., 2004].)

To calculate the information content (IC) of the motif, we use the same derivation of the Shannon entropy employed in the calculation of sequence LOGOs [Schneider and Stephens, 1990]. The information content (in bits) of a DNA PSPM, M, of length L, is given as

$$IC(M) = \sum_{i=1}^{L} \Big(2 + \sum_{a \in \mathcal{A}} M_{a,i} \log_2(M_{a,i})\Big).$$

This is equal to the average log-odds score of motif instances when the background base distribution is uniform.

2.3 Yeast multiple alignments and known TF binding sites

In the final part of our study we validate the theoretical model proposed here, using the Motiph algorithm to search for TFBSs in yeast intergenic regions. For the motif, we selected the MCM1 motif from the SCPD database due to the large number of validated binding sites listed for this TF. We searched the multiple alignments of orthologous intergenic regions from four yeast species constructed by Kellis *et al.* [Kellis et al., 2003]. These alignments were constructed using a manual procedure involving the multiple alignment algorithm ClustalW [Chenna et al., 2003]. Each known MCM1 site in *S. cerevisiae*, as specified in SCPD, was considered a "positive", and all other positions in *S. cerevisiae* are considered "negative" in computing the empirical FN and FP rates at various score thresholds. Positions in the multiple alignments containing gaps or ambiguous characters were ignored in the analysis.

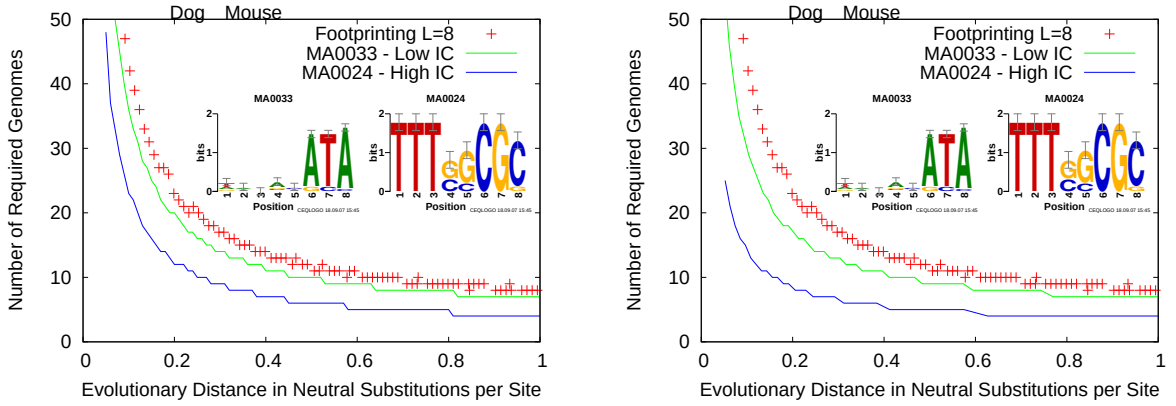

Fig. 2. Statistical power of phylogenetic motif models compared with footprinting. The plots show results for length-eight motifs using the ancestor-in-the-center model (panel **a**) or the target-in-the-center model (panel **b**), compared with the length-eight footprinting task. In each panel, the top curve is a reproduction of Eddy's results for a length-eight conserved region. The middle line shows the results for our phylogenetic motif model using the low information content JASPAR motif MA0033. The lower line shows the results of our model using the high information JASPAR motif MA0024. All curves are for accuracies of FP $= 10^{-4}$ and FN $= 10^{-2}$. The x position of first letters of 'Dog' and 'Mouse' correspond to their approximate evolutionary distances to human.

3 Results

3.1 Phylogenetic motif search has more statistical power than footprinting

In our first study, we take two TFBS motifs from the JASPAR database [Sandelin et al., 2004a], both of length eight, but with different information contents. We compare the theoretical number of genomes required for accurately detecting sites of each of these motifs (computed by our approach) with the number of genomes required to simply predict the conservation of a length-eight region by phylogenetic footprinting. Our chosen motifs are the length-eight motifs with the highest and lowest information content in the JASPAR database: MA0033 (FOXL1) and MA0024 (E2F1), respectively. The LOGOs for these two motifs are shown in Fig. 2. We use a single setting for the statistical power, which is the most stringent used in Eddy's study: an FP rate of 10^{-4} and an FN rate of 10^{-2}, and we consider placing either the target genome or an unknown ancestor genome in the center of the phylogenetic star.

The statistical power of phylogenetic motif discovery is higher than that of phylogenetic footprinting for both high- and low-IC motifs. Using the more biologically plausible ancestor-in-the-center model introduced here (Fig. 2**a**), the theoretical power of the low-IC motif is predicted to be only slightly better than the power predicted by Eddy's model of phylogenetic footprinting for features of the same length as the motif ($L = 8$). At evolutionary distances greater than 0.3 substitutions per site, only one or two fewer comparative genomes are required for identifying TFBSs compared to footprinting length-eight features (at the stated FP and FN rates). The high-IC motif, however, has much higher statistical power, requiring substantially fewer comparative genomes at all evolutionary distances modeled. We explore the relationship between motif IC and the statistical power of phylogenetic motif search further in Sec. 3.3.

The less biologically plausible target-in-the-center model, although shown to be appropriate for modeling phylogenetic footprinting by Eddy, overestimates the statistical power of phylogenetic motif search. This model predicts that many fewer comparative genomes are required than does the ancestor-in-the-center model (Fig. 2**b**). This effect is most pronounced with high-IC motifs (blue curves in the figure), but holds for low-IC motifs as well (green curves in the figure). For example, the target-centric model predicts that searches using the high-IC motif require only 13 genomes at an evolutionary distance of 0.1 substitutions per site (Fig. 2**a**), whereas the ancestor-centric model predicts that 23 genomes are required. These predictions, however, are in very poor agreement with the empirical validation results we present in Sec. 3.5, where the

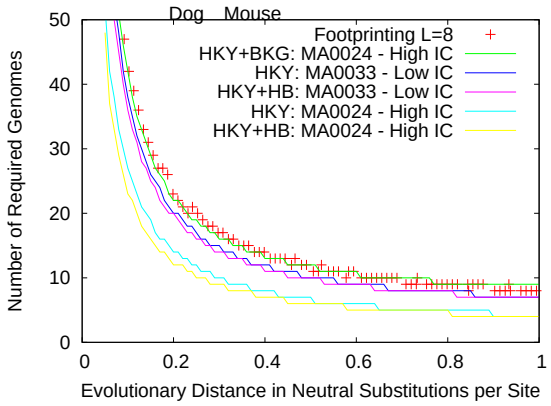

Fig. 3. Sources of statistical power of phylogenetic motif models. The plots show the results for the length-eight motifs using the ancestor-in-the-center model with modifications. The red points are a reproduction of Eddy's results for a length-eight conserved region. The green line uses the background base frequencies instead of the motif columns as equilibrium frequencies in the HKY model. The remaining lines use the motif columns as equilibrium frequencies in the evolutionary model without (HKY) and with (HKY+HB) the H-B modification. All curves are for accuracies of FP $= 10^{-4}$ and FN $= 10^{-2}$. The x position of first letters of 'Dog' and 'Mouse' correspond to their approximate evolutionary distances to human.

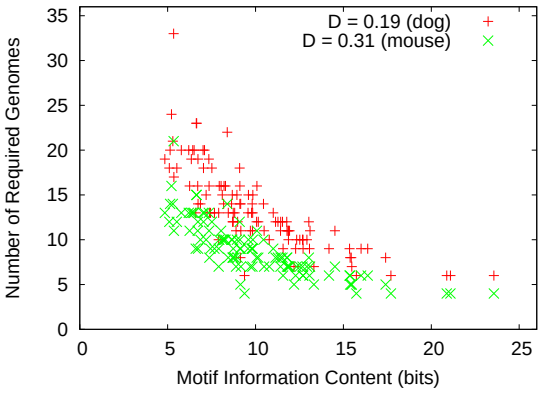

Fig. 4. Statistical power as a function of motif information content. Each point, (X, Y), represents an experiment using a different JASPAR motif. X is the information content of the motif, and Y is the minimum number of genomes required at a given evolutionary distance, D, in order to achieve a prediction accuracy of FP $= 10^{-4}$ and FN $= 10^{-2}$. We place the unknown ancestor genome at the center of the star and use the HKY model with the Halpern-Bruno modification for the motif substitution rates.

predictions of the ancestor-in-the-center model are shown to be reasonably accurate.

3.2 The sources of statistical power of phylogenetic motifs

Given that a consensus sequence can be represented by a probability matrix containing a probability of one for each of the specific residues, it is, in one sense, a motif with maximal information content. It might appear counter-intuitive that our model predicts fewer required genomes than for phylogenetic footprinting. This is compounded by the fact that Eddy confirmed his results numerically by sampling alignments generated using a HKY model with parameters almost identical to those used in our evolutionary model.

We therefore study the sources of statistical power of phylogenetic motif models. There are three main differences between our model and the footprinting model. Our evolutionary model places the motif at the root of the tree; it uses the frequencies of the bases in each motif column as the equilibrium frequencies in the HKY model; it used the H-B modification to the evolutionary model. We modify the study from Fig. 2**a** to isolate the contributions of each of these differences, and show the results in Fig. 3.

When we place the motif at the root of the tree, but use the background frequencies as the equilibrium frequencies in the HKY model, our model gives *nearly identical predictions* as the Eddy footprinting model, despite the fact that we are using a log-odds score function rather than counting differences with a target genome (green curve and red points in Fig. 3).

Most of the statistical power of phylogenetic motifs comes from using the motif columns for the equilibrium frequencies (light blue and dark blue lines in Fig. 3). Adding the H-B modification only results in about one fewer genome being required at evolutionary distances above 0.2 substitutions per site, at an accuracy of FP $= 10^{-4}$ and FN $= 10^{-2}$ (pink and yellow lines in Fig. 3).

3.3 Statistical power increases with motif information content

In order to obtain a greater indication of the influence of information content over the number of required genomes, we conducted a second study using all motifs from the JASPAR database [Sandelin et al., 2004a]. We calculate the number of genomes

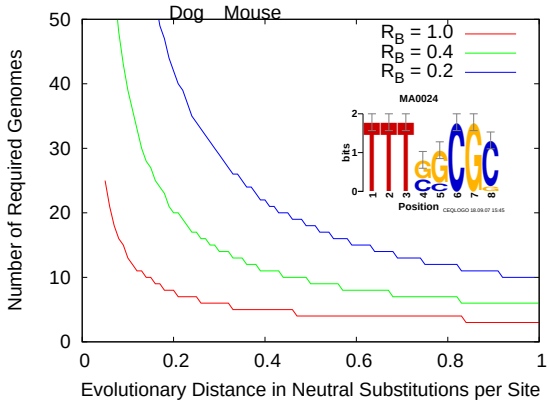

Fig. 5. Statistical power when background substitution rate is low. Using the high information content motif MA0024 and a motif substitution rate of $R_M = 0.2$, we estimated the number of genomes required for accuracy of FP $= 10^{-4}$ and FN $= 10^{-2}$. The three curves correspond to three different background substitution rates (R_B). We place the unknown ancestor in the center of the tree and use the HKY model with the H-B modification. First letters of 'Dog' and 'Mouse' correspond to approximate distance to human.

required to achieve a FP rate of 10^{-4} and a FN rate of 10^{-2} for each of the 123 motifs and plot it against the information content of the motif. We place the unknown ancestor in the center of the star and we use two different evolutionary distances (D)–0.19 and 0.31–the values of the independent branch lengths chosen by Eddy as representative distances corresponding to human-dog and human-mouse inter-genomic distances. (That is, D is approximately one-half the evolutionary distance between the human genome and the dog or mouse genomes.)

The results for both studies are shown in the scatter plot in Fig. 4. The most notable result is that there is a strong, general trend for the number of required genomes to decrease with information content. However, as the information content reaches and exceeds 17 bits, the plots appear to reach what appear to be limiting values dependent on the evolutionary distance.

3.4 Identifying motifs within conserved regions

One potential problem for identifying transcription factor binding sites is the fact that genomes often contain large regions evolving slower than the neutral rate. This means that the correct background model should have a substitution rate either not much faster or identical to the motif model. In the approach taken in our initial case studies we used a background rate of 1.0 and a motif rate of 0.2. To evaluate the effect of on statistical power when the TFBS motifs are embedded in highly conserved regions, we rerun the study shown in in Fig. 2**b** for the high information content motif MA0024 using background rates of $R_M = 0.4$ and $R_M = 0.2$. We keep the motif substitution rate constant at 0.2. The results are shown in Fig. 5.

As would be expected, the number of required genomes increases as we attempt to identify TFBSs within highly conserved sequence. In the worst case scenario, the background sequence is as conserved as the motif itself (ignoring the possibility that it might be more conserved) and three to six times the number of genomes are required to achieve the same statistical power compared with the case when the background is evolving five times slower than the motif.

3.5 Empirical validation of the phylogenetic motif model

There have been several attempts to determine the statistical power with which TFBSs in yeast can be identified. Kellis *et al.* [Kellis et al., 2003] and Cliften *et al.* [Cliften et al., 2003] both use four aligned *Saccharomyces* genomes. Eddy performs a follow up study using his model to determine the agreement with their results. We perform a similar study to measure the agreement of the ancestor-in-the-center model with reality using the task of identifying the sites of the TF MCM1 in yeast (*S. cerevisiae*), as described above in Methods. Briefly, we use the Motiph algorithm to search the multiple alignments of all intergenic regions of four yeast species with the MCM1 motif, using either the actual species tree [Kellis et al., 2003] (total independent branch length is 0.85), or a uniform star tree with *S. cerevisiae* in the center with branch lengths $D = \frac{0.85}{4} \approx 0.21$. For the background equilibrium frequencies, B, we use the yeast intergenic frequencies $B_A = B_T = 0.3$ and $B_C = B_G = 0.2$.

Reasonable agreement between the model and observed search accuracy is obtained when the motif substitution rate is assumed to be one-half the background rate–as seen in (Fig. 6**a**). The agreement between the model and the actual results is not very good when the motif substitution rate is assumed to be five times lower than the background (Fig. 6**b**). This suggests that the actual substitution rate for MCM1 TFBSs in yeast may lie somewhere in between 0.2 and 0.5 times the

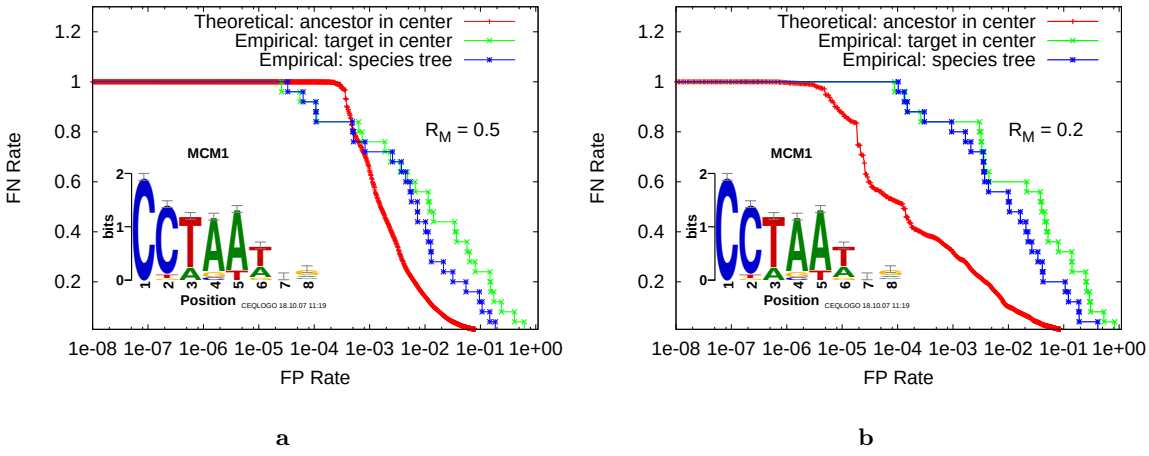

Fig. 6. Theoretical and empirical statistical power of TFBS discovery in yeast using Motiph. The plots show ROC-like curves for two experiments using SCPD motif MCM1. Panel **a** shows results when the motif substitution rate, R_M, is one-half the neutral rate. Panel **b** shows results when the motif substitution rate is only one-fifth the neutral rate. The red points show the achievable prediction accuracy when the unknown ancestor is placed in the center of the tree. The blue and green points show the actual accuracy achieved searching the yeast intergenic region alignments using motif MCM1. Green points are for a uniform star topology with *S. cerevisiae* in the center ($D = 0.21$); blue points give results using the actual species tree. In all experiments, the number of genomes is four, and the total independent branch length is 0.85.

background rate. This is in agreement with previous estimates of between one-half and one-third of the average intergenic rate [Moses et al., 2003].

It is to be expected that the theoretical FP rate predicted by the model (at a given FN rate) will be a lower bound on the rate achievable in an actual TFBS prediction task due to the numerous things that can go wrong in practice. For example, the model assumes that the orthologous regions have been correctly identified, that each multiple alignment is correct and that each TFBS is conserved in all of the species, and each of these assumptions may be partially (or completely) false. Assuming that the actual motif substitution rate is near $R_M = 0.5$ as suggested by the results in Fig. 6**a**, the discrepancy between the theoretical curve and the empirical curves can easily be explained by the existence, for instance, of inaccuracies in the multiple alignments or by TFBS "turnover" [Moses et al., 2006].

The difficulty of TFBS prediction, even using phylogenetic information, is apparent in this experiment. For example, at an FN rate of 0.5 the *theoretical* FP rate is around around 10^{-3} (0.0016) in Fig. 6**a** and about 10^{-4} (0.00012) in Fig. 6**b**. This means that even if we are satisfied with detecting only 50% of all TFBSs, we must deal with a false positive every 10000bp if the motif is evolving at one-fifth the background rate, and a false positive every 1000bp if it is evolving at one-half the background rate. In practice, the accuracy is even lower, as shown by the *empirical* curves in the figure. This points up the futility of genome-wide prediction of TFBSs for TFs such as MCM1 by using phylogenetic motif models to scan alignments of only four genomes of species as similar as the yeasts used here.

In this experiment, there is a clear advantage in using the actual species tree, rather than the simplified uniform star tree with the target genome in the center. This is evidenced by the fact that the FP rate using the species tree is generally lower at a given FN rate (Fig. 6**a**,**b**). The advantage is not huge, however, which lends support to the idea that the use of a star tree in our theoretical model is not particularly problematic. Any inaccuracy introduced by this simplification of biology made in our construction of the model will tend to make it more conservative, since it throws away some of the information contained in the actual species tree.

4 Discussion

Phylogenetic motif models are a specialization of profile phylogenetic hidden Markov models [Siepel and Haussler, 2004]. The introduction of phylogenetic relationships has been responsible for considerable improvement in the performance of *de novo*

motif discovery algorithms [Moses et al., 2004a, Sinha et al., 2004, Siddharthan et al., 2005]. However, the advantages of phylogenetic motif models for motif search are less clear. Even with the use of phylogenetic motif models and/or phylogenetic footprinting, transcription factor binding sites have remained difficult to identify due to their short lengths, low-specificity motifs and their presence inside highly conserved promoter regions. We have sought to analyse each of these limiting factors and present results on the number of genomes required at a variety of evolutionary distances to achieve reasonable statistical power.

In our first set of simulations, we show that fewer genomes are required to achieve given target levels of statistical significance in phylogenetic motif search than for phylogenetic footprinting. The amount of reduction depends on the information content of the motif. Compared to phylogenetic footprinting, phylogenetic motif search requires only about 50% as many genomes with high-information motifs, and 90% as many with low-information motifs.

We show that a simplified version of our model agrees almost perfectly with a very different model of phylogenetic footprinting proposed by Eddy. When we replace the site-specific equilibrium distribution in our model with the background base distribution, our model predicts statistical power virtually identical to Eddy's model of footprinting. This shows that the majority of the improvement in statistical power (compared with footprinting) is due to the use of site-specific equilibrium distributions in our phylogenetic motif model.

Our estimates of the statistical power are calculated under the assumption that nature is following a particular evolutionary model and phylogeny. This makes it dangerous to compare the statistical power of different models. For example, we found that the target-in-the-center model has higher statistical power than the ancestor-in-the-center model. It is wrong to conclude, however, that this model is preferable in practice. In our empirical test, the observed statistical power was well approximated by the ancestor-in-the-center model. The overly optimistic predictions of the target-in-the-center model shows the importance of using a generative model that is accordance with nature in this application. This notwithstanding, we observed that the Halpern-Bruno modification to the substitution model provides a small, but consistent, improvement to statistical power. Further empirical studies will show whether this theoretical improvement is real.

In a second set of simulations we explicitly explore the relationship between the information content of a motif and the number of genomes required to achieve a given statistical power. We see that the number of comparative genomes required is inversely proportional to the information content of the motif. However, at an information content of 17 bits or more, we appear to reach a limiting number of genomes that is dependent on the evolutionary distance. Although in the limit we would expect extremely high information content motifs to require the theoretical minimum of two genomes, it appears that in the practical range of information contents (judged by the entries in JASPAR) that there are limits dependent on the evolutionary distance.

In our third set of simulations we show that, under the worst case scenario where a TFBS is evolving at the same rate as the surrounding promoter region, the number of genomes required increases significantly to between three and six times the number required when the motif is evolving five times slower. This result may explain to some extent the great difficulty that has been encountered in identifying TFBSs accurately, while at the same time providing an upper limit on how many genomes we need at a given distance to identify these elusive features.

In our final study we evaluate the accuracy of the model on a TFBS prediction task in yeast. The theoretical accuracy curve is reasonably close to the observed curve, assuming that the motif sites are evolving about half as fast as the neutral, background rate. Both the theoretical and empirical accuracy curves show the inadequacy of four comparative genomes with total independent branch length of 0.85 for phylogenetic motif search. This agrees with Eddy's results on the footprinting task using these same genomes, and with our observation that the statistical power of phylogenetic motif search using "low" information content motifs is similar to that of phylogenetic footprinting of features the length of the motif.

In this work, we have developed a tool for computing the theoretical power of phylogenetic motif models. We intend to use the tool to create a database of ROC-like curves for a wide variety of motifs from the JASPAR, TRANSFAC and SCPD databases. These will be provided via the web, and will provide estimates of the achievable false positive versus false negative rates for each motif for different numbers of genomes at different, fixed evolutionary distances.

Acknowledgement

JH is funded by Australian Research Council grant DP0770471. TLB is funded by NIH grant RO-1 RR021692-01.

Bibliography

Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, Feb 2003. URL http://dx.doi.org/10.1126/science.1081331.

Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500, Jul 2003.

Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A Cohen, and Mark Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, Jul 2003. URL http://dx.doi.org/10.1126/science.1084337.

Sean R Eddy. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol*, 3(1):e10, Jan 2005. URL http://dx.doi.org/10.1371/journal.pbio.0030010.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

Martin C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200, 2004. URL http://dx.doi.org/10.1093/nar/gkh169.

Debraj GuhaThakurta. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res*, 34(12):3585–3598, 2006. URL http://dx.doi.org/10.1093/nar/gkl372.

D. L. Gumucio, H. Heilstedt-Williamson, T. A. Gray, S. A. Tarl, D. A. Shelton, D. A. Tagle, J. L. Slightom, M. Goodman, and F. S. Collins. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol*, 12(11):4919–4929, Nov 1992.

A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*, 15(7):910–917, Jul 1998.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–174, 1985.

Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003. URL http://dx.doi.org/10.1038/nature01644.

W. James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002. URL http://dx.doi.org/10.1101/gr.229102.

Gabriela G Loots and Ivan Ovcharenko. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W217–W221, Jul 2004. URL http://dx.doi.org/10.1093/nar/gkh383.

A. M. Moses, D. Y. Chiang, and M. B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*, pages 324–335, 2004a.

Alan M Moses, Derek Y Chiang, Manolis Kellis, Eric S Lander, and Michael B Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3:19, Aug 2003. URL http://dx.doi.org/10.1186/1471-2148-3-19.

Alan M Moses, Derek Y Chiang, Daniel A Pollard, Venky N Iyer, and Michael B Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12):R98, 2004b. URL http://dx.doi.org/10.1186/gb-2004-5-12-r98.

Alan M Moses, Daniel A Pollard, David A Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput Biol*, 2(10):e130, Oct 2006. URL http://dx.doi.org/10.1371/journal.pcbi.0020130.

Albin Sandelin, Wynand Alkema, Pr Engstrm, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004a. URL http://dx.doi.org/10.1093/nar/gkh012.

Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–W252, Jul 2004b. URL http://dx.doi.org/10.1093/nar/gkh372.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100, Oct 1990.

Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen. PhyloGibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, Dec 2005. URL http://dx.doi.org/10.1371/journal.pcbi.0010067.

Adam Siepel and David Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. *J Comput Biol*, 11(2-3):413–428, 2004. URL http://dx.doi.org/10.1089/1066527041410472.

Saurabh Sinha, Mathieu Blanchette, and Martin Tompa. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, Oct 2004. URL http://dx.doi.org/10.1186/1471-2105-5-170.

R. Staden. Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol*, 183: 193–211, 1990.

G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004. URL http://dx.doi.org/10.1038/nrg1315.

J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611, 1999.