BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2369–2380
DOI: 10.1093/bioinformatics/btg329

Combining phylogenetic data with co-regulated genes to identify regulatory motifs

Ting Wang and Gary D. Stormo*

Department of Genetics, Washington University Medical School, St. Louis, MO 63110, USA

Received on April 3, 2003; revised on June 2, 2003; accepted on June 12, 2003

ABSTRACT

Motivation: Discovery of regulatory motifs in unaligned DNA sequences remains a fundamental problem in computational biology. Two categories of algorithms have been developed to identify common motifs from a set of DNA sequences. The first can be called a 'multiple genes, single species' approach. It proposes that a degenerate motif is embedded in some or all of the otherwise unrelated input sequences and tries to describe a consensus motif and identify its occurrences. It is often used for co-regulated genes identified through experimental approaches. The second approach can be called 'single gene, multiple species'. It requires orthologous input sequences and tries to identify unusually well conserved regions by phylogenetic footprinting. Both approaches perform well, but each has some limitations. It is tempting to combine the knowledge of co-regulation among different genes and conservation among orthologous genes to improve our ability to identify motifs.

Results: Based on the Consensus algorithm previously established by our group, we introduce a new algorithm called PhyloCon (Phylogenetic Consensus) that takes into account both conservation among orthologous genes and co-regulation of genes within a species. This algorithm first aligns conserved regions of orthologous sequences into multiple sequence alignments, or profiles, then compares profiles representing non-orthologous sequences. Motifs emerge as common regions in these profiles. Here we present a novel statistic to compare profiles of DNA sequences and a greedy approach to search for common subprofiles. We demonstrate that PhyloCon performs well on both synthetic and biological data.

Availability: Software available upon request from the authors. http://ural.wustl.edu/softwares.html

contact: stormo@ural.wustl.edu

INTRODUCTION

Finding regulatory motifs in unaligned DNA sequences is a fundamental problem in computational biology. Over the past few years, several algorithms have been developed for this purpose, including Gibbs sampler (Lawrence *et al*., 1993), Consensus (Stormo and Hartzell, 1989; Hertz and Stormo, 1999), MEME (Bailey and Elkan, 1994, 1995), ANN-Spec (Workman and Stormo, 2000), AlignACE (Hughes *et al*., 2000), Projection (Buhler and Tompa, 2002), MITRA (Eskin and Pevzner, 2002), and FootPrinter (Blanchette and Tompa, 2002; Blanchette *et al*., 2002). These algorithms can be placed in two main categories. The first is a 'single species, multiple genes' approach. When genes of a single species share a common regulatory mechanism, the task of finding their regulatory motifs can be abstracted to a 'planted motif problem' (Pevzner and Sze, 2000), i.e. finding one or more degenerate motifs embedded in otherwise unrelated random sequences. To solve this problem, algorithms usually define an objective function to measure the difference between a representation of the motif and the background and then find a representation that maximizes the function score through local search. Promoter sequences of co-regulated genes (identified by microarray profiling or chromatin immunoprecipitation) are often subjected to such algorithms. The second approach can be called 'single gene, multiple species', or phylogenetic footprinting. It relies on conservation of regulatory mechanisms across species. Background sequences are not regarded as random but as orthologous sequences linked by a phylogenetic tree. Motifs are found as unusually well conserved substrings by comparative genomic analysis.

While these algorithms are useful for understanding regulatory networks, they have limitations. For the planted motif problem, current algorithms usually perform well in finding motifs with strong conservation. When the degeneracy is high [such as 30% mismatches within a motif, a typical (14,4) problem as described in Pevzner and Sze (2000)], they often fail. Additional algorithms have been developed to solve these more difficult problems, but their performance usually drops significantly on long background sequences (Buhler and Tompa, 2002). Statistical analysis of current motif finding algorithms suggests that their performance is influenced by the distribution of mutations within a motif and the spurious occurrences of random motifs in the background, which may pose a theoretical limitation for this type of algorithm (Workman and Stormo, 2000; Buhler and Tompa, 2002; Sze

*To whom correspondence should be addressed at 4566 Scott Avenue, Campus Box 8232, St. Louis, MO 63110, USA.

et al., 2002). Phylogenetic footprinting, by contrast, ignores experimental evidence of gene co-regulation and relies on species selection. Species need to be chosen at an appropriate evolutionary distance that maintains the regulatory mechanisms and motifs without excessive background conservation.

The rapid accumulation of genomic sequences from multiple organisms and the mounting evidence from microarray gene profiles make it possible to combine knowledge of co-regulation among different genes with conservation among orthologous genes to improve motif finding. Regulatory motifs frequently reside in evolutionarily conserved regions. For example, Wasserman *et al*. (2000) found that 98% of known skeletal-muscle-specific transcription factor (TF) binding sites are confined to 19% of human sequences that are most conserved in the orthologous rodent sequences. Cliften *et al*. (2001) sequenced six *Saccharomyces* species and showed that many TF binding sites are conserved across species and align well in conserved blocks, although the blocks are often much longer than the binding sites. Aside from taking all the promoter regions of multiple genomes into a single set on which motifs are searched (Gelfand *et al*., 2000; McGuire *et al*., 2000), two methods have proven effective when one or more reference genomes are available. One is to align orthologous sequences and identify conserved regions, which are then subjected to conventional motif finders (Wasserman *et al*., 2000). This reduces the search space to the range where conventional motif finders perform well. The other approach begins by finding many motifs in one species, then confirming their existence in reference species. False positives are eliminated because a motif is accepted only when it occurs in multiple species (GuhaThakurta *et al*., 2002; Cliften *et al*., 2003). However, no improvement has been made to the motif finding algorithm *per se*, and those methods do not take optimal advantage of both types of data. A recent approach developed for motif finding in multiple yeast species does utilize both types of information together (Kamvysselis *et al*., 2003). They start with an exhaustive enumeration of conserved words, which is quite different from our approach.

In this paper, we present a new algorithm called PhyloCon (Phylogenetic-Consensus) that takes into account both conservation among orthologous genes and co-regulation of genes within a species. The key idea of PhyloCon is to compare aligned sequence profiles from orthologous genes, rather than unaligned sequences. In brief, we usually know a few genes that share a regulatory mechanism and can obtain one or several orthologous sequences for each promoter. PhyloCon first locally aligns orthologous sequences. Alignments of conserved regions, including many suboptimal alignments, are converted to sequence profiles. Then, PhyloCon compares profiles generated from different genes. It identifies the common sections between two profiles and merges them into a new profile for subsequent comparison. It essentially samples all the possible sections of all profiles and identifies the commonly shared sections, which are reported as motifs (Fig. 1). This idea is similar to the one behind Pietrokovski's LAMA search tool (Pietrokovski, 1996). He showed that aligning protein multiple-alignments increases sensitivity in searching for conserved regions. LAMA uses the Pearson correlation coefficient to score similarities between alignments, while we propose a novel, fully probabilistic scoring scheme that incorporates thermodynamics of protein and DNA binding.

This approach has several advantages. PhyloCon does not consider a single instance of a motif as a string of letters. Instead, it sees any position of such an instance as a probabilistic distribution over all possible nucleotides. Random mutations that could disrupt the significance of any copy of the motif are much less devastating to a probabilistic profile. Spurious random profiles are much less likely than spurious random motifs. Thus PhyloCon has a low false positive rate and is very tolerant of background sequence length. Extended background genomic conservation beyond the motif helps not only to reduce search space, but also to correctly align conserved motifs (Fig. 1). By saving suboptimal alignments and comparing all of them, PhyloCon reduces its false negative rate. Finally, the statistics we use to compare profiles often precisely locate the boundaries of common sections; therefore PhyloCon does not need the length of the motif a priori.

ALGORITHM

We now present three components of the PhyloCon algorithm: initial profile generation, profile comparison, and a greedy approach to combine common regions in different profiles. Figure 1A shows a diagram of the PhyloCon algorithm, and Figure 1B shows an example of finding LEU3 sites in three groups of genes from four yeast species.

Initial profile generation

We generate initial multiple sequence alignments using the previously described Wconsensus program (Hertz and Stormo, 1999). Wconsensus is a variant of the Consensus program, and was designed to find multiple sequence alignments without prior knowledge of the alignment length. It does so by maximizing the crude information content, defined as the total information content subtracting two biases: the information content expected from the aligned sequences, and some multiple of the standard deviation of the information content expected from prior letter distribution. Wconsensus was chosen because, unlike other alignment tools such as BLAST (Altschul *et al*., 1990) and CLUSTALW (Thompson *et al*., 1994), it gives many ungapped suboptimal alignments. If the real motifs are correctly positioned in any of these alignments, they will emerge during subsequent profile comparisons. However, any other local alignment tool could be used at this step with proper modifications.

Initial multiple sequence alignments generated by Wconsensus are transformed into profiles, or position specific scoring matrices. Each column is a vector of four elements,

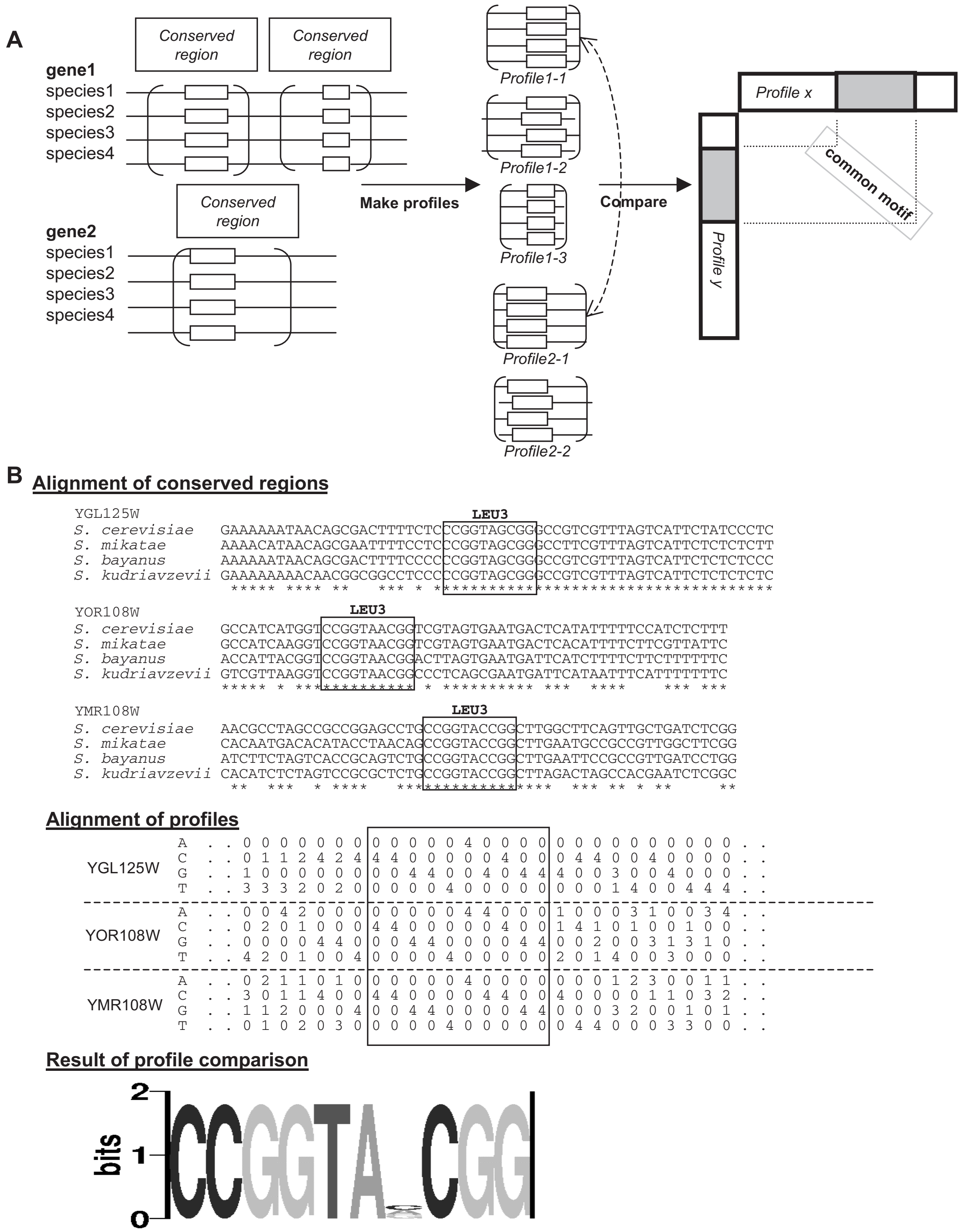
A
Conserved region
Conserved region
gene1
species1
species2
species3
species4
Conserved region
gene2
species1
species2
species3
species4
Make profiles
Profile1-1
Profile1-2
Profile1-3
Profile2-1
Profile2-2
Compare
Profile x
Profile y
common motif
B
Alignment of conserved regions
YGL125W
LEU3
S. cerevisiae GAAAAAATAACAGCGACTTTTCTCCCGGTAGCGGGCCGTCGTTTAGTCATTCTATCCCTC
S. mikatae AAAACATAACAGCGAATTTTCCTCCCGGTAGCGGGCCTTCGTTTAGTCATTCTCTCTCTT
S. bayanus AAAAAATAACAGCGACTTTTCCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCCC
S. kudriavzevii GAAAAAAAACAACGGCGGCCTCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCTC
***** **** ** *** * ************************************
YOR108W
LEU3
S. cerevisiae GCCATCATGGTCCGGTAACGGTCGTAGTGAATGACTCATATTTTTCCATCTCTTT
S. mikatae GCCATCAAGGTCCGGTAACGGTCGTAGTGAATGACTCACATTTTCTTCGTTATTC
S. bayanus ACCATTACGGTCCGGTAACGGACTTAGTGAATGATTCATCTTTTCTTCTTTTTTC
S. kudriavzevii GTCGTTAAGGTCCGGTAACGGCCCTCAGCGAATGATTCATAATTTCATTTTTTTC
***** * ************** * ********** *** **** *** ***
YMR108W
LEU3
S. cerevisiae AACGCCTAGCCGCCGGAGCCTGCCGGTACCGGCTTGGCTTCAGTTGCTGATCTCGG
S. mikatae CACAATGACACATACCTAACAGCCGGTACCGGCTTGAATGCCGCCGTTGGCTTCGG
S. bayanus ATCTTCTAGTCACCGCAGTCTGCCGGTACCGGCTTGAATTCCGCCGTTGATCCTGG
S. kudriavzevii CACATCTCTAGTCCGCGCTCTGCCGGTACCGGCTTAGACTAGCCACGAATCTCGGC
** *** * **** ***************** *** ** * ** **
Alignment of profiles
YGL125W
A . . 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . .
C . . 0 1 1 2 4 2 4 4 4 0 0 0 0 0 4 0 0 0 4 4 0 0 4 0 0 0 0 . .
G . . 1 0 0 0 0 0 0 0 0 4 4 0 0 4 0 4 4 4 0 0 3 0 0 4 0 0 0 . .
T . . 3 3 3 2 0 2 0 0 0 0 0 4 0 0 0 0 0 0 0 0 1 4 0 0 4 4 4 . .
YOR108W
A . . 0 0 4 2 0 0 0 0 0 0 0 0 4 4 0 0 0 1 0 0 0 3 1 0 0 3 4 . .
C . . 0 2 0 1 0 0 0 4 4 0 0 0 0 0 4 0 0 1 4 1 0 1 0 0 1 0 0 . .
G . . 0 0 0 0 4 4 0 0 0 4 4 0 0 0 0 4 4 0 0 2 0 0 3 1 3 1 0 . .
T . . 4 2 0 1 0 0 4 0 0 0 0 4 0 0 0 0 0 2 0 1 4 0 0 3 0 0 0 . .
YMR108W
A . . 0 2 1 1 0 1 0 0 0 0 0 0 4 0 0 0 0 0 0 0 1 2 3 0 0 1 1 . .
C . . 3 0 1 1 4 0 0 4 4 0 0 0 0 4 4 0 0 4 0 0 0 0 1 1 0 3 2 . .
G . . 1 1 2 0 0 0 4 0 0 4 4 0 0 0 0 4 4 0 0 0 3 2 0 0 1 0 1 . .
T . . 0 1 0 2 0 3 0 0 0 0 0 4 0 0 0 0 0 0 4 4 0 0 0 3 3 0 0 . .
Result of profile comparison
2
1
0
bits
CCGGTA CGG

Fig. 1. How PhyloCon works. (**A**) A diagram of how PhyloCon organizes and processes data. Sequences are grouped based on orthology. Many initial profiles are generated for conserved regions. Comparison of profiles from different orthologous groups reveals common motifs. (**B**) Alignments of orthologous sequences of four yeast species show high conservation in the 5′UTR of three genes. Asterisks indicate positions where at least three out of four letters are identical. Conservation extends beyond the true motifs (LEU3), making it difficult to identify the motif by simply examining the phylogenetic relationship. However, the motif emerges after comparing profiles from different orthologous groups.

representing either counts or observed frequencies of different nucleotides at a position in the alignment. Each profile represents a conserved region in the initial orthologous sequences. A conserved region can be represented by more than one profile based on suboptimal alignments.

Profile comparison

A profile is treated as a sequence of columns, so the alignment between profiles is analogous to the alignment between sequences. Assuming position independence and a suitable scoring scheme, the score of an alignment between two profiles is the sum of the scores from comparing corresponding columns. Since the binding site profiles are ungapped, they can be found by a simple algorithm that identifies the high scoring pairs (HSPs) along the diagonals of an alignment matrix. Therefore, the problem converts to comparing two columns.

We developed a new statistic called 'Average Log Likelihood Ratio' (ALLR) to compare two columns in different profiles. Given two probability distributions, the likelihood ratio (LR) is a standard statistic to measure the probability that the observed data belongs to one distribution versus the other. A column of a profile can be denoted as a distribution vector $f_b = \{f_\mathrm{A}, f_\mathrm{C}, f_\mathrm{G}, f_\mathrm{T}\}$, which represents the estimation of base frequencies at this position; $n_b = \{n_\mathrm{A}, n_\mathrm{C}, n_\mathrm{G}, n_\mathrm{T}\}$ denotes the observed base count at this position; $p_b = \{p_\mathrm{A}, p_\mathrm{C}, p_\mathrm{G}, p_\mathrm{T}\}$ denotes background base frequencies. Then, $\mathrm{LR} = \prod_{b=\mathrm{A..T}} (f_b/p_b)^{n_b}$ measures the likelihood that the observed base counts are from f_b rather than the background. Since f_b is a maximum likelihood estimation from n_b, LR actually measures how different f_b is from the background. Log likelihood ratio (LLR) is usually used instead of LR: $\mathrm{LLR} = \sum_{b=\mathrm{A..T}} n_b \ln(f_b/p_b)$. In practice, f_b is estimated from n_b plus some pseudocounts to reduce small sample biases.

Now we consider two columns i and j. Each column has a base count vector n_{bi} (or n_{bj}), and a base frequency vector f_{bi} (or f_{bj}) estimated from observed counts plus some pseudocounts. If these two columns are similar, the data observed for column j should fit the base distribution estimated for column i, and vice versa. Therefore, $\sum_{b=\mathrm{A..T}} n_{bj} \ln(f_{bi}/p_b)$ measures the likelihood of the observed data for column j being generated by the distribution estimated for column i. Similarly, we can perform a LLR test with observed data from column i and the estimated distribution for column j. The ALLR is given by the weighted sum of these two LLRs:

$$\mathrm{ALLR} = \frac{\sum_{b=\mathrm{A..T}} n_{bj} \ln(f_{bi}/p_b) + \sum_{b=\mathrm{A..T}} n_{bi} \ln(f_{bj}/p_b)}{\sum_{b=\mathrm{A..T}} n_{bi} + n_{bj}} \tag{1}$$

ALLR can be used to distinguish probability distributions from each other, as well as from the background. It measures the joint probability of observing the data generated by one distribution given the LR of the other distribution over the background distribution (derivation not shown). It is important to note that the expected value of ALLR is negative (substituting p_b for n_{bi} and n_{bj} in formula (1) gives negative value). Therefore, as in the Smith–Waterman algorithm (Smith and Waterman, 1981), the highest scoring local alignment is obtained by setting all negative scoring column pairs to zero and tracking back from the highest scoring column pair to the first positive pair. This procedure usually returns the correct motif end points without setting the length a priori.

ALLR is closely related to the information content used in the original Consensus algorithm and is also related to the thermodynamic properties of the protein/DNA binding system. For a given position i in a DNA binding site, the information content I_i of the sequence alignment at this position is $I_i = -\sum_{b=\mathrm{A..T}} f_{bi} H(b,i)$, where $H(b,i) = -\ln(f_{bi}/p_b)$, representing the relative entropy of base b at position i. $H(b,i)$ is a maximum probability estimate for the binding energy contribution of base b at position i, and I_i is the average binding energy of all known sites contributed by position i (Berg and von Hippel, 1987; Stormo and Fields, 1998; Hertz and Stormo, 1999; Stormo, 2000). If we let n_i and n_j denote the total base count at position i and j, formula (1) can be written as

$$\mathrm{ALLR} = \frac{-n_j \sum_{b=\mathrm{A..T}} f_{bj} H(b,i) - n_i \sum_{b=\mathrm{A..T}} f_{bi} H(b,j)}{n_i + n_j} \tag{2}$$

Positions i and j are in two different collections of binding sites. This formula is a measurement of the average binding energy of two different proteins binding to each other's functional sites. Only when f_{bi} and f_{bj} are similar is the expected binding energy negative, making the condition thermodynamically favorable. This can be interpreted to mean that the proteins accept each other's sites or that their sites are equivalent. When f_{bi} and f_{bj} are disimilar, the energy estimation will be positive on average, or thermodynamically unfavorable and the proteins reject each other's sites.

Given this scoring statistic and the assumption of position independence, the score of aligning local regions of two profiles is simply the sum of comparison scores of position pairs. A dynamic programming algorithm is implemented to identify high similarity regions using the ALLR statistic.

Profile merging

PhyloCon uses a greedy algorithm to combine profile comparison results. The following is a general description of how the alignment algorithm proceeds. Input sequences are grouped based on orthology. Let N be the number of groups, each containing a few orthologous sequences.

STEP 1. Create multiple sequence alignments for each orthologous group using Wconsensus (Hertz and Stormo, 1999). Save a user-defined number of suboptimal alignments. Convert multiple sequence alignments to profiles.

Profiles generated at STEP 1 each contain sequences from one orthologous group.

STEP 2. Pairwise compare all profiles to profiles from other orthologous groups using ALLR statistic. Save and sort some number of HSPs that exceed a threshold ALLR score. Merge profile components of a HSP into a new profile by simply summing them together, trimming off the sections not contained in the HSP. New profiles generated at STEP 2 contain sequences from two groups. They are ranked by their corresponding ALLR scores.

STEP 3. Compare each new profile saved from STEP 2 to profiles from STEP 1 that it does not already contain. Save HSPs and create new profiles, up to a user-defined number. New profiles generated at STEP 3 contain sequences from three groups.

STEP N. Compare profiles from previous cycles that do not share a common orthologous group and contain *N* orthologous groups if merged. Save HSPs and create new profiles. New profiles generated at STEP N contain sequences from *N* groups, and sorted by corresponding ALLR scores.

Each profile corresponds to a multiple sequence alignment. Whenever the alignments representing two profiles are identical, only one profile is saved. Top scoring profiles generated are reported as common motifs shared by sequences in all orthologous groups summarized by such profiles. If samples are corrupted (with only a fraction of sequences containing sites), correct results can be easily identified in profiles generated in early cycles. The program offers default values for all user defined parameters, such as the number of profiles to be saved at each cycle.

RESULTS

Scoring statistics

At least three properties of the scoring statistic make it an ideal objective function for profile comparison: (1) It is additive. The similarity score between two aligned profiles is simply the sum of the scores resulting from comparing each position pair. (2) The expected value is less than zero unless two distributions being compared are sufficiently similar (Fig. 2A), facilitating detection of the end points of an alignment. (3) Scores of HSPs generated in profile comparisons appear to follow an extreme value distribution (Fig. 2B). An HSP with a score higher than that expected from the background represents a common region shared by two profiles. For aligned profiles, this statistic gives a probabilistic estimation of their similarity; for unaligned profiles, it allows fast and accurate identification of local similarity between them with any standard sequence alignment algorithm.

Two other scoring methods have been used for the purpose of comparing two profiles. One is to use Pearson correlation coefficient as a similarity score between two profiles (Pietrokovski, 1996). The values of this statistic are between -1 and 1 and the expected value is 0 (Fig. 2A). Some constant factor (such as 0.5) should be subtracted from the values in order to make the expected value below zero. Alternatively, one can take advantage of nucleotide substitution scores and develop a 'Sum of all pairs' method (Thompson *et al.*, 1994). To align two columns is essentially equivalent to substituting all nucleotides in one column with all nucleotides in the other column. The total score is the sum of all pairwise substitution scores weighted by frequencies of all possible pairs. Such a statistic should have a negative expected value and specific substitution scoring matrices can be used for specific purposes. We plotted in Figure 2A the value distribution of all three statistics. We enumerated all possible columns from alignments of 10 DNA sequences and calculated scores of all pairwise comparisons. These scores were binned and a histogram was generated for each statistic. We also implemented all three statistics in PhyloCon and briefly compared the performance using simulated data. ALLR is clearly the best among the three, and the correlation coefficient based program has the least efficacy (see below).

Simulated data

We generated synthetic data based on a star topology model. A consensus motif was mutated and planted in a number of random sequences representing co-regulated genes in a species. For each sequence, some orthologous sequences were created to provide background homology of varying length around the planted motif. Key parameters include motif length, sequence length, number of co-regulated groups, orthologous sequences per group, background identity, and percentage of mismatches allowed between any instance of the motif and the consensus. By adjusting these parameters, we generated motif finding problems of varying complexity. To evaluate PhyloCon's performance, we compared it with several established programs, including Consensus, Wconsensus, Gibbs Sampler (Version 1.01.009) and Projection (Projection Genomics Toolkit version 0.42pre2). All programs except PhyloCon and Wconsensus were given the precise length of the planted motif. Projection was also given the number of mutations allowed in each motif as required by the program. Input sequences are organized into co-regulated groups for PhyloCon, while for the other programs they are simply pooled together. Following Pevzner and Sze (2000), we used the performance coefficient $|K \cap P|/|K \cup P|$ to evaluate the performance. K is the set of positions in known motifs in the data set, and P is the set of predicted positions. Sensitivity is defined as $|K \cap P|/K$, and specificity is defined as $|K \cap P|/P$.

Figure 3A illustrates results from experiments using five orthologous groups, each with four sequences of 600 nt. Planted motifs range from 8 to 20 nt in length with 60–90% identity, corresponding to a 40–10% mismatch rate. A 30% mismatch rate on average corresponds to a difficulty level of (14,4) to (15,5) problems (Pevzner and Sze, 2000).

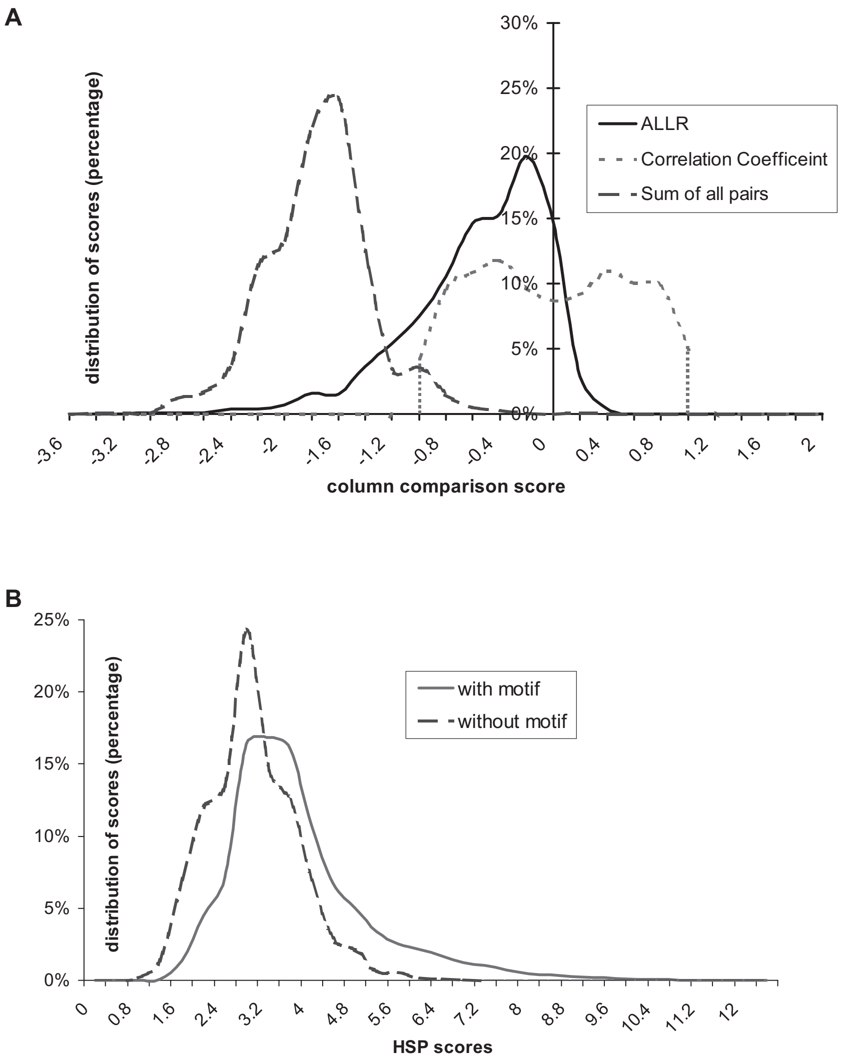

Fig. 2. (**A**) Distribution of the values of three statistics. To get these distributions, we enumerated all possible columns from alignment of 10 DNA sequences and calculated all pairwise comparison scores based on the three different statistics. Then we binned the scores based on their values and created a histogram for each statistic. Solid line: ALLR statistic, expected to be less than zero on average. Dotted line: Pearson correlation coefficient, values are between −1 and 1 and centered around 0. Dashed line: sum of all pairs, match score of 5 and mismatch score of −4 are used (BLASTN default scoring matrix). (**B**) Run time distribution of HSP scores. To get these distributions, we created two synthetic data sets, one with a motif of length 14 planted, the other with no motif planted. The motif percentage identity is 70%, and the background identity is 50%. With the program run as described, all scores of identified HSPs were recorded. These scores were binned and a histogram was generated for each case. Solid line: with planted motif. Dashed line: no planted motif. Both curves appear to fit extreme value distribution. A planted motif results in a much larger tail on the curve. HSPs in this tail usually correspond to identification of the planted motif.

Background identity ranges from 25% (no background homology) to near the motif percentage identity. When motifs have high conservation, PhyloCon performs about as well as the original Consensus. Indeed, most conventional motif finders can solve relatively easy problems with high accuracy. When the motifs are weak, containing more mismatches, the performance of all programs drops, but they do so to varying degrees. Consensus' performance coefficient drops from 0.9 to 0.5 when the percentage of mismatches in the motif increases from 10 to 35%, consistent with Sze *et al.*'s (2002) benchmark. Gibbs Sampler and Projection perform very similar to Consensus throughout the tested spectrum. Without background homology, PhyloCon performs similarly to Consensus, but introducing this factor provides a significant increase in performance. For example, for a challenging problem with about 30% mismatches in the motif,

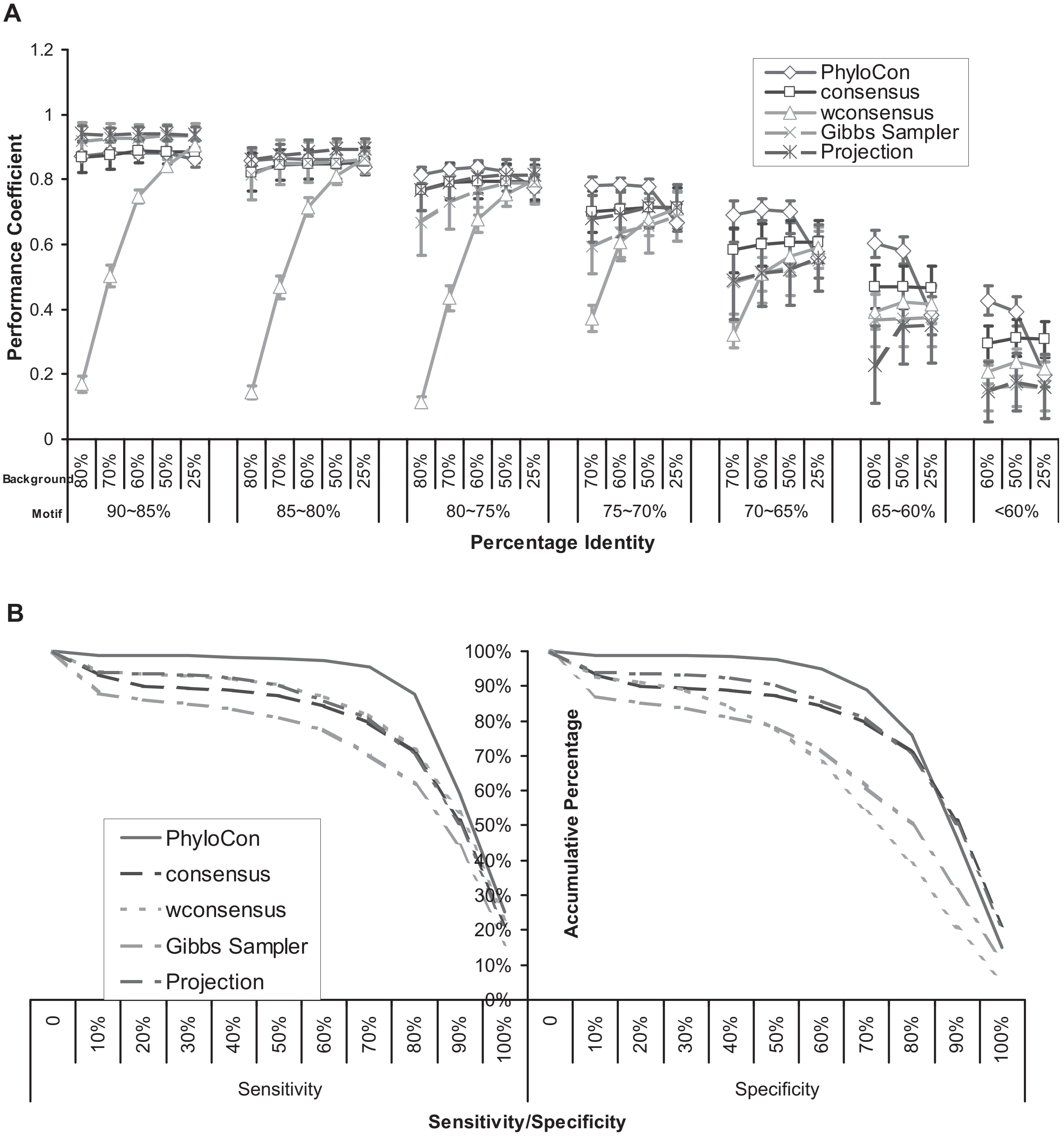

Fig. 3. (**A**) Performance coefficient comparison resulting from varied percentage identity of both planted motifs and background. Length of the planted motif is drawn from 8 to 20, and each point represents 100 independent experiments. (**B**) Plotting sensitivity and specificity separately. Ninty percent of PhyloCon's predictions achieve 80% sensitivity, and 80% of PhyloCon's predictions achieve 80% specificity.

a 50–70% background identity makes PhyloCon 20% better than Consensus. Many biological observations suggest that related sequences do provide this level of background information. Cliften *et al*. (2001) sequenced genomes of six yeast species and observed about 60% identity in the non-coding homologous regions. Human and mouse also share 50% identity in non-coding homologous regions averaging 2000 nt in length (Shabalina *et al*., 2001). PhyloCon demonstrates a great potential to increase the accuracy of motif searching when such biological phenomena are considered. In all these experiments, Consensus, Gibbs Sampler and Projection were given the motif length, while PhyloCon detected the length by itself. Based on our definition of sensitivity and specificity, for each prediction we know exactly what sensitivity and specificity values are. We plotted in Figure 3B sensitivity and specificity values separately. In this case, we examined all the predictions from the data set in which motifs have about 25–30% mismatches and backgrounds have about 50–60% identities. For each independent run on the simulated data, we calculated the percentage of the predictions that exceeds a certain level of sensitivity and specificity. As shown in the figure, about 90% of PhyloCon's predictions achieve 80% sensitivity, and 80% of the predictions achieve 80% specificity. PhyloCon has both good sensitivity and specificity, indicating its ability to accurately detect the motif length. PhyloCon nearly always outperforms the other programs by both criteria.

Most conventional motif finders show decreased performance when a relatively weak motif is planted in a very long sequence (Buhler and Tompa, 2002) due to the spurious occurrence of random motifs. However, PhyloCon is able to tolerate long sequences. Figure 4A compares performance in relation to background sequence length. In this scenario, there are five groups each containing four orthologous sequences, motifs have a 25–30% mismatch rate, and background sequences have 50–60% identity. Sequence length ranges from 600 to 10 000 nt. While the performance of all the other tested programs drops as expected, PhyloCon yields a surprisingly good performance coefficient of 0.8.

The initial profile generation determines how individual motifs are positioned in the final alignment, so we examined how many orthologous sequences are needed for the initial alignment. Figure 4B shows the performance coefficient when the number of orthologous sequences in each group ranges from two to six, while the number of groups is kept at five. Clearly, more orthologous sequences make the initial profile generation more reliable, resulting in better performance. However, even with two or three sequences in each group, PhyloCon displays a performance coefficient of 0.7, better than the other four programs. Therefore, PhyloCon could be a useful tool even when there are only one or two reference genomes available.

PhyloCon employs a greedy approach to merge profiles. We examined how many related groups are needed to generate a good representation of the motif. Figure 4C illustrates the performance coefficient when the number of groups ranges from two to six, while the number of orthologous sequences in each group is constant at four. PhyloCon is able to more accurately identify motifs when additional sequences are provided. However, even with a limited number of sequences, fewer than four, PhyloCon still gives a performance coefficient of about 0.7.

In another experiment we compared the performance of PhyloCon when three different statistics (ALLR, correlation coefficient, sum of all pairs) were implemented. In this case, motifs are 14 bases long and have a 25–30% mismatch rate, background sequences are 1000 bases long and have 50–60% identity. By using ALLR statistics, PhyloCon achieved on average a performance coefficient of above 0.8 (0.82 ± 0.03). When Pearson correlation coefficient was used, we always subtracted 0.5 from it to make the expected value negative, and the performance coefficient was about 0.5 (0.51 ± 0.02). The 'Sum of all pairs' method was implemented using default BLAST scoring matrix, i.e., match score was 5 and mismatch was -4. The performance coefficient averaged at 0.7 (0.71 ± 0.04). We concluded that the ALLR statistic is the best choice for PhyloCon.

Biological data

We applied PhyloCon to several sets of biological data where multiple reference genomes are available. Some of the results are illustrated in Table 1.

We first chose to evaluate a few known bacterial transcription factor binding sites. For example, *Escherichia Coli* metJ is a repressor of all met genes except metF. Sequences of three intergenic regions in *E.coli* and their orthologous sequences in *Haemophilus influenzae*, *Salmonella typhimurium LT2* and *Vibrio cholerae* were used (K. Tan, personal communication). PhyloCon identified the top site as AGACRTCYRGACGKCTA, which is almost identical to the documented metJ site (AGACGTYYAGAYGTCY) (Robison *et al.*, 1998) except for one extra base.

As demonstrated by simulation, PhyloCon works well even with a limited number of genes or species. *Caenorhabditis elegans* genes F44E5.5, T27E4.2 and M01B12.1 are regulated by heat shock factor (HSF). Sequences of 2000 nt upstream of the translation start sites of these genes and their orthologs in *Caenorhabditis briggsae* were used to predict HSF sites. The top prediction was CTTCTAGRAS, overlapping with the consensus site TTCTAGAA (GuhaThakurta *et al.*, 2002). PhyloCon also predicted TTCTRGAASCTTCTAGAAG in an intermediate cycle. This is a conserved two-copy repeat of HSF sites shared by F44E5.5 and T27E4.2 but absent in M01B12.1. This suggests that PhyloCon provides correct predictions for corrupted samples in intermediate cycles.

We also applied PhyloCon to Wasserman's muscle-specific transcription factor data set (Wasserman *et al.*, 2000). As noted in their paper, when a Gibbs sampling approach was applied to non-coding sequence of many kilobases around the human genes, biologically meaningless patterns were produced. However, when PhyloCon was applied to this data set plus orthologous mouse genome sequences (a total of more than 100 kb), we identified CA repeats, MEF2 sites, SP1 sites, SRF sites and MYF sites in different cycles. When using subsets of genes regulated by MEF2, SRF or MYF, the correct site was always identified as the most significant motif.

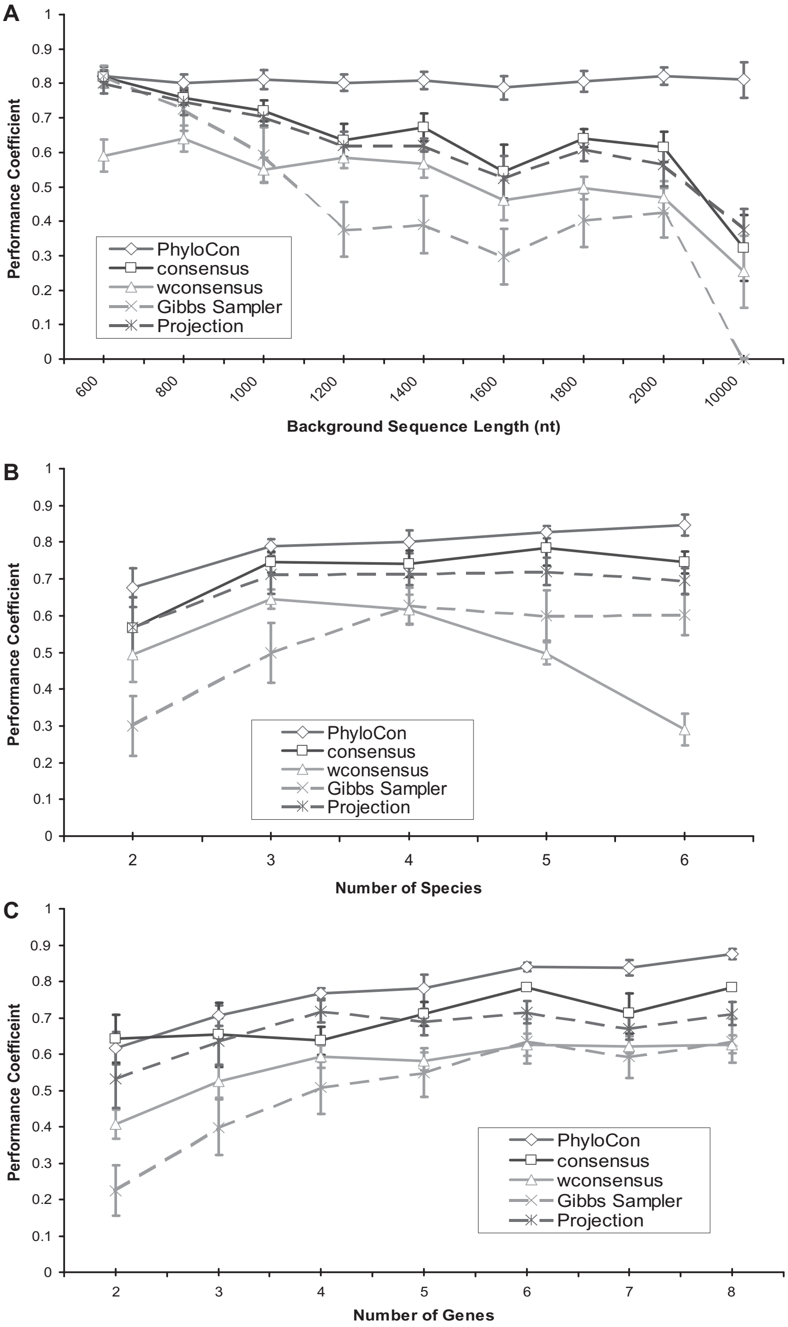

Fig. 4. Performance coefficient comparison correlating to variation of other parameters. In all experiments, motif mismatch rate is 25–30%, background identity is ∼50%, and length of motif is drawn from 12 to 14. In (**B**) and (**C**), length of each background sequence is 1000 nt. (**A**) Performance coefficient correlating to variation of length of background sequences. (**B**) Performance coefficient correlating to variation of number of species. (**C**) Performance coefficient correlating to variation of number of genes.

We obtained sequences of four yeast species of the sensu stricto group (*Saccharomyces cerevisiae*, *Saccharomyces mikatae*, *Saccharomyces bayanus* and *Saccharomyces kudriavzevii*) (Cliften *et al*., 2001). Using Lee *et al.*'s (2002) summary of literature evidence for known regulator–gene interactions, we applied PhyloCon to some of the listed groups. PhyloCon almost always identified the documented sites. It also found additional sites that may represent novel transcription factor binding sites or other important sequence features. We list a few examples in Table 1. The majority of

Table 1. Predictions of PhyloCon on biological data

Binding site	Number of genes	Number of species	Input size (nt)	PhyloCon predictions	Published reference motif
metJ	3	4	2768	AGACRTCYRGACGKCTA	AGACGTYYAGAYGTCY[a]
lexA	9	2–4	5038	TACTGTAWATRNRKACAGTA	TACTGTATWTANATMCAGY[a]
HSF	3	2	12 000	CTTCTAGRAS TTCTRGAASCTTCTAGAAG	TTCTAGAA[b]
MEF2	11	2	56 622	TATTTTTA	TATWWWWA[c]
SRF	6	2	31 838	CCATAYAAGG	CCWWANANGG[c]
MYF	10	2	37 446	GACAGCTG	RRCAGCTG[c]
CBF1	11	4	31 726	TCACGTGAR	NRTCACRTGA[d]
GCN4	11	4	27 451	TGACTC	NARTGACTCW[d]
LEU3	5	4	12 600	CCGGTASCGG	CCGGTACCGG[d]
MIG1	10	4	33 913	AAATGYGGGG	KANWWWWATSYGGGGWN[d]
ZAP1	3	4	11 170	ACCTTGAAGGT	ACCYYNAAGGT[e]

[a] Robison *et al*., 1998.
[b] GuhaThakurta *et al*., 2002.
[c] Wasserman *et al*., 2000.
[d] TransFac Release 6.2—licensed—2002-07-01, © Biobase GmbH.
[e] Lee *et al*., 2002.

Table 2. Comparison of PhyloCon's prediction to other programs

Binding site	PhyloCon	Consensus	Gibbs Sampler	Projection
metJ	AGACRTCYRGACGKCTA	CATCYRGACGKCTAAA	CRTCYAGACGKCTARA	CRTCYAGACGKCTARA
lexA	TACTGTAWATRNRKACAGTA	ACTGKATKAATATACAGTA	ACTGTATRWATAWACAGTW	TACTGTATGTATATACAGT
HSF	CTTCTAGRAS	GAGACGCA	GCAGCTSS	GAGACGCA
MEF2	TATTTTTA	TATTTTTA	GGGGTGGG	TATTTTTA
SRF	CCATAYAAGG	ACTAAAAAAA	AGAGRAGSAGA	CCATAMAAGG
MYF	GACAGCTG	AAATAAGG	GGGGWGGG	GASAGCTG
CBF1	TCACGTGAR	CACGTGRSYR	CACGTGRCYA	CACGTGRCYA
GCN4	TGACTC	GATGACTCRT	TTTYTTTTTC	TTTTTTTTTC
LEU3	CCGGTASCGG	AGAAAMGGAM	TTCTTTTTCT	TCTTTTTCTY
MIG1	AAATGYGGGG	AAATGCGGGG	AAARRAAAAA	AAARRAAAAA
ZAP1	ACCTTGAAGGT	ACCYTGAAGGT	CCTTGARGGTG	ACCTTGARGGT

this data will be presented elsewhere (Wang and Stormo, in preparation).

In order to see if PhyloCon represents a true improvement over other programs when applied to biological data, we tested Consensus, Gibbs Sampler and Projection on some of the biological data sets we gathered. In each case, the same input sequences were given to all four programs, and the motif length documented in published references was given to the other three programs. The best motif found is recorded in Table 2. The predictions of all four programs often overlap and represent the known sites. However, in many cases where one or all of the other three programs predicted apparently spurious patterns, PhyloCon predicted the documented sites. When the entire Wasserman's muscle-specific transcription factor data set (Wasserman *et al*., 2000) was provided to programs other than PhyloCon, only biologically meaningless patterns were produced. PhyloCon clearly has advantages over other tested programs when searching for relatively weak motifs in relatively long background sequences when both co-regulation and conservation data are available, in addition to its ability to predict motif length.

CONCLUSION

Conventional motif finding algorithms take either a 'multiple gene, single species' approach or a 'single gene, multiple species' approach. In this paper, we have presented a 'multiple genes, multiple species' approach that we have termed PhyloCon. PhyloCon integrates knowledge of co-regulation of genes in a single species and sequence conservation across multiple species to improve performance of motif finding. The presented data establish that PhyloCon circumvents the problem of spurious random motifs that confounds many other motif finders and also predicts the length of the motif with no prior information. We also presented a new statistic that we call the ALLR statistic. It is an extension of the LLR statistic and may be used as a general tool for hypothesis testing.

PhyloCon represents a first ‘multiple genes, multiple species’ approach, and it may be further improved. For example, the initial profile generation can be made more accurate with any approach that explicitly takes the phylogenetic tree into account. Instead of using a greedy approach comparing two profiles at a time, common profile identification can be implemented as a Gibbs sampling approach when the length of the motif is predetermined. Profile comparison can also be extended to a gapped version which will aid in identification of motifs with a flexible internal linker, as well as simultaneous identification of multiple motifs in a region.

ACKNOWLEDGEMENTS

We would like to thank P. Cliften and M. Johnston for providing sequences of yeast genomes, and K. Tan and D. GuhaThakurta for providing other sequences. Thanks to B. Cohen and S. Eddy for insightful discussions and comments that improved the manuscript. We also thank Chip Lawrence for providing the Gibbs Motif Sampler and Jeremy Buhler for providing Projection Genomics Toolkit. One unknown reviewer is thanked for bringing to our attention several relevant publications. This work is supported by HG00249 from NIH. TW is partially supported by a NIH training grant in genomic science 2T32HG00045.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Blanchette,M., Schwikowski,B. and Tompa,M. (2002) Algorithms for phylogenetic footprinting. *J. Comput. Biol.*, **9**, 211–223.

Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.

Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. *J. Comput. Biol.*, **9**, 225–242.

Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterston,R.H. and Johnston,M. (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.

Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18**, S354–S363.

Gelfand,M.S., Koonin,E.V. and Mironov,A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.*, **28**, 695–705.

GuhaThakurta,D., Palomar,L., Stormo,G.D., Tedesco,P., Johnson,T.E., Walker,D.W., Lithgow,G., Kim,S. and Link,C.D. (2002) Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res.*, **12**, 701–712.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

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

Kamvysselis,K., Patterson,N., Birren,B., Berger,B. and Lander,E. (2003) Whole-genome comparative annotation and regulatory motif discovery in multiple yeast species. *RECOMB*, **2003**, 157–166.

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

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

McGuire,A.M., Hughes,J.D. and Church,G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, **10**, 744–757.

Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 269–278.

Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.

Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.

Shabalina,S.A., Ogurtsov,A.Y., Kondrashov,V.A. and Kondrashov,A.S. (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.*, **17**, 373–376.

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

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

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Stormo,G.D. and Hartzell,G.W., 3rd (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.

Sze,S.H., Gelfand,M.S. and Pevzner,P.A. (2002) Finding weak motifs in DNA sequences. *Pac. Symp. Biocomput.*, 235–246.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, 225–228.

Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.