Reliable prediction of transcription factor binding sites by phylogenetic verification

Xiaoman Li[a,b], Sheng Zhong[c], and Wing H. Wong[a]

[a]Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065; and [c]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved October 3, 2005 (received for review May 20, 2005)

We present a statistical methodology that largely improves the accuracy in computational predictions of transcription factor (TF) binding sites in eukaryote genomes. This method models the cross-species conservation of binding sites without relying on accurate sequence alignment. It can be coupled with any motif-finding algorithm that searches for overrepresented sequence motifs in individual species and can increase the accuracy of the coupled motif-finding algorithm. Because this method is capable of accurately detecting TF binding sites, it also enhances our ability to predict the cis-regulatory modules. We applied this method on the published chromatin immunoprecipitation (ChIP)–chip data in *Saccharomyces cerevisiae* and found that its sensitivity and specificity are 9% and 14% higher than those of two recent methods. We also recovered almost all of the previously verified TF binding sites and made predictions on the cis-regulatory elements that govern the tight regulation of ribosomal protein genes in 13 eukaryote species (2 plants, 4 yeasts, 2 worms, 2 insects, and 3 mammals). These results give insights to the transcriptional regulation in eukaryotic organisms.

cross-species conservation | motif finding | ribosomal protein genes

Although significant advances have been made in the past 10 years on computational prediction of transcription factor (TF) binding sites in eukaryote genomes, the accuracy of such predictions has reached a plateau where achieving significant improvements seems difficult. Dramatically improving the state-of-the-art accuracy of TF binding site predictions, so that the *in silico* results standing alone become convincing evidence rather than mere guidance to biological tests, would significantly accelerate the pace of scientific discovery by means of computational biology techniques.

Up to now, the best sensitivity and specificity of cis-motif-finding programs are achieved by computational methods (1–6) that jointly use the cis-motif overrepresentation property in a cluster of coregulated genes[d] and the higher conservation property of bona fide TF binding sites in the regulatory regions of orthologous genes; these two properties were previously taken into account separately (7–15). Although the methods (1–6) using both of the aforementioned properties already provide useful results, they still suffer from one or more of the following limitations. Some of these methods can only be applied to two species (3); some treat orthologous sequences as statistically independently (1); some neglect the difference in the divergent time among species (1, 4, 5); and some try to find motifs in the aligned orthologous sequences and therefore require motif instances to be aligned correctly with the orthologous counterparts in the alignments[e] (1, 2, 4–6). Moreover, to our knowledge, given the many motifs output from the above methods, current practice of selecting motifs based on their *P* values or scores from these methods will make incorrect selections in >28% of the cases from yeast chromatin immunoprecipitation (ChIP)–chip data (16) (see Table 5, which is published as supporting information on the PNAS web site). This poor performance is because the *P* values or the scores of these motifs are calculated by using the same information that was used to find them.

We propose here a previously undescribed method, the cross-species conservation (CSC) method, which also jointly utilizes the motif overrepresentation property and the high conservation property. The method goes beyond the existing methods in that it finds motifs by using overrepresentation information first and then models the motifs in the context of the evolutionary divergence of neutral and functional sequences to judge their significances. Such modeling circumvents some overly simplistic assumptions used in current methods. As the results presented below demonstrate, our approach offers substantial improvement in prediction accuracy over two current methods where software is available.

Methods

Schematically, for a set of coregulated genes, CSC first uses a motif overrepresentation based method, e.g., MEME (8), with a nonstringent threshold to find potential motifs in the anchor species[f] from which the coregulated genes are obtained, as well as in other closely related species. We call the identified motifs in any species the marginally significant motifs (MSMs) of that species. Because the threshold is set to be very low, most motifs that are overrepresented in the genomic regions of the selected genes would likely be included in MSMs; i.e., we expect there to be many false-positive motifs, but the genuine motifs are likely to be included. CSC then models the evolutionary paths of the neutral intergenic regions and poses a null hypothesis that the MSMs are not functional motifs, and therefore their instances evolved like neutral intergenic regions. CSC then tests whether there are MSMs that are much more conserved in the multiple species than what are expected under the null hypothesis. CSC performs the tests by enumerating all of the groupings[g] of MSMs and calculating the probability that the current grouping of MSMs evolved from the same common ancestral motif under the null model. In the end, CSC reports the significantly conserved MSMs as putative TF binding motifs. Fig. 1 gives the overall strategy of CSC, the details of which are below.

Identification of Coregulated Genes. The coregulated genes are identified either with ChIP–chip data or microarray data for the

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: ChIP, chromatin immunoprecipitation; CSC, cross-species conservation; MSM, marginally significant motif; TF, transcription factor; TSS, transcription start site; RPG, ribosomal protein gene; *Sc*, *Saccharomyces cerevisiae*; NLC, network-level conservation.

[b]To whom correspondence should be sent at the present address: Division of Biostatistics, Department of Medicine, Indiana University, Indianapolis, IN 46202. E-mail: shawnli@iupui.edu.

[d]Coregulated genes are those that are regulated by the same TF or TF modules.

[e]*Appendix 1*, which is published as supporting information on the PNAS web site, gives an example showing that motif instances may not always align correctly.

[f]The anchor species is where the motif-finding problem arises; i.e., if we are interested in finding the motifs in a certain species, then this species is called the anchor species. We give this name to this species to differentiate it from all other species that are used to help finding the motifs (the genes from the anchor species are called anchor genes).

[g]A grouping of MSMs is a collection of similar MSMs, where each MSM in the group belongs to a different species. See *Appendix 2*, which is published as supporting information on the PNAS web site, for how to obtain groupings of MSMs.

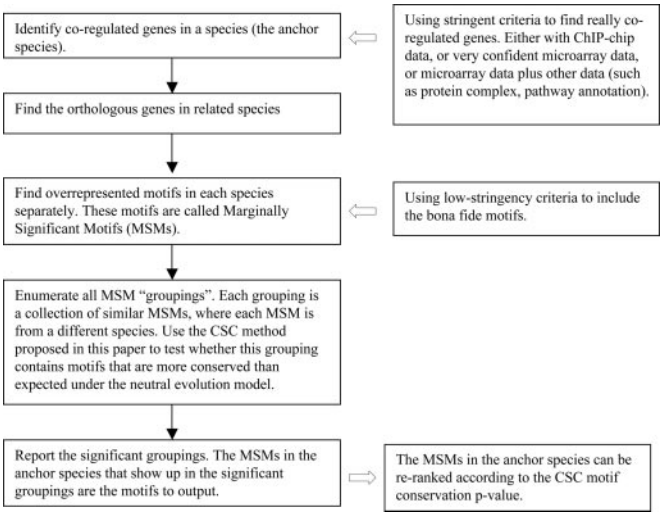

Fig. 1. CSC strategy diagram. See *Appendix 2* for the details of each step.

anchor species. For ChIP–chip data, genes that are bound by the TFs with $P < 0.001$ for binding are used (17). For microarray expression data, the following two criteria are applied. First, these genes should be clustered together across many expression profiles taken from various tissues or experimental conditions. Second, there should be sound biological reasons for these genes to be coexpressed, which could be that the protein products of such genes form equimolar protein complex, that the protein products sustain the same signal transduction pathway, or others. We apply the second criterion to ensure these genes are indeed coregulated. This criterion will not be necessary if the expression data are so informative that from them alone coregulated genes can be confidently inferred.

For the coregulated genes obtained from the anchor species, we then extract the corresponding orthologous genes in other species (see *Data* below).

Data. ***Data summary.*** We applied the CSC method on two sets of data. The first set contains ChIP–chip data for 53 TFs in *Saccharomyces cerevisiae* (*Sc*), for which we have at least 5 candidate target genes ($P < 0.001$) (16–18), and we know the experimentally verified motifs. The second data set is the collection of ribosomal protein genes (RPGs) in 13 eukaryote species (2 plants, 4 yeasts, 2 worms, 2 insects, and 3 mammals).

Microarray data for RPGs. We have observed the coexpression of RPGs in the analyses of many expression profiles. Here we exemplify this phenomenon with the mouse neural differentiation data (19), the mouse time course profiles in liver and heart (20), and the profiles of preimplantation mouse embryos (21). We used DCHIP (22) software to make hierarchical clustering on genes in every data set (see Fig. 4, which is published as supporting information on the PNAS web site).

Protein complex and genomic organization. The ribosome is a large protein–RNA complex present in bacteria and all eukaryote species. Fifty to 90 RPGs were identified in various species. Because all ribosomal proteins contribute to the formation of the same protein complex, it is expected that their production is governed by the same machinery in transcription, translation, or protein modification stages, or in some combination of these. In bacteria, RPGs have remarkable patterns of gene order conserved across 2 billion years of evolution (23). In *Sc*, the RAP1 TF has been shown to work on several promoters of RPGs (24). Taking the sources of information together with our observation in mammalian gene expression profiles, we hypothesized that the RPGs are likely to share some cis-regulatory signal in each species.

Sequence data. The upstream sequences for all of the genes in the ChIP–chip data set were obtained from the "phylogenetic footprinting" web site (25), which uses *Saccharomyces* Genome Database's (www.yeastgenome.org) sequence data.

We collected the RPG IDs for *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* from the Kyoto Encyclopedia of Genes and Genomes database (26) and downloaded their 800-bp upstream sequences of the transcription start sites (TSSs) from Ensembl (www.Ensembl.org). With the gene sequences in *C. elegans* and *D. melanogaster*, we used Ensembl to look for homologous genes in *Caenorhabditis briggsae* and *Anopheles gambiae*. In case that one RPG has multiple homologous genes, we manually checked the literature to nail down the genuine orthologous copy. If we could not find proper information about orthologs in literature, reciprocal best hits are assumed as the orthologs. We then retrieved the 800-bp upstream sequences of TSSs of these genes from Ensembl. For *Oryza sativa*, we downloaded the genome sequence and the annotated gene sequences from TIGR (www.tigr.org/tdb/e2k1/osa1) and then aligned the annotated gene sequences with the genome sequence to identify the TSSs and obtained the upstream 800-bp sequences. For human, mouse, and rat, we obtained all of the RPG LocusLink IDs from the National Center for Biotechnology Information, then used EZRETRIEVE 2.0 (27) to obtain the 1,000-bp upstream and 200-bp downstream sequences of the TSS of every gene. For yeast species *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, and *Sc*, we used the upstream sequences of translational start sites of the RPGs defined by Cliften *et al.* (25). The number of RPGs we obtained for every species is listed in Table 6, which is published as supporting information on the PNAS web site.

Discovery of MSMs in Each Species. For the gene neighborhood sequences obtained above in every species, we used REPEATMASKER (http://ftp.genome.washington.edu/RM/RepeatMasker.html) to mask the repeats within them and then used MEME (8) to search for overrepresented motifs among them. MEME uses the expectation maximization algorithm to detect motifs that have enriched instances in the input sequence sets compared with the genomic background. The current implementation of MEME enumerates all potential motif seeds (28), which minimizes the drawback of potentially missing good motifs for the reason of computational limitations. We set the parameters so that MEME would report motifs, which are also called MSMs, of 6 to 14 bp in length and *E*-value statistic of <1E8.

Note that the methodology we are about to describe does not rely on MEME. MEME is a tool used to search the overrepresented motifs in each species, and it can be replaced by any other method with the same purpose (see refs. 10, 11, etc.).

Determination of the Significance of the MSMs in the Anchor Species. If an MSM, identified by means of overrepresentation in the anchor species, is a bona fide motif, it is likely that it also will be identified in other species by means of overrepresentation; moreover, motif instances for the corresponding bona fide motifs in different species should often occur simultaneously and have high conservation if the species are properly chosen and the motifs are conserved. Thus, a motif in the anchor species is determined to be bona fide if the conservation of the motif instances far exceeds the conservation of the overall orthologous upstream sequences[h] of the coregulated genes used. The following five-step procedure is applied to select MSMs of the anchor species based on the motif conservation *P* value defined below.

Step 1: Modeling the evolution of neutral sequences. We first align the upstream sequences of orthologous genes and obtain well aligned

[h]Although we use "upstream sequence" in describing the method, in practice, the method should be applied to any regions that may contain cis-regulatory elements.

Table 1. Example of the base substitution matrix to describe the evolution of neutral nucleotide from the common ancestor of *C. elegans* and *C. briggsae* to *C. elegans*

	C. elegans			
Ancestor	A	C	G	T
A	0.7529	0.0369	0.1457	0.0645
C	0.0622	0.6522	0.0373	0.2483
G	0.2426	0.0369	0.6560	0.0645
T	0.0622	0.1418	0.0373	0.7587

For instance, the first number 0.7529 means the probability that the nucleotide A in the common ancestor of the two worm species evolved into an A in *C. elegans* is 0.7529.

regions.[i] Based on these well aligned regions and by using the species tree as the phylogenetic tree, we estimate the branch lengths of the phylogenetic tree as well as the background nucleotide distribution of the ancestral species at the root of the tree by maximal-likelihood estimation (29). Then, for every branch of the phylogenetic tree, a 4×4 base substitution matrix is calculated from the background nucleotide distribution and the branch lengths (30).

A base-substitution matrix for one branch obtained in this way, the entries of which give the probabilities of a neutral nucleotide in the parent species evolving into a base in the child species, is a measurement of the divergent time of the species after speciation on the branch in the way that the divergent time is shorter if the matrix is more similar to the 4×4 identity matrix. Table 1 gives an example of a base-substitution matrix.

Step 2: Finding all groupings of MSMs. We first find similar MSMs in two species by pulling out the motif instances of both MSMs and then aligning the two sets of motif instances without gaps (see *Motif Profile Alignment* in *Appendix 2*). A similar MSM pair is called a grouping of MSMs, which is stored and then judged on whether it contains a bona fide motif by the motif conservation *P* values calculated in Step 3 and 4 below. If a grouping of MSMs is believed to potentially contain a bona fide motif (has a motif conservation $P < 1 \times 10^{-9}$),[j] it will be considered as a new MSM and will be compared with all MSMs in species other than the two where the new MSM comes from. That is, groupings of MSMs containing motifs from two species are constructed first; then some good groupings of MSMs (those with motif conservation $P < 1 \times 10^{-9}$) are expanded by adding a similar MSM from another species to form groupings of MSMs containing motifs from three species, and so on. The details of finding all groupings of MSMs are described in Steps 2a and 2b by using the four species *Sc*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* as an example (*Sc* is the anchor species).

Step 2a: Constructing groupings of MSMs from Sc *and another species.* The motif instances of every MSM in each species are pulled out from MEME output and aligned as a profile. For every MSM in *Sc*, we will align its motif instance profile with that of every MSM in other species, and then we will store the pair of MSMs as a grouping of MSMs if the alignment of the two motif instance profiles is good. The detailed procedure of constructing the motif instance profile, the profile alignment and the criteria of good alignment are in *Motif Profile Alignment* in *Appendix 2*.

Step 2b: Constructing groupings of MSMs consisting of MSMs from Sc *and two or more other species.* For every grouping of MSMs from Step 2a, the *P* value that the putative motifs in the grouping of MSMs are not bona fide motifs is calculated in Step 3 below. If $P < 1 \times 10^{-9}$, this grouping of MSMs will be stored as a new MSM and the above Step 2a is applied to this new MSM. For instance, if this new MSM consists of motifs from *Sc* and *S. mikatae* currently, we will compare this new MSM with every MSM in the species *S. kudriavzevii* and *S. bayanus*, as is done in Step 2a, to form groupings of MSMs consisting of three MSMs from three different species. Note that the new MSM will not be expanded anymore if it contains MSMs from all different species under consideration (in our example the species number under consideration is four).

In summary, at the beginning, we will find all groupings of MSMs containing motifs from only two species; then the *P* values of each of the groupings of MSMs is calculated and the MSM groupings with $P < 1 \times 10^{-9}$ will be stored as new MSMs; then the new MSMs generated just now will be compared with MSMs from a different species to form a grouping of MSMs consisting of motifs from three species. Groupings of MSMs consisting of motifs from more species can be obtained in a similar fashion.

Step 3: Inferring the ancestral motif instances and the ancestral motif weight matrix for every grouping of MSMs. For a grouping of MSMs obtained at Step 2, no matter if it contains MSMs from two, three, or more species, the corresponding part of the species tree will be taken out as the phylogenetic tree for this grouping of MSMs, i.e., this phylogenetic tree will only have n leaves and $n - 1$ other nodes if the grouping of MSMs contains MSMs from n species. Based on this phylogenetic tree, we construct ancestral motif instances by using maximal parsimony for every orthologous gene group in which at least one motif instance was obtained from MEME (8) output. An ancestral weight matrix is subsequently compiled from the obtained ancestral motif instances. Then, every obtained ancestral motif instance is scored with this ancestral weight matrix, and the 20% quantile of these scores is defined as the threshold of scores of real motif instances of the ancestral motif. The details are illustrated in *Motif Profile Alignment* and *Constructing Ancestral Motif for One Grouping of MSMs* in *Appendix 2*.

Step 4: Assessing the significance of every grouping of MSMs. For any grouping of MSMs, CSC uses the following procedure to assess the statistical significance of the test statistic, the number of orthologous gene groups with at least two of the orthologous genes (must include the gene in the anchor species) containing real motif instances of the ancestral motif[k] (see Fig. 2 for the construction of the test statistic). The significance is represented by the motif conservation *P* value, which is the tail probability of the distribution of the test statistic under the neutral evolution model.

Note that all of the ancestral sequences are random variables with the same nucleotide distribution estimated from the current sequences by using maximal-likelihood estimation (29). Thus, the test statistic is a random variable, say X, when given the phylogenetic tree, the base substitution matrices, the ancestral motif weight matrix, and the threshold for a segment to be a real motif instance of the ancestral motif. The distribution of X can be derived by assuming that the ancestral sequences evolved into the current sequences by following the above estimated base-substitution matrices (Step 1).

In *Calculate the Motif Conservation P Value for a Grouping of MSMs* in *Appendix 2*, we show how to calculate the motif conservation *P* value $\Pr(X > x)$, where x is the observed value of the test statistic X. This motif conservation *P* value indicates the chance of observing at least as many orthologous gene groups that contain

[i]For the yeast species, we downloaded alignment of upstream orthologs from ref. 25. For the two plant species, we did local alignment of orthologous upstream and used the best-aligned regions of 100-bp length for every orthologous pair. The 100-bp cutoff is arbitrarily chosen, but, to our knowledge, our method is not so sensitive to the background-substitution matrices. For the three mammalian species, two insect species, and two worm species, we download the available alignments of the RPG upstream sequences from University of California, Santa Cruz genome browser web site.

[j]This cutoff is arbitrary. From our experience, this cutoff works well for all the data sets from different species we used. In the text, we have another empirical *P* value cutoff, 1×10^{-19}, which is used to report motifs.

[k]To avoid overfitting, we exclude the motif instances on the current group of orthologous genes (the group of genes to be scanned by the ancient motif) from constructing the ancient motif.

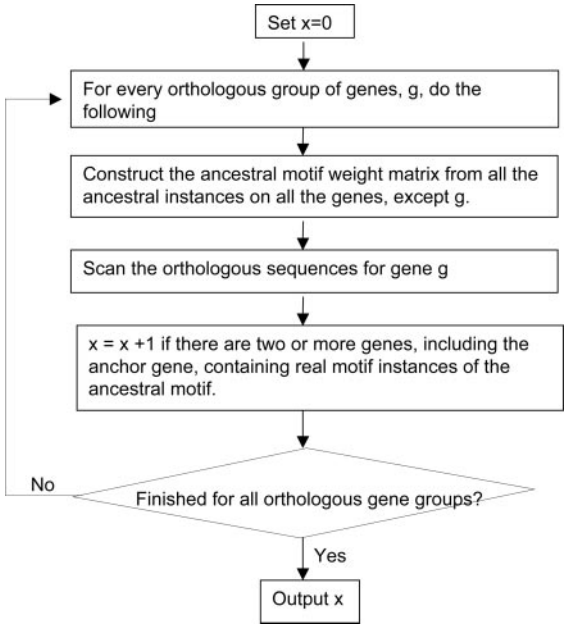

Fig. 2. The construction of the test statistic for testing whether a group of MSMs are derived from the same ancestral motif.

conserved real motif instances of the ancestral motif in multiple species, including the anchor species, as we have actually observed. ***Step 5: Selecting MSMs and reporting their significance.*** We enumerate all of the MSM groupings and apply the procedure in Step 4 to calculate the motif conservation P value to judge whether the current grouping of MSMs contains a bona fide motif that is derived from a common ancestral motif. In general, if the P value of a grouping is $<1 \times 10^{-19}$, the motif from the anchor species in this grouping will be reported as a bona fide motif (see *Basic Procedure of Predicting Motifs* in *Appendix 2* for the detailed criteria of reporting MSMs that are considered as bona fide motifs).

Results

Results for the ChIP–Chip Data. For every target gene set of the 53 TFs, we applied CSC as well as two recently published methods (1, 5) to predict the putative motifs. We compared CSC with these two methods because they were considered the state-of-the-art motif-finding methods and have available software. For the comparison, we used the experimentally verified motifs recorded in the TRANSFAC database (31) as the gold standard motifs. Table 5 lists all of the predicted motifs by the three methods and the known motifs in the TRANSFAC database (31) for each of the 53 TFs. We gave both PHYLOCON (1) and COMPAREPROSPECTOR (5) human help by doing the following. In the list of their output motifs,[l] we manually searched for the motif that best matched the recorded motif in TRANSFAC. As long as one motif found by PHYLOCON (or COMPAREPROSPECTOR) matches the known one in TRANSFAC well,[m] we agreed that PHYLOCON (or COMPAREPROSPECTOR) predicted the correct motif. Conversely, no manual intervention was given to CSC; we only counted the motifs identified by CSC as correct if the top reported motif was correct. CSC correctly reported the motifs of 30 TFs (Table 5). Both COMPAREPROSPECTOR and PHYLOCON predicted correct motifs for 25 TFs. CSC made 5 incorrect predictions of 35 predictions, whereas COMPAREPROSPECTOR and PHYLOCON made 21 and 10 incorrect predictions of 46 and 35 predictions, respectively. From the comparison, we can see that even with the advantage we gave to COMPAREPROSPECTOR and PHYLOCON, CSC significantly outperformed them (see Table 2). Besides the 35 predictions, CSC did not make predictions for the remaining 18 TFs; in 16 of those, the known motif is not included in MEME output, and in the remaining 2 cases the known motif is ranked as the top 1 motif, yet their P values were $>1 \times 10^{-19}$ because of the small number (<5) of target genes containing motif instances. Notice that in all MEME did not find the known motifs for 17 TFs, which may be due to noisy target gene sets, because the other 6 independent methods did not find the known motifs in 9 of the 17 cases either (18). Moreover, the incorrect predictions in the 5 cases are not necessarily incorrect, although the top reported motifs do not match the known motifs associated with those TFs. For instance, the top motif predicted by our method for CIN5, TGCGGTGTGTGGGT, occurs in 127 different CIN5 target genes and has motif instances in at least two species in the 127 genes, whereas the motif provided by the literature occurs in 122 genes with only 88 genes containing motif instances in at least two species, by allowing one mismatch with the known CIN5 consensus, TTACATAA.

Table 2. Comparison of sensitivity and specificity

Trait	CSC (ours)	COMPAREPROSPECTOR	PHYLOCON
Sensitivity	(29 + 1)/53 = 56.6%	(24 + 1)/53 = 47.2%	(24 + 1)/53 = 47.2%
Specificity	(29 + 1)/35 = 85.7%	(24 + 1)/46 = 54.3%	(24 + 1)/35 = 71.4%

+1 means that one motif predicted for SUM1 looks similar to the corresponding experimentally verified motif in TRANSFAC, although they do not satisfy our criteria of correct prediction.

Very recently, Harbison *et al.* (18) used six different computational methods to look for candidate motifs from ChIP–chip data. With the help of the prior knowledge, they made educated guesses to pick one motif from the reservoir of candidate motifs. This human-intervened fine tuning may achieve the highest imaginable accuracy, because a large amount of information from different sources has been used manually. The obvious drawbacks of this method are that it cannot be systemized, and it is not easy to redo the exercise for other researchers on other species. Even this approach combined with human intervention may not necessarily give better results as compared with CSC (see Table 5, legend). Moreover, CSC can report multiple motifs for one set of genes because it will assume all of the MSMs with P values less than some threshold to be bona fide motifs. Therefore, cis-motif modules, which consist of multiple binding sites and attract cooperative TFs, are often evident from our result (Table 3).

Results for RPGs. CSC reported a motif in each of the four yeast species (TACATCCGTACATT for *Sc*) with the motif conservation $P = 1.1 \times 10^{-120}$. Their consensus sequences were almost identical both among themselves and to the previously known RAP1 binding sites in *Sc* (see Table 6). Motif instances were found in almost every upstream sequence. Because all biologically verified RAP1 binding sites in *Sc* were included in the predicted motif instances in the TRANSFAC database (31), it is not unreasonable to assume that most of these predicted motif instances are bona fide RAP1 binding sites. Several interesting observations were made to them:

1. Their relative distances to translation start sites are conserved, ranging from 200 to 500 bp.
2. The motif instances on upstream sequences of orthologous genes appear in the same strand in most cases (399 of 433 pairs).

[l] PHYLOCON on average outputs 60 predictions with some redundancies. COMPAREPROSPECTOR outputs ranked ordered motifs, for which we performed the manual search within the top three motifs.

[m] The criterion for matching with TRANSFAC motifs is that there should be at most one mismatch when we compare the putative motifs with the TRANSFAC ones (see the legend of Table 5 and the supplementary files of ref. 18).

Table 3. Predicted cis-regulatory modules by CSC

TF	Motif 1	Other motifs	Cooperative factor
ABF1	ATCACTATATACGA(ABF1)	CTGAAAAATTTTCG	UME6,
		CGGCGGCAATT(UME6)	Unknown
FKH1	GCCGTTGTTTACG(FKH1)	CCCTGGCGCGTCTT	Unknown
GCN4	ATGACTCAGC(GCN4)	CGGGACCGGCTCTG	Unknown
HAP4	GCGGGCCAATCAGA(HAP4)	TTCCCGTCCTAAT	Unknown
MBP1	ACGCGACGCGT(MBP1)	GCGTGGGCCCTCCT	Unknown
		CGTCTTGCCTACAC	
MCM1	CCTAATAAGGAAAT(MCM1)	GGCGGCTAAAAATA	Unknown
RAP1	TACACCCATACATC(RAP1)	TTCGGTTTCCTTC(GCR1)	GCR1
STE12	TGAAACAA(STE12)	AAGAAAAAGCCGCC	Unknown
SUM1	TATTTACTGACAC(SUM1)	GCTGACGCTGTCGC	Unknown
SUT1	ATATACGTATATAT	GAAGGCACAGT(SUT1)	Unknown
SW16	GGAAACGCGACGCG(SWI4)	CGCGAAAGACC(MBP1)	MBP1,SWI4
		TTCCCTTTTCGGAA	Unknown
UME6	CTTCGGCGGCTAAT(UME6)	GGAAGAAAAGAAAG	Unknown

The first column gives the TF on which the ChIP experiment was performed. Motif 1 gives the most significant motif identified by CSC. Other motifs include all the other motifs identified by CSC. They may form cis-regulatory module with the first motif. Cooperative factor is factor that can bind onto the other predicted motifs.

3. Motif instances on the orthologous upstreams tend to be more similar than motif instances on different genes of the same species.

In each of the two worm species (*C. elegans* and *C. briggsae*), MEME (8) reported 12 motifs. Among them, 6 pairs of motifs pass our conservation threshold. *Appendix 3*, which is published as supporting information on the PNAS web site, gives the MSMs and their CSC *P* values.

In insects, 8 mosquito motifs and 11 Drosophila motifs were found by MEME. One significant grouping was reported by CSC (*Appendix 3*). The most conserved pair was

Drosophila: GCGGTCACAct
Mosquito: gcaGCTGTCAAAtg.

We found the two consensuses shared seven bases from a core of nine bases. Even though the consensuses did not appear to be extremely similar, CSC reported their conservation *P* value as 1×10^{-20}. This result is because CSC automatically adjusted for the fact that the local conservation of the upstream sequences of the two species is low, because of the 250-million-year divergence of the two insect species. We observed that the motif instances had relatively conserved loci relative to the corresponding TSSs. Interestingly, the motif consensus in *D. melanogaster* ribosomal genes matched an experimentally verified binding site, CAGTCACA, in *Schizosaccharomyces pombe* (32). This result indicates that other than the RAP1 TF, there is another TF that also regulates RPGs in yeast but does not play as important a role as RAP1. However, it seems that the motif in *S. pombe* has taken over the importance in insects.

The two plant species *A. thaliana* and *O. sativa*, respectively, had seven and six MSMs, and one grouping of MSMs passed our conservation criterion (*Appendix 3*). The two motifs in this grouping have consensus ATTAGGGTTTT (*A. thaliana*) and GCTAGGGTTTC (*O. sativa*), which are more conserved compared with metazoan ones, in the senses of both the resemblance among the motif instances and the conserved relative locations of the motif instances relative to the corresponding TSSs. One instance of this identified motif in *A. thaliana* has been experimentally verified in the upstream sequence of an *A. thaliana* gene (33).

Human, mouse, and rat had 5, 7, and 10 MSMs, respectively. CSC reported 4 significant groupings (Table 4 and *Appendix 3*). The motifs that constituted the significant groupings are the CSC-reported motifs (referred to as "motifs" hereafter). The instances

Table 4. The consensuses of the MSMs in three mammals

Marginal order	Human	Mouse	Rat
1	ATCCGCCGCCATCC	ATCCGCCGCCATCC	TCTTCCTTTCC
2	CCAACATGGTGAGT	TCCCTTCCTCTTCC	ATCCGCCGCCATCC
3	AAATATTTTTTAAA	GTTCTCGCGAGAGC	CAGTGCGCATGCGC
4	AATCTCGCGAGAAC	CAAGATGGTGAGTG	AAACTACAATTCCC
5	AACTACAATTCCCA	TGGGAATTGTAGT	TCCGCCATCTTCG
6		GGGGTGAGGGGGAG	TATTTTAAAAA
7		GGAAGCCAAGGCC	GTAGGGGCGGGGCT
8			GCAGCCGCCAAGGT
9			CACCATGGTAAGTG
10			ACTTAAAACATTTT

The MSMs are computed and ordered by the MEME program. The colored consensuses represent the motifs with significant CSC motif conservation. *P* values (1×10^{-19}). The motifs with the same color form a significant grouping.

of the most significant motif (ATCCGCCGCCATCC) in mouse have been shown to have regulatory function on mouse genes *RPL32*, *RPL30*, and *RPS16* (34). Our finding shows that the positions at the two ends of the motif are more important than the middle, contradicting the assertions made by the authors (34). In human, there are 78 RPGs containing at least one instance of this motif. Many instances of this motif in three mammalian species are located in the first intron. The genes with the motif instances in the first intron usually have a noncoding first exon. Guofu Hu *et al.* at Harvard University tested our predicted motifs with luciferase reporter assays in HeLa cells. They inserted each predicted element into a pGL3P vector that contained a SV40 promoter. They found two of our predicted motifs not mentioned in literature before (CCAACATGGTGAGT and AATCTCGCGAGAAC) having enhancer activities (Guofu Hu, personal communication). Taken together, of the four motifs that CSC reported in the three mammalian species, one has been reported in literature, and two have been verified *in vitro*.

Discussion

Many software programs can produce an exhaustive list of putative motifs, but determining which motifs are bona fide is quite difficult and is an urgent topic requiring extensive research. A routine strategy is to choose the top one or few putative motifs with the smallest *P* values or best scores. This idea does not work in >28% cases by applying two recent methods to 53 gene data sets (see Table 5, legend). We have similar findings with other software on yeast coregulated genes.

A fundamental problem consistently found in recent methods is that the motifs are identified and evaluated with the same information, i.e., the *P* values or scores output from those methods are biased to be used for selecting motifs. In this work, we use overrepresentation information to find motifs first and then use the independent information of conservation to verify the found putative motifs. Preliminary results show our method is much better in determining the bona fide motifs. Two aspects contribute to the better performance. First, using MEME (8) to find motifs by lowering the threshold is more likely to include all bona fide motifs. It is true that methods based on the expectation maximization algorithm may be trapped by local optima, but this local optima phenomena is greatly minimized by enumerating every *w*-mer (for $w = 4, 8, 16, 32$) in the input sequences, then running the expectation maximization algorithm starting at each *w*-mer. Then, by picking the *w*-mers for each value of *w* that have high likelihood ratios to continue, the computations converge to local optima (28). Thus, genuine motifs are most likely included if we can lower the threshold enough. Second, we model the neutral evolution by using a base-substitution matrix for every branch. These matrices reflect the differences of divergence time. By using them, on one hand, some overrepresented and

conserved instances may not be statistically significant because the species are very close; on the other hand, some not well conserved motifs may be picked out because they are more conserved than the average background sequences.

We constructed the base-substitution matrices on every branch by using alignments from well aligned regions. Note that there is no contradiction with our belief to try not to align orthologous sequences first and then find motif instances from the aligned sequences. The rationale is as follows. The background segments are long, whereas the motif instances are short. In general, it is much easier to identify longer regions that really should be aligned than to judge the correctness of short segments like motif instances in an alignment. Moreover, the base-substitution matrices are not required to be very accurate. Our experience shows that CSC is robust to the perturbations of the substitution matrices. Of course, we also can estimate the base-substitution matrices from other sources such as synonymous sites in the corresponding protein sequences or ancient repeats in the neighborhood regions.

When the first version of CSC was implemented in 2004, we noticed that Pristker *et al.* (35) proposed a similar method, network-level conservation (NLC). NLC pools together upstream sequences from two different species and identifies candidate motifs in the pool using motif overrepresentation-based methods. Then, NLC screens candidate motifs by assuming that the distribution of the number of the orthologous upstreams containing the motif in two species follows a hypergeometric distribution. The difference between NLC and CSC is that CSC can be applied on multiple species, whereas NLC can be used only on two species; CSC models the evolution of motif instances by taking the divergence time into account to determine motif significances, whereas NLC neglects the divergence time and uses a hypergeometric distribution to calculate motif significances; and CSC assumes orthologous sequences evolved from the common ancestral sequence, whereas NLC assumes orthologous sequences are independent. In our opinion, CSC not only can select motifs from very distant species such as fly and mosquito, which cannot be done by NLC, but also is more powerful than NLC in selecting motifs in close species.

A recent method (36) that also used the phylogenetic trees and the base-substitution matrices to find motifs can find very degenerate motifs. Nonetheless, the highest ranking motif found by that method is not guaranteed to be a bona fide motif, and is it not clear exactly how many of the output motifs are indeed genuine. On the contrary, CSC may not be able to find very degenerate motifs, but it can rank the motifs well and can choose multiple correct motifs from any software output (see Table 5).

The comparison of conserved motifs in RPGs among close species using CSC suggests two things. First, there are some dominant TFs regulating the RPGs in each species. Second, the TFs may be different for different species, although the ribosome has the same function across all species. These two suggestions extend our understanding of the transcriptional regulation of RPGs. Although the RPG sequences and functions are highly conserved across species, the cis-elements and probably the TFs behind them are different.

The comparison of conserved motifs in RPGs in 13 species by using CSC (see Fig. 5, which is published as supporting information on the PNAS web site) also sheds light on the evolutionary properties of cis-regulatory motifs. The three mammalian species, two worm species, and four yeast species are close enough to share exactly the same dominant elements, whereas the two insects, separated by ≈250 million years, share less similarity in the dominant sites. Conversely, the two plants, far from each other (>250 million years), share exactly the same dominant site. These data suggest the evolutionary differences of cis-regulatory motifs not only depend on the divergent time, but also on the species genus. This understanding may help refine current phylogenetic footprinting methods.

X.L. thanks Arthur Berg at the University of California, San Diego for his help on the writing and insightful suggestions. This work was supported by a National Institute of General Medical Sciences Grant GM67250 (to W.H.W.), and the computation in this work was supported by National Institutes of Health Grant NIH R01-HG02518-01 (to J. S. Liu).

1. Wang, T. & Stormo, G. D. (2003) *Bioinformatics* **19,** 2369–2380.
2. Moses, A., Chiang, D. & Eisen, M. (2004) *Pacific Symp. Biocomput.*, 324–335.
3. Prakash, A., Blanchette, M., Sinha, S. & Tompa, M. (2004) *Pacific Symp. Biocomput.*, 348–359.
4. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. (2004) *J. Comp. Biol.* **2-3,** 319–357.
5. Liu, Y., Liu, X. S., Wei, L., Altman, R. B. & Batzoglou, S. (2004) *Genome Res.* **14,** 451–458.
6. Sinha, S., Blanchette, M. & Tompa, M. (2004) *BMC Bioinformatics* **5,** 170.
7. Stormo, G. D. & Hartzell, G. W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 1183–1187.
8. Bailey, T. L. & Elkan, C. (1994) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36.
9. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993) *Science* **262,** 208–214.
10. Hughes, J. D., Estep, P. W., Tavarzoie, S. & Church, G. M. (2000) *J. Mol. Biol.* **296,** 1205–1214.
11. Liu, X., Brutlag, D. L. & Liu, J. S. (2001) *Pacific Symp. Biocomput.*, 127–138.
12. Eskin, E. & Pevzner, P. A. (2002) *Bioinformatics* **18,** Suppl. 1, 354–363.
13. Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M. & Frazer, K. A. (2000) *Science* **288,** 136–140.
14. Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M. & Frazer, K. A. (2000) *Genome Res.* **10,** 1304–1306.
15. Blanchette, M. & Tompa, M. (2003) *Nucleic Acids Res.* **31,** 3840–3842.
16. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. R., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
17. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000) *Science* **290,** 2306–2309.
18. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004) *Nature* **431,** 99–104.
19. Bachoo, R. M., Kim, R. S., Ligon, K. L., Maher, E. A., Brennan, C., Billings, N., Chan, S., Li, C., Rowitch, D. H., Wong, W. H. & DePinho, R. A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 8384–8389.
20. Storch, K. F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H. & Weitz, C. J. (2002) *Nature* **417,** 78–83.
21. Wang, Q. T., Piotrowska, K., Ciemerych, M. A., Milenkovic, L., Scott, M. P., Davis, R. W. & Zernicka-Goetz, M. (2004) *Dev. Cell* **6,** 133–144.
22. Li, C. & Wong, W. H. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 31–36.
23. Wang, N., Chen, R. & Wong, W. H. (2000) *Sci. China Ser. C* **43,** 120–128.
24. Lascaris, R. F., Mager, W. H. & Planta, R. J. (1999) *Bioinformatics* **15,** 267–277.
25. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) *Science* **301,** 71–76.
26. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. (2002) *Nucleic Acids Res.* **30,** 42–46.
27. Zhang, H., Ramanathan, Y., Soteropoulos, P., Recce, M. L. & Tolias, P. P. (2002) *Nucleic Acids Res.* **30,** e121.
28. Bailey, T. L. & Elkan, C. (1995) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3,** 21–29.
29. Felsenstein, J. & Churchill, G. A. (1996) *Mol. Biol. Evol.* **13,** 93–104.
30. Hillis, D. M., Moritz, C. & Mable, B. K. (1996) *Molecular Systematics.* (Sinauer, Sunderland, MA), pp. 1–167.
31. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. (1996) *Nucleic Acids Res.* **24,** 238–241.
32. Witt, I., Straub, N., Kaufer, N. F. & Gross, T. (1993) *EMBO J.* **12,** 1201–1208.
33. Manevski, A., Bertoni, G., Bardet, C., Tremousaygue, D. & Lescure, B. (2000) *FEBS Lett.* **483,** 43–46.
34. Hariharan, N., Kelley, D. E. & Perry, R. P. (1989) *Gene Dev.* **3,** 1789–1800.
35. Pritsker, M., Liu, Y. C., Beer, M. A. & Tavazoie S. (2004) *Genome Res.* **14,** 99–108.
36. Li, X. & Wong, W. H. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 9481–9486.