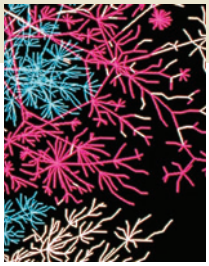

Discovery of regulatory elements in vertebrates through comparative genomics

Amol Prakash[1] & Martin Tompa[1–2]

We have analyzed issues of reliability in studies in which comparative genomic approaches have been applied to the discovery of regulatory elements at a genome-wide level in vertebrates. We point out some potential problems with such studies, including difficulties in accurately identifying orthologous promoter regions. Many of these subtle analytical problems have become apparent only when studying the more complex vertebrate genomes. By determining motif reliability, we compared existing tools when applied to the discovery of vertebrate regulatory elements. We then used a statistical clustering method to produce a computational catalog of high quality putative regulatory elements from vertebrates, some of which are widely conserved among vertebrates and many of which are novel regulatory elements. The results provide a glimpse into the wealth of information that comparative genomics can yield and suggest the need for further improvement of genome-wide comparative computational techniques.

Understanding gene regulation has been and remains one of the major challenges for the molecular biology community. Gene regulation is mediated by a variety of short DNA sequences called regulatory elements, which include transcription factor binding sites. A first step toward understanding regulation is the identification of the regulatory elements present in the genome.

The approach called phylogenetic footprinting[1] is based on the observation that regulatory elements are under selective pressure, which causes them to evolve at a slower rate than the surrounding nonfunctional sequence. Thus, by comparing orthologous regulatory regions from related species, phylogenetic footprinting predicts highly conserved sub-sequences that could function as potential regulatory elements.

The list of vertebrate genomes that have been completely sequenced will soon grow to well over a dozen, so that phylogenetic footprinting on a whole-genome scale will become an important tool in their analysis. In particular, this will yield a catalog of potential human regulatory elements, most of which are certain to be novel. Two recent studies[2,3] have applied phylogenetic footprinting on a whole-genome scale to a variety of yeast species, yielding catalogs of yeast regulatory elements. In contrast, whole-genome phylogenetic footprinting of vertebrates is a more complex proposition, as we demonstrate here through a study of some of the vertebrate genomes currently available.

Here we identify two of the main problems of applying phylogenetic footprinting to vertebrates. (i) The annotated sequence databases used to identify orthologous regulatory regions among vertebrates lack completeness and accuracy, prerequisites for phylogenetic footprinting; (ii) the tools and parameters for phylogenetic footprinting are as likely to predict equally well-conserved motifs in nonorthologous data sets as in orthologous data sets.

We introduce two methods useful for phylogenetic footprinting in the vertebrates. The first one provides an empirical approach to determining which phylogenetic footprinting tools and which parameters are likely to produce reliable motifs, and the second is a statistical method for clustering the well-conserved motifs produced, based on the Karlin-Altschul system[4].

Finally, we apply the above principles (i) to a performance comparison of existing phylogenetic footprinting tools when applied to genome-level discovery of regulatory elements in vertebrate promoters and (ii) to generate a list of motifs conserved in the vertebrate promoters, many of which are novel candidates for regulatory elements.

The phylogenetic tree relating the various vertebrate genomes used in this study (human, chimp, mouse, rat, chicken and fugu) can be seen as **Supplementary Figure 1** online. We report on three separate studies, one involving just the four mammals, a second including chicken and the last including all six vertebrates. In these studies we focus on promoter regions, but similar ideas apply to other regulatory regions, such as untranslated regions or introns.

The CORG datatabase[5] (http://corg.molgen.mpg.de/) also uses computational methods to extract conserved noncoding blocks from the upstream regions of orthologous gene pairs from human, mouse, rat, fugu and zebrafish. In contrast, we compare various tools for their ability to identify regulatory elements and choose the best one. Our motifs are extracted using multiple alignments rather than pair-wise alignments and the resulting high-quality motifs are clustered to turn them into a catalog. Another recent paper[6] also identifies regulatory elements in human and mouse using a genome-wide analysis.

Most closely related to the present work is the very recent paper of Xie *et al.*[7], in which the authors use a whole-genome multiple alignment of human, mouse, rat and dog to identify regulatory motifs. Their focus is on predicting motifs that are both conserved across these species and

[1]Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle, Washington 98195-2350, USA. [2]Department of Genome Sciences, Box 357730, University of Washington, Seattle, Washington 98195-7730, USA. Correspondence should be addressed to M.T. (tompa@cs.washington.edu).

Published online 6 October 2005; doi:10.1038/nbt1140

Box 1 Watching out for nonorthologous promoters

The first step in phylogenetic footprinting is to obtain a set of reliably orthologous promoter regions. Several idiosyncrasies of vertebrate comparative genomics make this process more difficult than with simpler genomes.

1. It is frequently difficult to identify the orthologous genes among vertebrates. To illustrate this we use the homology match defined between pairs of genomes in Ensembl[8] (http://www.ensembl.org). Ensembl's homology definition is based on computational techniques such as protein sequence similarity and synteny, and thus is subject to error. We started by listing all human genes that have an annotated homolog for every mammal in the study. Among these genes, 16% showed some inconsistency in Ensembl's homology mappings; for example, the mouse and rat homologs of a particular human gene are not annotated as being homologous to each other. This makes it unclear how to choose truly orthologous genes.

2. Once a set of orthologous genes has been determined for every species, it is common practice to extract and analyze the regions just upstream of the annotated translation start sites (e.g., ref. 6). However, we find that it is surprisingly common that the annotated translation start sites of orthologous genes are indeed not orthologous positions. This may be the result of a loss of the first exon in some species, errors in annotation or lack of experimental evidence for start sites.

To explore the extent of this phenomenon, we extracted the site in mouse that is aligned to the human translation start site in the human-mouse whole genome local alignments provided by Ensembl. Its genomic distance from the annotated translation start site of the orthologous mouse gene provides a measure of the skewness in the two translation start sites. The same experiment was repeated for all the other species (chimp, rat, chicken, fugu) by taking their whole genome local alignments against human. The histogram of these distances for each of the species is plotted in **Figure 1a**. This figure also shows the fraction of data sets for which there is no local alignment containing the human translation start site (because it did not score above Ensembl's threshold). As expected, the skewness is the minimum for chimp. For many mouse genes (25%), rat genes (31%) and chicken genes (16%), the skewness is more than 1,000 residues, which suggests nonorthology of the annotated first coding exons. Thus if we were to extract the sequences upstream of the annotated translation start sites for these data sets, they would likely not be orthologous to each other. Also, we observed a huge skew (more than 100 kb) for a substantial fraction of chimp, mouse, rat and chicken data sets, suggesting errors in annotations. For a large fraction of human-fugu orthologs (78%) and human-chicken orthologs (45%), Ensembl does not have any alignments containing the human translation start site. Thus, identifying orthologous promoter regions is even harder among such distant species.

Figure 1a further illustrates potential problems with the chimpanzee annotation, where the skewness in the region of misannotations (100 kb or more) seems to be larger than mouse or rat. This further justifies the need for better annotations and complete sequencing of the chimpanzee genome[25].

3. To extract orthologous promoter regions, sequences upstream of the translation start site are commonly used as a proxy for sequences upstream of the transcription start site. However, for vertebrates these two may be very different, as the genomic distance between transcription and translation start sites can be very large, so that a point 1,000 bp upstream of the translation start site may lie downstream of the transcription start site. We verified this by analyzing these distances in the human genome. Ensembl has annotated transcription and translation start sites for 85% of all human genes. **Figure 1b** plots the distribution of the distance between them. For a large fraction of the genes (33%), this distance is more than 1,000 bp. Thus if one is looking in particular for transcription factor binding sites, it would be better to extract the region upstream of the transcription start site rather than the translation start site itself.

overrepresented across the genes of each species. Our study is primarily focused on determining the best methodology for predicting conserved regulatory elements, whether or not they are overrepresented.

Phylogenetic footprinting

As suggested above, one can think of phylogenetic footprinting as consisting of two steps, the identification of orthologous promoter regions, and the discovery of well-conserved motifs within those regions. We discuss these two steps separately.

Table 1 Number of data sets and motifs for each of the three studies

	4 mammals	4 mammals + chicken	6 vertebrates
Annotated human genes	22,242	22,242	22,242
Genes passing orthologous gene filter	14,215	10,600	8,418
Genes passing orthologous start site filter	5,073	945	21
Genes having a parsimony 0 or 1 motif	4,181	258	9
Motifs	25,317	633	48

Orthologous promoter region identification. The first step in phylogenetic footprinting is to obtain a set of reliably orthologous promoter regions. **Box 1** details the potential problems associated with applying this process to vertebrate genomes, and in this section we present two stringent filters that overcome these problems and result in sets of high confidence orthologous promoter regions.

To identify orthologous genes, we used the homology match defined between pairs of genomes in Ensembl[8] (http://www.ensembl.org). Choosing all human genes that have an annotated ortholog for every species (the 'orthologous gene filter') leads to 14,215 data sets for human/chimp/mouse/rat, 10,600 data sets for human/chimp/mouse/rat/chicken and 8,418 data sets for human/chimp/mouse/rat/chicken/fugu. These and subsequent statistics are collected in **Table 1**.

To ensure orthologous regulatory sequences, we apply a stringent 'orthologous start site filter.' The human gene should have an annotated transcription start site in Ensembl (true for 85% of annotated human genes) and, for every species, there should be an Ensembl local alignment block that overlaps both its upstream sequence and the human upstream sequence, where the upstream sequence is defined to be the 1,000 base pairs upstream of the annotated gene start (which in Ensembl can be anywhere between the translation start site and transcription start site). The first part of this filter implies that we have sequences upstream of the transcription start site and the second

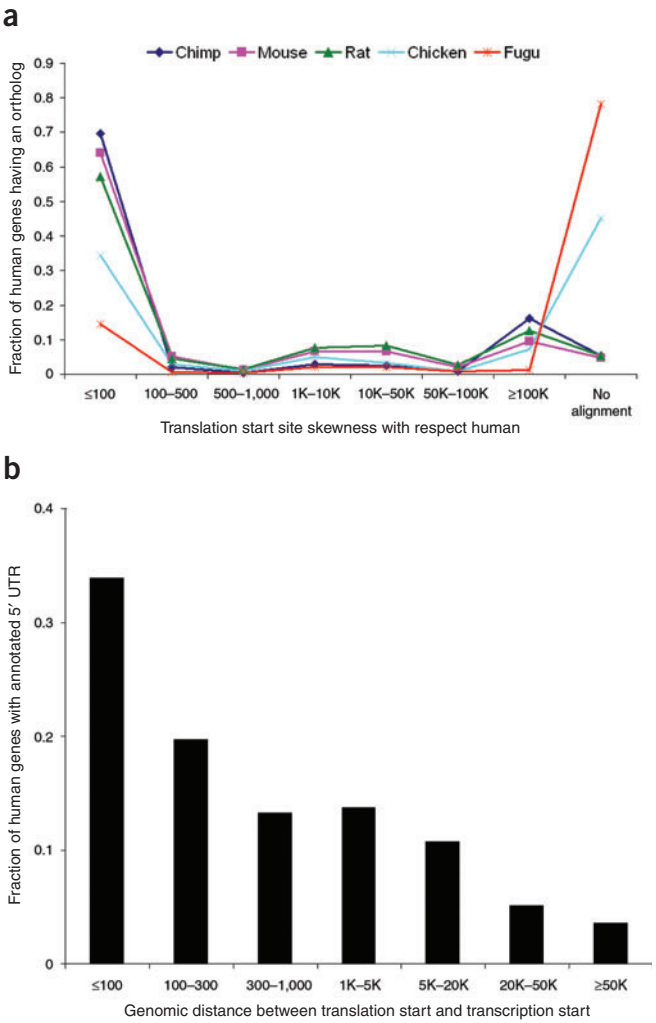

Figure 1 Potential hazards in choosing orthologous promoter regions in vertebrates. (**a**) Skewness in the translation start sites for various species with respect to the human translation start site. The human translation start site is mapped onto the other genomes using the pair-wise alignment blocks provided by Ensembl[8]. Skewness is defined to be the genomic distance to the annotated translation start site of the orthologous gene. The distribution of this skewness is plotted against the fraction of data sets having orthologous genes from both species. (**b**) The distribution of distances between transcription and translation start sites in the human genome.

implies that we have orthologous sequences (see **Box 1** and **Fig. 1**). One could make this filter even more stringent, but that results in a significant reduction of the number of data sets (e.g., only 15% of rat genes have an annotated transcription start site).

Applying this filter, we were left with 5,073 human/chimp/mouse/rat data sets, 945 human/chimp/mouse/rat/chicken data sets and 21 human/chimp/mouse/rat/chicken/fugu data sets. For each of these data sets, we extracted the 1,000-bp region upstream of each annotated gene start. Note the small number of data sets passing our stringent filters, particularly for the distant species. These numbers should improve substantially as tools and annotations improve.

Performance comparison of tools for phylogenetic footprinting. Once reliable data sets of orthologous promoter regions have been obtained, the next step in phylogenetic footprinting is to find well-conserved motifs in each of the data sets. For this we have a variety of multiple alignment tools available. These range from the global multiple alignment tool ClustalW[9] to anchor-based global multiple alignment tools MAVID[10] and MLAGAN[11], to anchor-based local multiple alignment tools DIALIGN[12] and TBA[13], to the phylogenetic footprinting tool FootPrinter[14]. We evaluate the performance of all six of these multiple alignment tools with regard to their ability to predict regulatory elements using real vertebrate promoter sequences on a whole-genome level.

Most of these tools are not designed specifically for the discovery of short motifs, so first we needed an automatic method of extracting well-conserved motifs from the multiple alignments they produce. We used the simple measure of parsimony[15] with respect to the species tree to measure conservation across an aligned column; one of the results that emerged was that we needed to identify motifs of length 10 with parsimony 0 or 1 (that is, nearly identical motif instances in all species), so the particular choice of parsimony as a measure of conservation would be unimportant. A parsimony zero motif means that the motif is identical across all species, whereas a parsimony 1 motif translates into a motif being identical across the species analyzed with the exception of one aligned column that contains one mutation on some branch of the phylogeny. Note that we use such well-conserved motifs because they are strong candidates as transcription factor binding sites. There are surely some bona fide transcription factor binding sites that are not this well conserved, and our method will miss them. However, bona fide binding sites are not characterized on a genome-wide scale so, like all phylogenetic footprinting methods, ours hinges on good conservation as an indication of functionality.

Figure 2a shows the performance of the various alignment tools on the 5,073 human/chimp/mouse/rat data sets. The upper graph plots $h_A(0)$ and $h_A(1)$, that is, the fraction of orthologous data sets for which the best length 10 motif had parsimony score 0 and 1, respectively, for each tool A. This yields the (true plus false) positive rates for each tool and parsimony score. The lower graph shows the likelihood of failure, a measure of the false discovery rate, for the corresponding tools and parsimony scores. (See Methods for the precise definition of likelihood of failure.) TBA and MLAGAN had the best performance, having a high enough $h_A(0)$ and $h_A(1)$ while keeping the corresponding likelihood of failure low. In fact, for the highest quality motifs, we chose TBA alignments containing parsimony 0 and 1 motifs only. These have a small likelihood of failure and still were found in more than 85% of the orthologous data sets.

These tests provide another piece of evidence for the effectiveness of the filter described above: before applying the orthologous start site filter, the various alignment tools found parsimony 0 motifs in only 40–60% of the data sets.

Figure 2b shows the performance of the various alignment tools on the 945 data sets with chicken added. The results and conclusions are quite similar to those on the mammals above: the highest quality predictions are the parsimony 0 and 1 motifs produced by TBA, though in this case they cover only 27% of the data sets. This suggests that chicken is too distant from the mammals to admit many reliable alignments, even when the promoter regions are truly orthologous. MLAGAN also performed well on these data sets.

Adding fugu, we started with only 21 data sets. This was too small a data set on which to do statistical analysis, so instead we analyzed these data case by case. Out of 21 data sets, 8 data sets had a parsimony 0 motif that all tools but ClustalW identified. ClustalW could identify only four of these. Whereas TBA found another two data sets whose best motif had parsimony score 1, MLAGAN found four. The previous analyses without fugu revealed TBA and MLAGAN to have

comparable performance, so we decided to pursue MLAGAN's motifs.

A point to note here is that this comparison of tools was done only on promoter sequences and only with the aim of identifying short motifs: the assessment of tools should not be extrapolated beyond that.

A recent study[16] compared various alignment tools using simulated upstream sequences from evolutionary models of *Drosophila*. Binding sites were modeled as constrained blocks, and pair-wise alignments were used to compare the various tools. In contrast, we use multiple alignments of real upstream sequences to compare these tools. Multiple alignments have an obvious advantage for phylogenetic footprinting, as pair-wise alignments are attempting to predict conserved regions without using all available information. Margulies *et al.*[17] demonstrate that pair-wise alignments fail to identify the conserved motifs found by multiple alignments in this context.

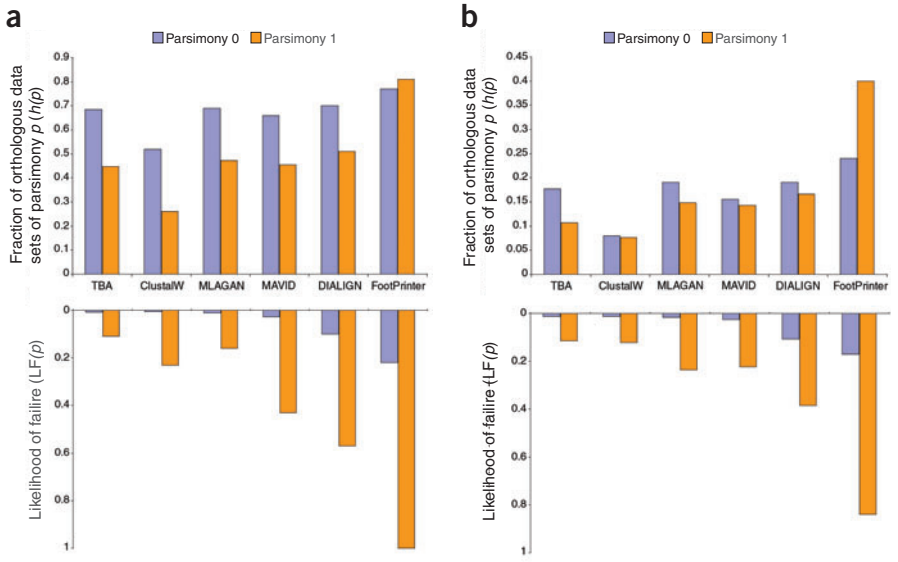

Figure 2 Performance of alignment tools. (**a**,**b**) Comparison of the performance of alignment tools on 5,073 human/chimp/mouse/rat (**a**) and 945 human/chimp/mouse/rat/chicken (**b**) data sets. The upper graph plots $h_A(p)$, the fraction of orthologous data sets for which the best length 10 motif has parsimony score p. The lower graph plots $LF_A(p)$, the likelihood of failure for parsimony score p. The performance is plotted for parsimony scores 0 and 1.

Categorizing high quality motifs

In the previous section we presented a method to identify orthologous data sets containing high quality motifs. The next step is to extract all such motifs, decide which ones are likely to be variants of the same regulatory element and identify which ones are novel. The motifs that occur multiple times are strong candidates for regulatory elements, since they are present upstream of multiple human genes and are perfectly (or nearly perfectly) conserved in the other vertebrates in each of these genes.

Motif clusters. To identify motif overrepresentation, we clustered all the motifs by sequence similarity, so that one cluster might represent potential binding sites of a single transcription factor. For this purpose we developed a greedy clustering algorithm, described in Methods. We also extracted the 1,791 human binding sites from TRANSFAC[18], a curated database of transcription factor binding sites. These were used to establish whether the cluster represented binding sites similar to ones already annotated in TRANSFAC (see Methods).

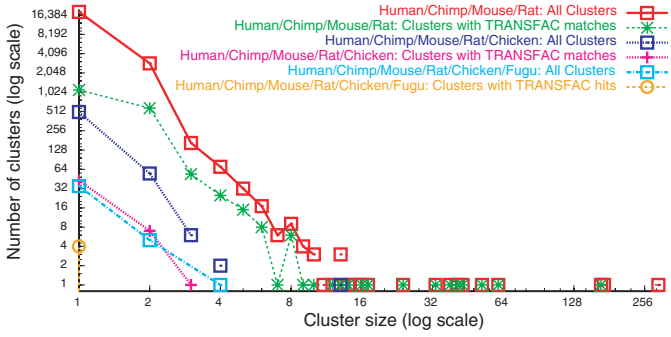

Figure 3 Distribution of cluster sizes for human/chimp/mouse/rat, human/chimp/mouse/rat/chicken, and human/chimp/mouse/rat/chicken/fugu data sets, for all clusters and for those clusters with a significant match in TRANSFAC[18].

It is well known that, if one lists all positions in a single genome having a good match to some TRANSFAC motif, the list will contain many false positives. Note, though, that in the current analysis a human motif instance must be both a good TRANSFAC match and (nearly) perfectly conserved in all the vertebrates under study. This cuts down the false-positive rate considerably.

For the human/chimp/mouse/rat study, TBA parsimony 0 and 1 motifs of length 10 were extracted and overlapping motifs were merged, resulting in 25,317 merged motifs (which we will refer to simply as motifs). The lengths of these motifs range from 10 to 389, with mean length approximately 17. The average number of motifs per gene is approximately 6, evidence that there are often multiple regulatory elements per gene.

These motifs were then clustered using the algorithm described in Methods, resulting in 3,215 motif clusters of at least size 2. **Figure 3** plots the distribution of these cluster sizes in red. Out of these we find 1,807 clusters to have a significant match in TRANSFAC. **Figure 3** also plots the size distribution of these clusters in green. Our web site (http://bio.cs.washington.edu/vertebrate/) shows a similar graph that is the result of clustering using a more permissive set of parameters. The clusters themselves and any significant TRANSFAC matches can also be found on that web site. **Table 2a** illustrates good matches in TRANSFAC for three particular clusters. These clusters contain one gene each that is itself annotated in TRANSFAC as containing a transcription factor binding site, and the motif that we find matches that annotated binding site. There are many other clusters (nearly 200) like this, and they can be found on our web site (http://bio.cs.washington.edu/vertebrate/). This shows that we have discovered the known TRANSFAC binding site, but we also predict novel binding sites for that transcription factor upstream of other genes.

Adding the chicken genome to our analysis, we extracted all maximal parsimony 0 and 1 motifs from the TBA alignments of the data sets exactly as we did for the mammals. This resulted in 633 motifs in human, chimp, mouse, rat and chicken, and these are again good

candidates for regulatory elements. We clustered these in the same way as we clustered the mammalian motifs and then identified clusters with a significant match to some human binding site in TRANSFAC. The size distributions of both are shown in **Figure 3**. **Table 2b** lists two sample clusters, showing the putative human binding sites and the TRANSFAC entries that they match.

Adding fugu, we again extracted the parsimony 0 and 1 motifs, but this time we parsed the MLAGAN alignments. This leads to 48 motifs that are conserved as far as fugu. **Figure 3** also shows the results of clustering these motifs and the ones having a significant TRANSFAC match. All the motifs having a significant match in TRANSFAC are listed in **Table 2c**.

Despite chicken and fugu being very distant from the mammals, we did not consider motifs with higher parsimony scores because analysis showed that such motifs had a very high chance of being false positives. Recent studies[17,19] reported that a very small fraction of regulatory elements are conserved between human and chicken. This further suggests the need for methods and tools that can help identify less conserved regulatory elements with high confidence.

Following are some conclusions that can be derived from **Figure 3**. First, and despite relaxing the threshold that we used to decide whether a cluster had a significant match in TRANSFAC, we were still left with a large list of motifs that had no significant similarity with sites in TRANSFAC. (See our web site http://bio.cs.washington.edu/vertebrate/.) These should be excellent candidates for novel regulatory elements, since they are well conserved in multiple vertebrates and are present upstream of multiple genes. Second, the data indicate that large clusters have a greater chance of matching a binding site in TRANSFAC than the smaller clusters, so that the more overrepresented the motif, the more likely it is to be an annotated binding site. Finally, many clusters of size one (19,537 in the case of human/chimp/mouse/rat) also qualify as candidates for regulatory elements. Having such a large number of unclustered motifs suggests we need a better understanding of the degeneracy of the various binding sites of a transcription factor, because it is unlikely that there are so many distinct transcription factors. Evidence of this can be seen in the results of the permissive clustering (see our web site http://bio.cs.washington.edu/vertebrate/), where often clusters that have binding sites for the same transcription factor

Table 2 Sample motif clusters that have a significant match to some human binding site in TRANSFAC[18]

Human gene description	Putative binding site
(**a**) Sample human/chimp/mouse/rat clusters[a]	
Chorionic somatomammotropin (*Sp1*)	ATGTGTGGGAGGAGCTTCT...
Growth hormone 1	ATGTGTGGGAGGAGCTTCT...
Growth hormone 2	ATGTGTGGGAGGAGCTTCC...
AMPK gamma-2	...CTCTGGGAATCTGTGGGAGGAGCCGAGA
PPP1R1B (also known as *DARPP-32*)	TGTGTGTGGGAGGACACGTG...
***HOX4C* (*HOXD9, HOXD10*)**	...ACACATTAATCTATAATCAAATAC...
Dedicator of cytokinesis protein 4	CATTAATCTATAATTAATTGTG
DNA polymerase beta (43K protein, *ATF-2, CREB*)	GACGCGTGACGTCAC
Chloride channel protein 3	ACGCGTGACGTCAC
F-box only protein 34	...CGGCGCGTGACGTCAC
PDZ domain containing 8	GCGCGTGACGTCAGAG
Adaptor-related protein complex 2	TCGCCGCGTGACGTCATC
NP_060625	GCGTGACGTCACG
(**b**) Sample human/chimp/mouse/rat/chicken clusters	
Interleukin-3 (*ATF/CREB, POU2F1*)	ATGAATAATTACGtct
Potassium channel tetramerization protein 5	...TTGAATGTGAATAATTAC
NP_079321	GTTGGTGAATAATTAA
Phosphoglycerate kinase 1	TGCTGAGCAGCCGCTATTGGCCACAG
Small conductance calcium-activated potassium channel protein 2	CCCAGAGCAGCCGC
ATP2A2	TATTAGAGCAGCCGCCG
Nuclear cap binding protein subunit 2	CAGCAGCCGCCG
(**c**) Sample human/chimp/mouse/rat/chicken/fugu clusters	
Platelet-derived growth factor chain B (*Sp1*)	GAAAGGCTGTCTCCACCCACCTCTCGCACTCTCCCTTCTCC
Homeobox protein Nkx-2.2	CAAATACTGTCTTCATCCACTTGACT
***CYP21B* (*ASP, Sp1*)**	GACCCGCCCACAGAG
Homeobox protein Nkx-2.2	CCAAGACCCGCCCAC
EGR2	GTCGCTGCCCATATATGGACT
Homeobox protein Nkx-2.2	CCAGCCTTATATGGACTG
Prolactin (*POU1F1a*)	CTTCCTGAATATGAATAAGAAATAAAA
Potassium channel tetramerization D5	TTGAATGTGAATAA

Each cluster shows the various Ensembl human genes along with our predicted motif in each. The boldface line contains the TRANSFAC gene and its annotated binding site to which the cluster matches. The binding factor is shown within parentheses (when provided by TRANSFAC) and the highly conserved regions are underlined. [a]These predicted clusters each contain the boldface gene that is annotated in TRANSFAC, and we predicted the same binding site as annotated.

in TRANSFAC are merged together. A heuristic analysis based on parsimony *P* value (Margulies *et al.*[17]) suggests, though, that clusters of size at least 5 in the human/chimp/mouse/rat study, and all clusters in the human/chimp/mouse/rat/chicken study, are statistically significant.

Enrichment for GO categories. Most of the clusters identified here are too small to represent a significant functional enrichment (**Fig. 3**). We studied the larger clusters from the permissive clustering for functional enrichment using the tool GOTM[20] (http://genereg.ornl.gov/gotm/), which reports enrichment with respect to GO categories. Illustrative results for some of the larger clusters are presented in **Table 3**. This table includes estimates of both *P* values and *E* values for each GO category enrichment. As a guiding example, an *E* value of 0.1 would mean that the number of GO categories would have to be ten times as great to see this much enrichment in the best category by chance.

Cluster 1 contains three known genes all involved in two processes related to the formation and rearrangement of protein disulfide bonds. Although this is a small number of genes for a cluster of this size, this is a specialized function shared by very few genes, and the level of enrichment is significant (*E* value, 0.08).

Cluster 14 is a particularly large cluster (762 instances) that shows enrichment in the regulation of transcription and associated RNA metabolism. The consensus motif matches the known Sp1 binding site well. Sp1 is known to act as a master regulator of transcription factors. Cluster 16 shows enrichment for various processes related to the initiation of translation.

Cluster 20 shows enrichment for processes related to receptor binding of hormones. Phosphorylation and dephosphorylation are key activities triggered by these events. There is a good match of this cluster's motif to CREB binding sites in TRANSFAC. CREB is known

Table 3 Functional enrichment for some of the larger human/chimp/mouse/rat clusters studied using GOTM[20].

No.	Cluster motif consensus, size	Genes with GO annotation	GO category enrichment (size, *P* value, *E* value)
1	CTGATTGG, 164	119	Protein disulfide isomerase activity (3, 3.7×10^{-5}, 0.08)
			Intramolecular oxidoreductase activity, transposing S-S bonds (3, 3.7×10^{-5}, 0.08)
14	GGGCGGGG, 762	461	Intracellular transport (36, 1.6×10^{-4}, 0.15)
			Intracellular protein transport (25, 3×10^{-4}, 0.35)
			RNA metabolism (32, 4×10^{-4}, 0.43)
			Nucleobase, nucleoside, nucleotide and nucleic acid metabolism (149, 3.1×10^{-5}, 0.03)
			Cellular metabolism (292, 1×10^{-5}, 0.009)
			Transcription regulator activity (67, 1.9×10^{-4}, 0.2)
			Transcription factor activity (53, 3×10^{-4}, 0.32)
			Transcription factor binding (22, 2×10^{-4}, 0.26)
16	GGAAGTGA, 116	80	Ribosome biogenesis (4, 2.8×10^{-4}, 0.38)
			RNA metabolism (11, 1.5×10^{-4}, 0.2)
			RNA processing (9, 6.2×10^{-4}, 0.41)
			Unfolded protein binding (6, 3.8×10^{-4}, 0.5)
20	GTGACGTCA, 105	82	Protein amino acid dephosphorylation (6, 6.7×10^{-5}, 0.09)
			Receptor binding (12, 1.7×10^{-4}, 0.23)
			Hormone activity (6, 4.7×10^{-5}, 0.06)
			ATP-dependent helicase activity (5, 2.6×10^{-4}, 0.35)
27	TGGGANTTGTAGT, 216	127	Protein transport (14, 1.7×10^{-4}, 0.23)
			Establishment of protein localization (14, 1.8×10^{-4}, 0.25)
			Protein localization (14, 2.3×10^{-4}, 0.33)
			Guanyl nucleotide binding (12, 8.9×10^{-5}, 0.09)
			GTP binding (12, 7×10^{-5}, 0.076)
			Endoplasmic reticulum (16, 1.5×10^{-6}, 0.003)
45	CCGGAAGT, 161	104	Macromolecule biosynthesis (16, 2.2×10^{-5}, 0.02)
			Protein biosynthesis (16, 5.5×10^{-6}, 0.006)
			RNA binding (12, 3×10^{-4}, 0.4)
			Translation factor activity, nucleic acid binding (6, 1.1×10^{-4}, 0.15)
			Structural constituent of ribosome (8, 1.4×10^{-4}, 0.19)
			Translation regulator activity (6, 1.2×10^{-4}, 0.16)
157	CCCCTCC, 235	184	Frizzled-2 signaling pathway (4, 1×10^{-4}, 0.2)
			Morphogenesis (33, 6.4×10^{-5}, 0.14)
			Organogenesis (29, 3.9×10^{-5}, 0.085)
			Development (51, 2.7×10^{-7}, 0.0005)
			Phosphatidylcholine-sterol O-acyltransferase activity (2, 1.9×10^{-4}, 0.41)
			Hormone activity (8, 1.1×10^{-4}, 0.24)

The last column shows the GO category for which the enrichment was observed,along with the number of genes, *P* value, and *E* value of the enrichment in parentheses. The third column shows the number of genes of the cluster having any GO biological process annotation.

to be transcriptionally silent in the unphosphorylated state: it needs to be phophorylated to be active[21]. It is likely that the dephosphorylation genes in this cluster that are transcriptionally activated by CREB serve as negative feedback to inactivate CREB.

Cluster 27, with motif consensus TGGGANTTGTAGT, is particularly likely to represent binding sites of a novel transcription factor, as its closest TRANSFAC match CGGGAAATGTCGA matches this cluster with the insignificant *E* value 17.4. The genes in this cluster are enriched for protein transport and localization. Such transport is GTP dependent, which naturally leads to enrichment in GTP and guanyl nucleotide binding.

Cluster 45 is enriched for processes related to translation and its regulation. There is a good match in TRANSFAC to binding sites of the human transcription factor ELK. Yamazaki *et al.*[22] identified the same regulatory element CCGGAAGT in the upstream region of the mouse gene *Cctq*, known to be regulated by transcription factor Elk-1. *Cctq* is involved in folding of newly synthesized proteins. Cluster 157 is enriched for processes involved in organ development. There is a good match to TRANSFAC binding sites of the transcription factor "Wilms tumor suppressor gene WT1," which is a master switch for organ development[23].

Conclusions

We have identified several potential problems with applying existing phylogenetic footprinting methods to the analysis of vertebrate genomes. Starting with a gene of interest, the first step requires the identification of orthologous genes, which becomes harder as more distant species are included in the analysis. The next step, the identification of orthologous promoter regions is also difficult because of unannotated or missing exons and misannotated start sites. Next, the alignment of orthologous promoter regions is difficult, especially for distant species where, for most genes, we seem unable to obtain better quality alignments for orthologous data sets than for partially orthologous data sets.

In part this work emphasizes the need for better annotation and better computational tools at each of these steps. For instance, many of the genes in **Figure 1a** with skewness over 100 kb are likely to indicate misannotated start sites, and the genes marked 'no alignment' in that figure might benefit from better alignment tools. Note that only 36% of the mammalian data sets and 9% of the data sets having chicken passed the orthologous start site filter, another indication of potential improvements in gene start annotation, especially when more distant vertebrates are included.

Using a very stringent filter (to ensure promoter sequence orthology), stringent alignments (very strongly conserved motifs, present close to the transcription start site in all genomes), a likelihood-based mechanism for evaluating alignment tools, and a statistical clustering algorithm, we produced a list of well-conserved promoter motifs from different sets of vertebrates. Although many of these have significant matches to known transcription factor binding sites, many are novel. Although some of the matches to sites in TRANSFAC may be false positives, this does not really matter: the purpose of this step is to identify those conserved sites that do not have a good match to sites in TRANSFAC. We believe that these are excellent candidates as novel functional regulatory elements.

Even using such stringent methods, we produced a long list of putative motifs in mammals. In part, this long list may be an artifact of our current limited selection of annotated mammalian genomes. As genome sequences from more species, especially mammals, become available, and as the annotations and computational tools improve, our collections of orthologous data sets and well-conserved motifs will grow more reliable. To analyze these data sets better, we will need to relax the conservation criteria and allow for motif losses. An added benefit at that future time is that the clusters of overrepresented and well-conserved motifs may become large enough that we will be able to test them in detail for enrichment for functional classes of genes, thus providing clues to the function of these genes' common regulatory factors. Also, a larger motif list will help identify regulatory elements working in conjunction with each other and thus enhance our understanding of regulatory mechanisms in vertebrates.

METHODS

Parameter settings for comparing alignment tools. For ClustalW and MAVID there are no parameters whose setting affects the motifs produced. DIALIGN has only one such parameter, and we used its default setting. TBA and MLAGAN have more such parameters, which we attempted to optimize as follows. We chose a small collection of orthologous data sets on which both ClustalW and FootPrinter identified well-conserved motifs. We then tuned the parameters of TBA and MLAGAN so that they would align those motifs identified by both ClustalW and FootPrinter. For MLAGAN the default setting performed best, whereas for TBA the best choice was more permissive than the default settings. The final parameters for all tools used in the comparison are listed in **Supplementary Table 1** online.

Likelihood of failure. For each data set of orthologous regulatory sequences and the alignment of those sequences produced by each of the tools, we parse the alignment using a sliding window of length 10. For every such window that contains no gaps and contains no low-complexity region or repeat identified using DUST (R.L. Tatusov & D.J. Lipman, National Center for Biotechnology Information, Bethesda, Maryland, USA, http://blast.wustl.edu/pub/dust/) or RepeatMasker (A.F. A. Smit, R. Hubley & P. Green, http://www.repeatmasker.org), we compute the parsimony score using Fitch's algorithm[15]. In this way we identify the motif with best parsimony score produced by the alignment.

To measure the false-positive rate, we introduce the notion of a 'partially orthologous data set.' In assembling a set of orthologous regions, the most likely place for an error in orthology to occur is at the longest branch of the phylogeny relating the species. Thus, for example, we define a partially orthologous human/chimp/mouse/rat/chicken data set to consist of orthologous upstream sequences from human, chimp, mouse and rat, together with a randomly chosen upstream sequence from chicken. Similarly, a partially orthologous human/chimp/mouse/rat data set consists of orthologous upstream sequences from human and chimp for some gene *g* together with orthologous upstream sequences from mouse and rat for some randomly chosen gene *h*.

For each orthologous data set, we generate a corresponding randomly chosen partially orthologous data set, as defined above. We run each of the alignment tools on this partially orthologous data set and identify the motif with best parsimony score, exactly as described above. The purpose of this is to get an indication of the false-positive rate of each alignment tool. Toward this end, for alignment tool *A*, let $n_A(p)$ be the number of orthologous data sets having a motif with best parsimony score *p*. Then $h_A(p)$, the fraction of orthologous data sets having a motif with best parsimony score *p* is defined in the following way:

$$h_A(p) = n_A(p) / \Sigma_{k \geq p}\, n_A(k)\ .$$

Similarly, $w_A(p)$ is defined as the fraction of partially orthologous data sets having a motif with best parsimony score *p*. Define the likelihood of failure for alignment tool *A* and parsimony score *p* as follows:

$$\mathrm{LF}_A(p) = \mathrm{w}_A(p) / h_A(p)\ .$$

If the likelihood of failure is close to 0, this means *A* is far less likely to produce a parsimony *p* motif in partially orthologous data sets than it is in orthologous data sets, so such well-conserved motifs serve to distinguish these two types of data sets. (Note the analogy of likelihood of failure to the well-known measure of false discovery rate.) This is a very stringent test, much more so than if we used independently chosen genes rather than a partially orthologous data set. Also, taking the presence of even a single parsimony *p* motif in partially orthologous data sets as a token of failure makes this test more stringent.

Clustering motifs. The ideas underlying our motif clustering algorithm are very similar to those used by PSI-BLAST[24]. The score of adding a new motif *m* to a cluster *C* of aligned motifs is the score of the best local ungapped alignment of *m* and the alignment *C*, the score of a motif column being its average-of-pairs score. For this score, a mismatch is assigned score –1.1 and a match scores +1. Each motif *m* is considered in both orientations. The significance of the local alignment score is computed using Karlin-Altschul statistics[4], which compute the probability of seeing such a score in a random database of similar size. We stop adding motifs to the cluster once the significance of the cluster falls below a certain threshold. The significance threshold used is an *E* value of 1/20 multiplied by the cluster size, for stringent clustering. This means that for every 20 cluster elements, we expect one element to cluster by chance. This is again a very stringent clustering technique, and it is likely that multiple clusters may contain the binding sites for a single transcription factor. For permissive clustering, we used the threshold 1/5 multiplied by the cluster size.

To search for a TRANSFAC[18] match, the score of the best ungapped local alignment of the cluster with the TRANSFAC binding sites is checked for significance using the same ideas as above, but with threshold *E* value 0.2 for stringent clustering and 0.5 for permissive clustering.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Mathieu Blanchette, Nan Li, Michal Linial, Larry Ruzzo, Saurabh Sinha, Zasha Weinberg, Zizhen Yao, the Ensembl Help Desk (in particular, Michael Schuster and Ewan Birney) and the anonymous reviewers for their contributions to this work. This material is based upon work supported in part by the National Science Foundation under grant DBI-0218798 and by the National Institutes of Health under grant R01 HG02602.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Tagle, D. *et al.* Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*); nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455 (1988).
2. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
3. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
4. Karlin, S. & Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268 (1990).
5. Dieterich, C., Wang, H., Rateitschak, K., Luz, H. & Vingron, M. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res.* **31**, 55–57 (2003).
6. Elemento, O. & Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* **6**, R18 (2005).
7. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
8. Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004).
9. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500 (2003).
10. Bray, N. & Pachter, L. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14**, 693–699 (2004).
11. Brudno, M. *et al.* LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
12. Morgenstern, B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218 (1999).
13. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
14. Blanchette, M. & Tompa, M. Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31**, 3840–3842 (2003).
15. Fitch, W.M. Toward defining the course of evolution: Minimum change for a specified tree topology. *Syst. Zool.* **20**, 406–416 (1971).
16. Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E. & Eisen, M.B. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 6 (2004).
17. Margulies, E., Blanchette, M., Haussler, D. & Green, E. NISC Comparative Sequencing Program, Haussler, D. & Green, E. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
18. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
19. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Genome Biol.* **432**, 695–716 (2004).
20. Zhang, B., Schmoyer, D., Kirov, S. & Snoddy, J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**, 16 (2004).
21. Thiel, G., Sarraj, J.A. & Stefano, L. cAMP response element binding protein (CREB) activates transcription via two distinct genetic elements of the human glucose-6-phosphatase gene. *BMC Mol. Biol.* **6**, 2 (2005).
22. Yamazaki, Y., Kubota, H., Nozaki, M. & Nagata, K. Transcriptional regulation of the cytosolic chaperonin θ subunit gene, *Cctq*, by Ets domain transcription factors Elk-1, Sap-1a, and Net in the absence of serum response factor. *J. Biol. Chem.* **278**, 30642–30651 (2003).
23. Scholz, H. & Kirschner, K.M. A role for the Wilms' tumor protein WT1 in organ development. *Physiology (Bethesda)* **20**, 54–59 (2005).
24. Schäffer, A.A. *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).
25. Olson, M.V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat. Rev. Genet.* **4**, 20–28 (2003).