Journal of Biology

BioMed Central

Open Access

Research article

Identification of conserved regulatory elements by comparative genome analysis

Boris Lenhard*†, Albin Sandelin*†, Luis Mendoza*‡, Pär Engström*, Niclas Jareborg*§ and Wyeth W Wasserman*¶

Addresses: *Center for Genomics and Bioinformatics, Karolinska Institutet, 171 77 Stockholm, Sweden. ‡Current address: Serono Research and Development, CH-1121 Geneva 20, Switzerland. §Current address: AstraZeneca Research and Development, S-151 85 Södertälje, Sweden. ¶Current address: Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC V5Z 4H4, Canada.

†These authors contributed equally to this work.

Correspondence: Wyeth W Wasserman. E-mail: wyeth@cmmt.ubc.ca

Published: 22 May 2003

Journal of Biology 2003, **2:**13

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/2/2/13

Received: 12 December 2002
Revised: 21 March 2003
Accepted: 8 April 2003

Abstract

Background: For genes that have been successfully delineated within the human genome sequence, most regulatory sequences remain to be elucidated. The annotation and interpretation process requires additional data resources and significant improvements in computational methods for the detection of regulatory regions. One approach of growing popularity is based on the preferential conservation of functional sequences over the course of evolution by selective pressure, termed 'phylogenetic footprinting'. Mutations are more likely to be disruptive if they appear in functional sites, resulting in a measurable difference in evolution rates between functional and non-functional genomic segments.

Results: We have devised a flexible suite of methods for the identification and visualization of conserved transcription-factor-binding sites. The system reports those putative transcription-factor-binding sites that are both situated in conserved regions and located as pairs of sites in equivalent positions in alignments between two orthologous sequences. An underlying collection of metazoan transcription-factor-binding profiles was assembled to facilitate the study. This approach results in a significant improvement in the detection of transcription-factor-binding sites because of an increased signal-to-noise ratio, as demonstrated with two sets of promoter sequences. The method is implemented as a graphical web application, ConSite, which is at the disposal of the scientific community at http://www.phylofoot.org/.

Conclusions: Phylogenetic footprinting dramatically improves the predictive selectivity of bioinformatic approaches to the analysis of promoter sequences. ConSite delivers unparalleled performance using a novel database of high-quality binding models for metazoan transcription factors. With a dynamic interface, this bioinformatics tool provides broad access to promoter analysis with phylogenetic footprinting.

Introduction

The information in genes generally flows from static DNA sequences to active proteins via an RNA intermediary. Depending upon the cellular context of physiological, developmental and environmental inputs, genes are selectively activated via regulatory sequences in the DNA. At their foundation, transcriptional regulatory regions in the human genome are characterized by the presence of target binding sites for transcription factors (TFs). Knowledge of the identity of a mediating TF can give important insights into the function of a gene via inference of the processes or conditions that lead to expression. Research in bioinformatics has developed reliable methods to model the DNA binding specificity of individual TFs. As most eukaryotic TFs tolerate considerable sequence variation in their target sites, simple consensus sequences fail to represent the specificity of binding factors. This realization led to the development of the quantitative representation of binding specificity with position weight matrices [1]. Such matrices can be highly accurate in identifying *in vitro* target sequences [2], but are insufficiently specific in the identification of sites with *in vivo* function to provide meaningful predictions [3]. The *in vivo* binding specificity of a TF depends upon additional properties not modeled by a weight matrix, such as protein-protein interactions, chromatin superstructures and TF concentrations.

Comparison of orthologous gene sequences has emerged as a powerful tool in genome analysis. 'Phylogenetic footprinting' [4] provides complementary data to computational predictions, as sequence conservation over evolution highlights segments in genes likely to mediate biological function. The utility of phylogenetic footprinting extends to a broad array of annotation challenges, but it is particularly suited to the identification of sequences with a functional role in the regulation of gene transcription [5,6]. Despite specific successes [7] in studies of gene regulation, the central algorithms for phylogenetic footprinting remain to be optimized and are thus the focus of continuing research. In particular, new algorithms based on phylogenetic footprinting have been presented for the alignment of genomic sequences, data visualization and the identification of exons [8,9]. Algorithms for the analysis of regulatory sequences have addressed the detection of over-represented patterns in the promoters of co-regulated genes [10], and the improved discrimination of regulatory modules [11], as well as comparative studies of orthologous promoters across collections of microbial genomes [12,13].

Here, we introduce a highly specific algorithm, ConSite, for the detection of transcription-factor-binding sites (TFBSs) that is based on phylogenetic footprinting. Three central components underlie the advance: first, a non-redundant set of transcription-factor binding models; second, a suitable alignment algorithm for orthologous non-coding genomic sequences; and third, modular software for the integration of binding-site predictions with analysis of sequence similarity. We show that our approach results in an increased specificity of predicted TFBSs as a result of a significant reduction of noise. The ConSite algorithm is thus particularly suited to the analysis of pairs of orthologous genomic sequences with limited or no experimental annotation of regulatory elements.

Results

A non-redundant set of high-quality transcription-factor binding models

Potential TFBSs can be identified within a genomic sequence by well-studied computational approaches based on quantitative profiles describing the binding site characteristics for TFs. The quality of matrix models is dependent upon the number of biochemically determined target sites. While the binding specificities of few eukaryotic TFs are described richly in the literature by multiple *in vivo* functional sites, a significant number of TF binding profiles have been produced through the application of *in vitro* target-site detection assays [14]. We collected available data of both types from the biological literature to construct 108 non-redundant high-quality profiles [15]. The profiles are derived from the super-classes vertebrates, insects or plants, but the majority (65%) of matrices model the binding of human or rodent factors. As the majority of the profiles originate from site-selection assays, the average number of TFBSs contributing to each profile is a robust 31.2 sites per model. Information content, in terms of bits of information, is commonly used within bioinformatics to describe the overall specificity of a profile. The models in the collection range in information content from 5.6 to 26.2 bits, with an average of 12.1 bits. All models are hyperlinked to corresponding sequence accession numbers and the PubMed abstract for the article describing the binding study.

Integrating binding-site prediction with analysis of sequence conservation in orthologous genomic sequences

Phylogenetic footprinting provides data complementary to binding-site predictions, for the analysis of gene regulation. The simple hypothesis that motivates phylogenetic footprinting is that important functional sequences will be under selective pressure to be retained over moderate periods of evolution. The classification of sequences as conserved or freely evolving (as proposed by Kimura [16]) is not yet a quantitative process. It should be noted that evolutionary rates vary dramatically between genes and the choice of species is an important consideration in phylogenetic footprinting studies. Too great an evolutionary distance can

result in regulatory alterations or difficulty in aligning short patches of similarity between long sequences. Inadequate evolutionary distance does not significantly improve the overall specificity of predictions. We have developed the ConSite method to integrate phylogenetic footprinting with profile-based predictions of TFBSs, in order to achieve specific predictions of functional regulatory elements in genes. As an example of the influence of species selection on the qualitative performance of the system, the human β globin promoter was compared to a diverse range of orthologs (Figure 1).

In this report, we focus on human-rodent comparisons, as several studies have suggested that only a small portion (17-20%) of non-coding regions are conserved (on average) at this evolutionary distance [10,17]. Furthermore, similarity is punctuated, with distinguishable segments of high similarity flanked by regions of apparently random sequence (roughly 33% nucleotide identity is observed between random genomic sequences, with wide variations dependent upon the applied alignment algorithm, settings, and sequence characteristics [18]). This compartmentalized pattern of similarity is consistent with the emerging emphasis on multiple TFs binding to locally dense site clusters termed regulatory modules [19], which suggests that distinct blocks of sequence are required for transcriptional regulation. In order to identify segments of preferential conservation in orthologous genomic sequences, a suitable set of classification criteria must be defined. As similarity or rates of evolution vary widely across genomic sequences, no single threshold will be perfectly suited. We elected to focus the algorithm on segments of high similarity. This refers to sliding windows of fixed size over the alignment, retaining only those where the sequence identity exceeds a default or user-specified threshold. If a cDNA sequence is available, the analysis program can exclude from consideration binding-site predictions situated within exons present in an alignment of genomic sequences.

Assessing the impact of phylogenetic footprinting on the specificity of binding-site predictions

In order to assess quantitatively the contribution of comparative sequence analysis to the specificity of TFBS predictions, a reference collection of 14 well-studied genes was assembled. We compared the selectivity and sensitivity of the TFBS predictions between those generated with isolated human sequences and those generated with the same human genes filtered by comparative analysis with orthologous mouse gene sequences (Table 1). The sequence pairs ranged in length between 680 and 2,900 base-pairs (bp), but all included the region -500 to +100 relative to the transcription start site. Within the 14 paired sequences are 40 experimentally defined TFBSs (Table 1) for 13 distinct TFs within the set of available matrices. For clarity, these binding sites were not utilized in the construction of the matrix models. A conservation cutoff was set to 70% for all tests, while the window size for conservation analysis was set to 50 bp.

Selectivity

Insufficient experimental data are available to confidently classify predictions as false, because many functional sites remain to be discovered. As the population of true TFBSs within a genomic sequence is anticipated to be small, we define the false-positive rate as the total number of predictions from all models divided by the length of the query sequence. The number of predicted TFBSs was determined for incrementally increasing relative matrix score thresholds (described in the Materials and methods section) between 65% and 90% for both single sequences and the corresponding orthologous pairs:

$$Sel(c) = \frac{\sum_{m \in M} P_{m,c}}{L}$$

where M is the set of 108 models, $P_{m,c}$ the number of predicted sites using model m and relative matrix score threshold c, and L the length of the analyzed sequence in base-pairs (Figure 2a).

Predictive selectivity (measured by the average number of predicted TFBSs per 100 bp of promoter sequence when scanning with all models) improved by 85% (average ratio: 0.15) when phylogenetic footprinting is applied. The ratios of the observed selectivity scores using phylogenetic footprinting to those obtained using single-sequence analysis modes are shown in Figure 2c.

Sensitivity

Sensitivity measures the ability to correctly detect known sites (that is, when a prediction and an annotated TFBS overlap by at least 50% of the width of the thinnest pattern), given a corresponding transcription-factor binding-profile model. Analyses were performed with incrementally increasing relative matrix score thresholds between 65% and 90%. The overall sensitivity (the fraction of known sites detected) was reduced slightly under the conservation requirement: 65.5% were detected with phylogenetic footprinting (settings of 75% relative matrix score threshold, 70% identity cut-off, 50 bp window) as compared to 72.5% when analyzing single sequences (Figure 2b). The fact that a few sites were not detected with the stringent requirements for both regional sequence and specific-site conservation can be attributed to multiple causes. For instance, TFBSs may not be conserved or may be present but not detected by the profile under the thresholds. We conclude that most

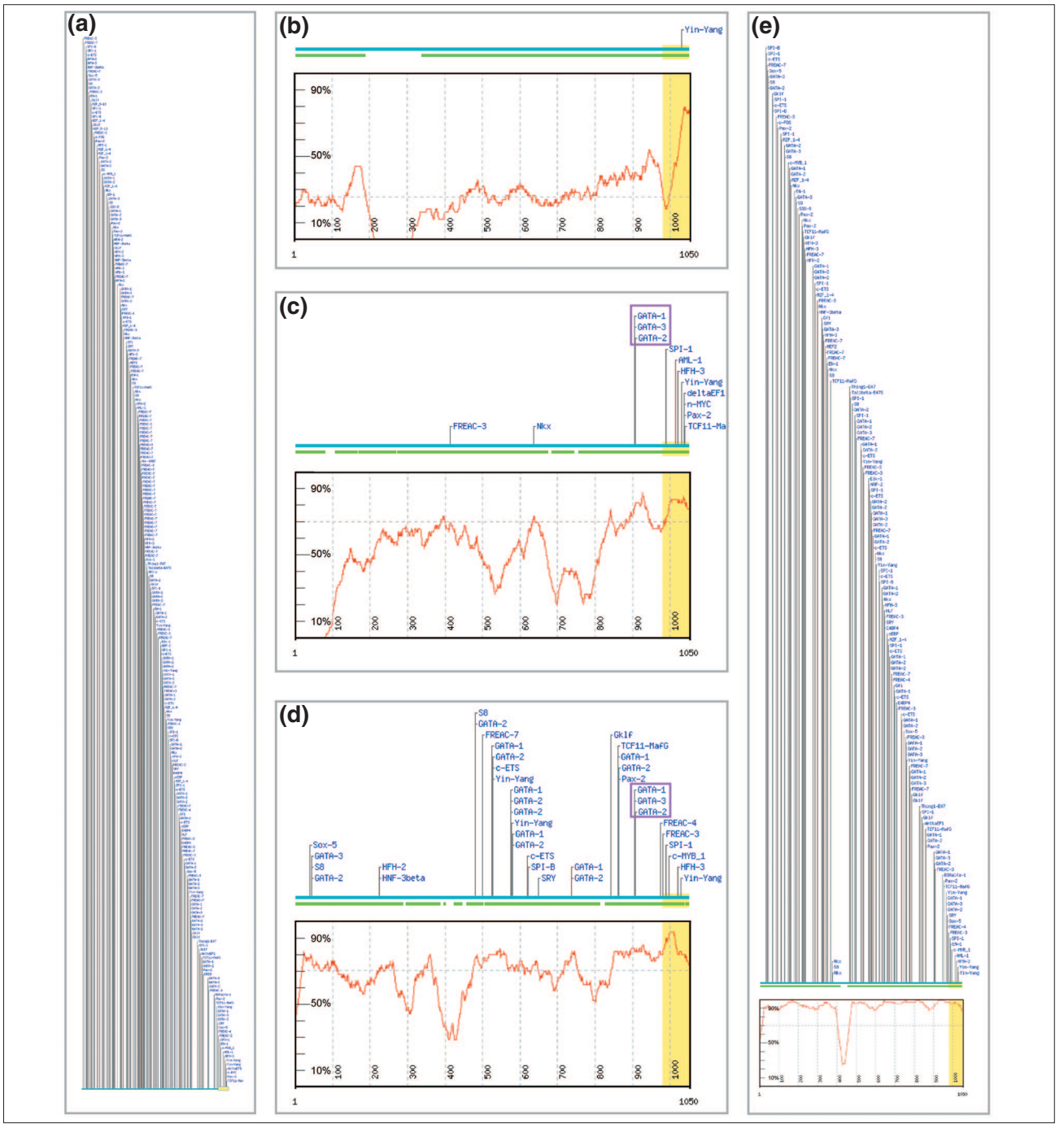

Figure 1
Cross-species comparisons of the β-globin gene promoter. **(a)** Analysis of the human promoter without phylogenetic filtering generates numerous predictions, most of which are biologically irrelevant. **(b)** Comparison with the chicken promoter fails to detect conserved sites (screened with the artificially low conservation cutoff of 25%). **(c)** Comparison with the mouse promoter sequence identifies conserved sites, including a documented GATA-binding site [49] (boxed). **(d)** Comparison with the cow promoter identifies more conserved sites. **(e)** Comparison to the Macaque monkey (*Macaca cynomolgus*) promoter results in a plot similar to the single sequence analysis. Unless indicated, all plots were generated using all available matrices from vertebrates, with 70% conservation cutoff, 50 base-pair window size and 85% transcription factor score threshold settings. The *y* axis in all graphs specifies the percentage of identical nucleotides within a sliding window of fixed length (using the default of 50 base-pairs). The *x* axis refers to the nucleotide position in the human sequence at which the window initiates.

Table 1

The reference collection of 14 gene pairs and 40 verified transcription-factor-binding sites used for testing

Gene name	Human sequence	Rodent sequence	Transcription factors	Binding sequence	Location	MEDLINE ID [49]
Skeletal muscle actin	AF182035*	M12347	SP1	GCGGGGTGGCGCG	-64/-51	11017083
			SRF	ACCCAAATATGGCT	-100/ -86	1922033
			TEF-1	GACATTCCTGCG	-73/-51	11017083
Aldolase A	X12447*	J05517	MEF2	CCTAAATATAGGTC	-125/-111	8413246
α B crystallin	M28638*	U04320	SP1	AGGAGGAGGGGCA	-343/-330	11017083
			SRF	GCCCAAGATAGTTG	-393/-379	11017083
Cardiac α myosin heavy chain	Z20656	U71441 and M62404*	MEF2	TTAAAAATAACTGA	-327/-313	8366095
			TEF-1	AGGAGGAATGTGC	-239/-226	7961957
			SRF	CTCCAAATTTAGGC	-62/-48	8782063
CEBPα	U34070*	M62362	AP2 α	GGCCGGGGGCGGA	-243/-232	9520389
			TBP	TATAAAA	-30/-24	96003748
Cell division cycle protein 2	L06298 and X66172*	U69555	E2F	TCTTTCGCGC	-131/-119	94094909
			cETS	GGGAAG	-109/-104	951721551
Cholesterol 7 α hydroxylase	L13460	U01962*	HNF3β	TCTGTTTGTTCT	-175/-166	9799805
			cEBP	ATGTTATGTCA	-227/-217	28182075
Early growth response protein 1	AJ243425	M22326*	SRF	TGCTTCCCATATATGGCCATGT	-88/-67	90097904
			SRF	CCAGCGCCTTATATGGAGTGGC	-358/-337	90097904
			SRF	GAAACGCCATATAAGGAGCAGG	-412/-391	90097904
Glucose-6-phosphatase	AF051355*	U57552	HNF3 β	CCAAAGA	-72/-66	9369482
			HNF3 β	ACAAACG	-91/-85	9369482
			HNF3 β	GTTTTTGAG	-82/-74	9369482
			HNF3 β	TGTGTGC	-180/-174	9369482
			HNF3 β	TGTTTGC	-139/-133	9369482
			HNF1	AGTTAATCATTGGCC	-226/-212	9369482
Leptin	U43589	U36238*	SP1	GGGCGG	-100/-95	9492033
			cEBP	GTTGCGCAAG	-58/-49	9492033
			TBP	TATAAG	-33/-28	9492033
Lipoprotein lipase	M29549*	M63335	NFY	CAAT	-65/-61	1918010
			cEBP	TAGCCAAT	-68/-61	1918010
			TBP	TATAA	-27/-23	1918010
Muscle creatine kinase	M21487	AF188002 and M21390*	SRF	CCATGTAAGG	-1236/-1227	93233638
			AP2α	GGCCTGGGGA	-1220/-1211	93233638
			MEF2	TCTAAAAATAAC	-1078/-1067	93233638
			MYF	GGGCCAGCTGTCCC	-253/-240	96347575
			MYF	CCAACACCTGCTGC	-1157/-1144	96347575
			P53	ATACAAGGCC	-176/-167	96047120
			P53	ATACAAGGCC	-158/-149	96047120
Rb susceptibility gene	L11910*	M86180	SP1	GGGCGG	-202/-188	1881452
Troponin I	L21905*	U49920 and S66110	MEF2	AGACTATAATAGCC	-976/-962	9774679
			MYF	TAAACAGGTGCAGC	-879/-865	9774679

GenBank accession numbers [41] are given for the human and rodent sequences. The transcription-factor-binding sequences refer to the human or rodent sequence(s) marked with an asterisk. 'Location' refers to the position of the TFBS relative to the transcription start site.

experimentally annotated binding sites are located within conserved regions, as we can correctly detect 82.5% of the TFBSs with a score threshold of 60%, using orthologous gene pairs (data not shown). Ratios of the sensitivity results obtained using single-sequence analysis to those obtained using phylogenetic footprinting, are shown in Figure 2c.

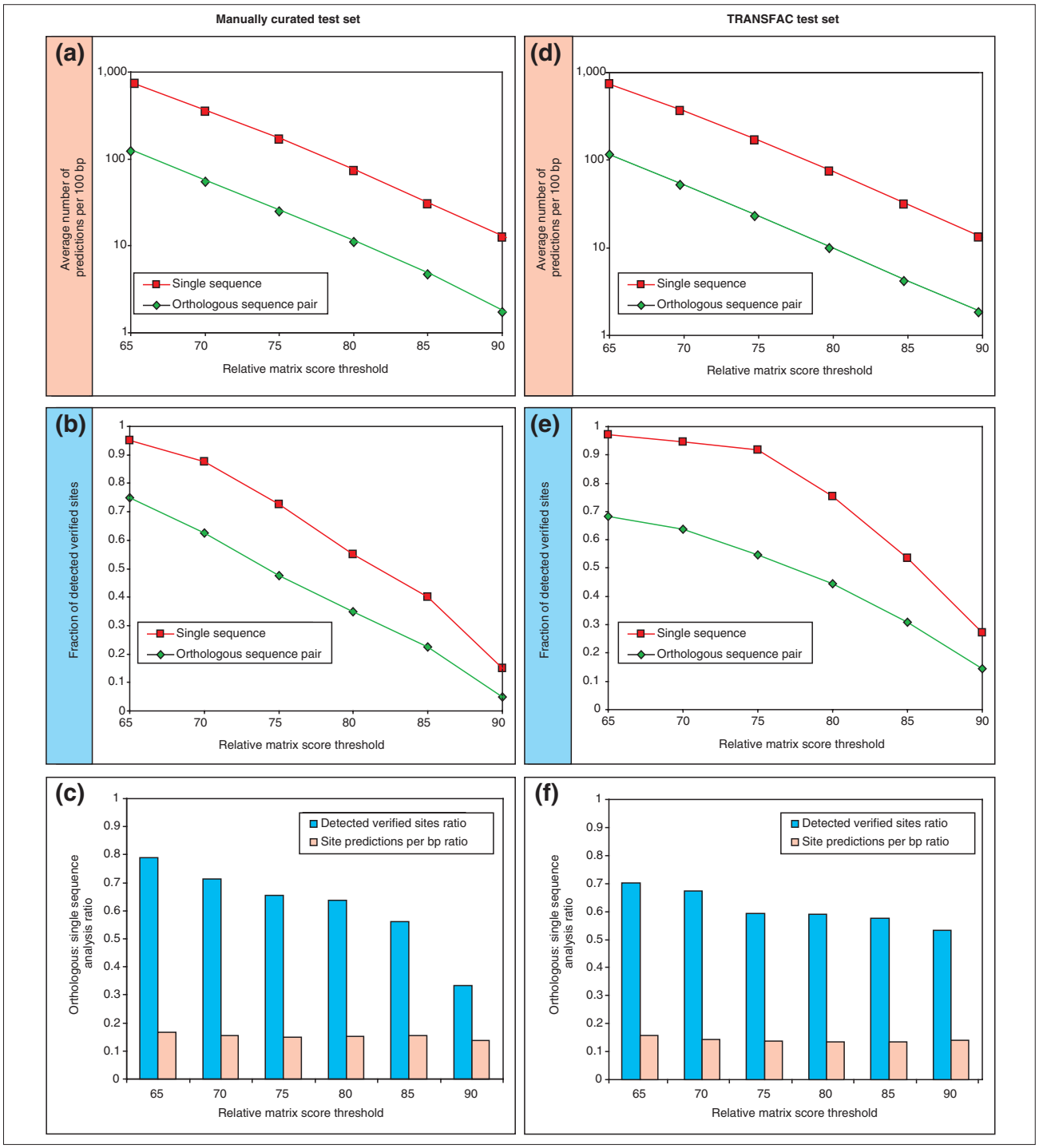

Figure 2
The impact of phylogenetic footprinting analysis. Both **(a-c)** a high-quality set (14 genes and 40 verified sites), and **(d-f)** a larger collection of promoters (57 genes and 110 sites, from the TRANSFAC database [20,21]) were analyzed. (a,d) Comparison of the selectivity (defined as the average number of predictions per 100 bp, using all models) between orthologous and single-sequence analysis modes. (b,e) Comparison of the sensitivity (the portion of 40 or 110 verified sites, respectively, that are detected with the given setting) between orthologous and single-sequence analysis modes. (c,f) Ratios of the number of sites detected in single-sequence mode to the number detected in orthologous-sequence mode; the pair: single-sequence ratios are displayed for both sensitivity (detected verified sites) and selectivity (all predicted sites).

Performance assessment with an extended phylogenetic footprinting TFBS reference collection

Assessment of comparative genome analysis methods requires a broad collection of reference data to insure that algorithms and settings are not overly oriented towards a few genes or factors. A phylogenetic footprinting reference collection was assembled on the basis of the TRANSFAC database [20,21] (as described in the Materials and methods section). For the identification of orthologous genes, only intragenic regions (exons and introns) were used (that is, no potential promoters were included). In any such large-scale mapping, it is of critical importance to find truly orthologous sequences, as opposed to pseudogenes or homologs which have no selective pressure to retain functional binding sites. Our selection process resulted in 110 uniquely mapped TFBSs in 57 promoters of human-mouse orthologous gene pairs (available at [22]). The reference collection does not overlap with the initial set of 14 reference genes described above.

The promoter regions from the reference set were analyzed using the same procedures as were applied above (Figure 2d-f). In spite of the likelihood that the new reference collection will have greater noise than the small set collected by detailed literature analysis, the performance results are comparable between sets. The sensitivity is slightly lower for the large collection (Figure 2e,f), which in addition to the potential difference in annotation standards could be attributable to the TFs associated with the sites. The average information content of the models for TFs linked to sites in the reference collection is lower than that for the factors associated with the small test set (median information content: 9.7, as compared to 15.3 bits in the first test set). Selectivity performance is virtually identical to the test (Figure 2d,f).

Web implementation

The algorithm described for the identification of regulatory regions by comparative sequence analysis has been implemented as an intuitive and easy to use web service named ConSite [23]. The implementation allows for three analysis modes: first, alignment and conserved-site analysis of two orthologous genomic sequences applying one or more TF profiles; second, conserved site analysis on a submitted alignment, which allows users to generate alignments from their preferred tools and allows for the analysis of longer genomic sequences; and third, a single-sequence analysis tool. The single-sequence service is functionally comparable to the TESS system [24], but utilizes the JASPAR profile collection [15]. Alignment submission accepts the *de facto* standard CLUSTALW format [25]. In all operating modes, users are allowed to submit a cDNA sequence to define exon locations. Users may also submit new matrix profiles of their own construction.

Results can be obtained in three distinct report formats. Graphical view (Figure 3a) displays an alignment overview and conservation plots with *x*-axis reference for each submitted sequence. Positions of conserved TFBSs are indicated above the plot. The transcription-factor labels are equipped with mouse-over function to display additional data (the name and structural class of the factor, and the absolute and relative site scores), and are hyperlinked to further information on the TF and its binding profile (Figure 3b). The pop-up windows provide data summaries, including a sequence logo (graphical representation of the specificity of the profile based on position-specific information content [26]) with the corresponding profile from the database. Alignment view (Figure 3c) provides a detailed overview of the detected potential TFBSs displayed on the sequence. The numbering indicates positions in the actual sequences, and the predicted TFBSs are marked. For convenience, a tabular output of detected sites with associated details is also provided in Table view.

Discussion

Comparison of orthologous genomic sequences is an effective method for the identification of segments likely to mediate a sequence-specific biological function. The performance of phylogenetic footprinting methods for the detection of TFBSs is dependent upon multiple factors, including the alignment algorithm, the available binding profiles and the evolutionary distance between the target sequences. Two key data resources are introduced in this study: a novel collection of transcription-factor binding profiles compiled from the biological research literature and a reference test set for phylogenetic footprinting methods. The ConSite web interface to the system facilitates user control, an essential feature for users studying diverse genomes.

The binding profile collection is an important resource for bioinformatics projects. Like the TFBS programming system [27], the JASPAR profile collection is available freely to the research community [15]. The profiles are non-redundant and are restricted to those cases for which sufficient binding data were available to generate a meaningful representation of the binding specificity of a TF. Continuing expansion of the collection is anticipated, given the strong research progress in modeling DNA binding sites [28].

The new phylogenetic footprinting reference collection of TFBSs allows for quantitative assessment of the performance of new methods. This is the largest collection of its kind available for broad use. In our study, we could detect around 68% of the experimentally defined TFBSs in conserved segments (at 65% relative matrix score threshold; see Figure 2). This differs slightly from the outcome of a study of conservation

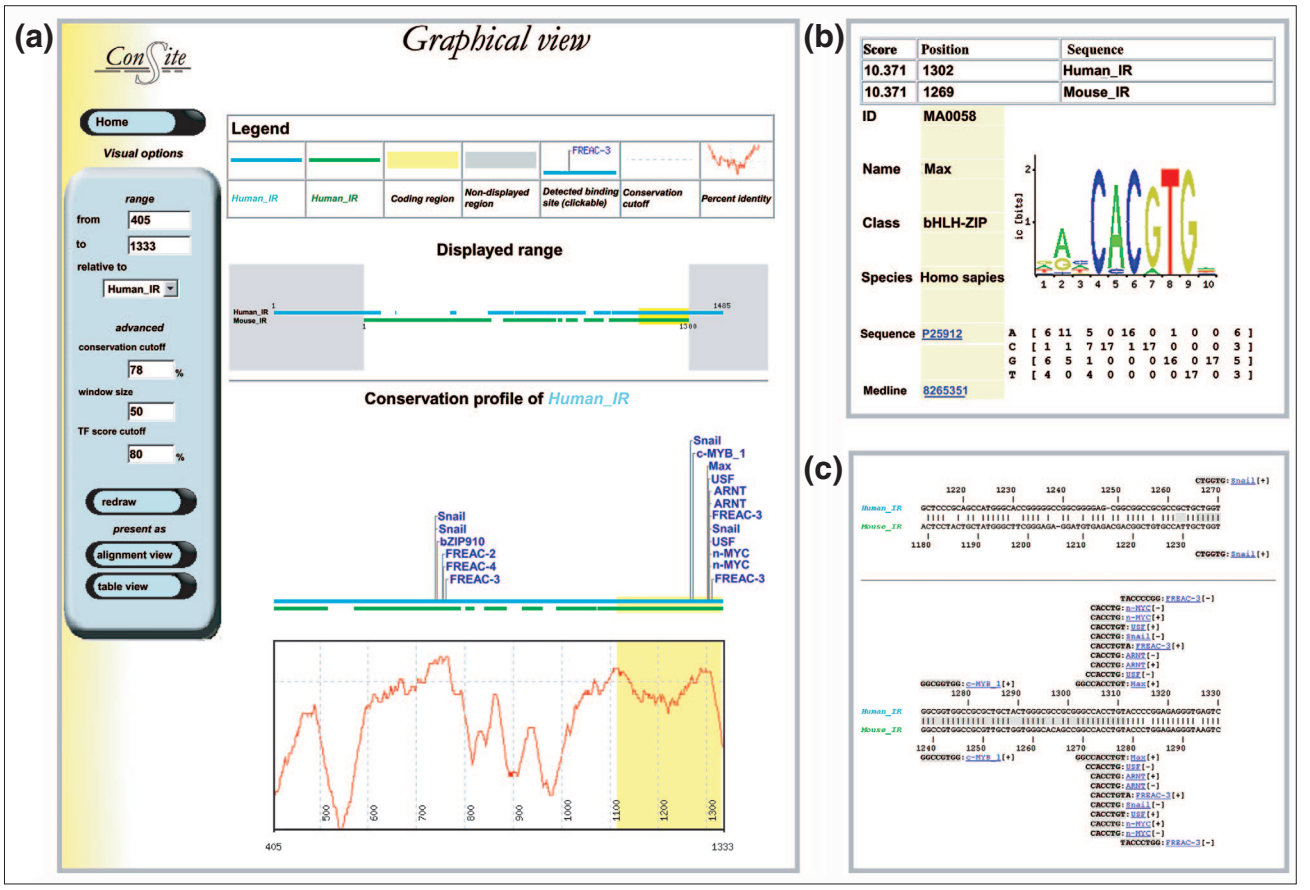

Figure 3
The ConSite result report and visualization tools for the analysis of two orthologous genomic sequences. **(a)** Graphical view, with conservation profile plots for the two orthologous sequences, as well as the control panel for altering the visualization parameters. **(b)** Pop-up window containing information about individual TFBSs. **(c)** Detailed alignment view, providing sequence-level details on putative TFBSs conserved between two orthologous sequences.

properties proximal to TFBSs [29], which indicated that only around 50% of sites are situated in conserved regions. There are several key factors that may account for this difference. The procedures for defining the collections were different. For instance, the amount of flanking sequence used for mapping the locations of the sites onto genome sequences was lower in the previous study. These short fragments were mapped onto a commercial human genome assembly and the mapped regions compared to shotgun-generated fragments of mouse genomes from multiple strains. The alignment procedures were also different, with the older set aligned by BLAST [30] and assessed by a stringent similarity threshold (> 80% identity over 40 bp). There was no exclusion of pseudogenes or paralogous genes indicated in the previous study, which would result in decreased sensitivity due to the erroneous application of phylogenetic footprinting to genes evolving under distinct evolutionary pressures.

While the work presented here focuses on mammalian sequence comparisons, there is no limitation within the ConSite system precluding studies of other organisms (the ConSite website includes samples with insect and nematode sequences). In the future it will be important to develop methods capable of analyzing multiple genomic sequences in parallel, but this is a non-trivial task. Such a system must allow for weighting based on evolutionary distances to preserve sensitivity, and requires advances in multiple sequence alignment algorithms. Some steps in this direction are beginning to emerge [31,32].

No single resource offers the same set of functions or integration as ConSite. The only similarly scoped resource is the recently published rVista [33], which searches for TFBSs in a reference sequence and filters the results for sites in regions of high conservation with respect to a second genomic

sequence. Unlike rVista, ConSite searches both sequences for TFBSs, for better specificity, and enables easy modification of the parameters for interactive analysis, as well as providing different output formats to aid the design and interpretation of experiments in molecular biotechnology. ConSite's publicly available collection of transcription-factor profiles allows users to access information about the TFs associated with the predicted sites. Given that many users focus on a specific TF and have developed high-quality models of their own, ConSite also allows for user-defined profiles.

We present an algorithm that uses phylogenetic footprinting to identify potential TFBSs. The approach to identifying regulatory elements presented here yields greater specificity than previous approaches that were based purely on profile searches of single genomic sequences. In short, using phylogenetic footprinting to filter the computational predictions significantly reduces noise at the price of a slight decrease in sensitivity. The web application we present enables researchers to utilize this approach in a straightforward manner. With the culmination of the human and mouse genome sequencing efforts [34,35], we believe this new algorithm will be of significant use in the ongoing efforts to ascribe function to non-coding sequences.

Materials and methods

Genomic sequence alignment

As a result of the low overall similarity of non-coding regions across moderate evolutionary distances (for example, between human and mouse), many alignment algorithms will fail to produce biologically meaningful alignments or will require an arduous process to tune the algorithm parameters. In order to obtain high-quality global alignments, we utilized the DPB algorithm (L.M. and W.W., unpublished; see [23]), which is optimized for the global alignment of long genomic sequences containing short, colinear segments of similarity.

Measurement of local similarity in global alignments

The most common approach used to measure local similarity between two globally aligned orthologous sequences utilizes a fixed-size sliding window to scan an alignment and identify segments containing a minimum number of identical nucleotides. The difficulties that arise with sliding-window approaches are related to the treatment of edges and gaps in the alignment. Sliding a window along the alignment itself will assign a low identity score to short regions of high identity flanked by long regions of greater variation (for example, a large gap or insertion in one of the sequences). We elected to collapse the gaps in the alignment (that is, to remove the positions containing gaps in the sequence in question) and to calculate a separate conservation profile for each orthologous sequence.

Classification of motif-match conservation within aligned genomic sequences

Within the conserved segments, conserved sites are detected by, firstly, scanning each of the two orthologous sequences with position-specific weight matrices [1] for the TFs of interest, and secondly, retaining only those predicted sites (for each given TF model) that are in equivalent positions in the alignment. The scores for matches to the position-specific weight matrix models must exceed the user-defined relative matrix score threshold.

Collection and annotation of binding models

All profiles are derived from published collections of experimentally defined TFBSs for multicellular eukaryotes. The database, named JASPAR [15], represents a curated collection of target sequences. The motif-detection program ANN-Spec [36] was used to align each binding site set. The ANN-Spec alignments were performed with a range of motif widths, using three random seeds and 80,000 iterations. The profile matrices and associated information are stored in a relational database (MySQL); a flat file representation of the data is available for academic use [22]. Users may also submit their own profiles for private use within the ConSite system.

Identification of relative matrix score thresholds

Candidate TFBSs in individual sequences have a score as determined by the position weight matrix for the given sequence, which has been reviewed elsewhere [1]. The score ranges are unique for each binding model, so it is advantageous to convert the score range to a common, relative unit scale as given by

$$100 \times \frac{score - score_{\min}}{score_{\max} - score_{\min}}$$

Score ranges are used for defining relative matrix score thresholds. The applied scoring method is in direct relation to the protein-DNA binding energy [1], and it therefore does not take into account statistical significance of an observed motif in relation to the local nucleotide composition (for example, GC-rich regions). The influence of the background distribution on the protein-DNA interaction is poorly understood. This is recognized as an open problem within the field, as it is highly controversial whether the surrounding base composition could have any influence on the thermodynamics of binding [37]. For these reasons, we choose to score the matrix profiles using a uniform base composition.

Parameter settings and manipulation

In all three analysis modes the user can choose relative matrix score thresholds (default 80%). In alignment analysis modes, one can also choose the size of the sliding window (default 50 nucleotides) and the conservation cutoff (percentage sequence identity within the window for the definition of conserved regions). There is no fixed default value for the latter parameter; instead, the conservation cutoff is set to retain the top 10% of conserved windows (based on nucleotide identity within a window of sequence in the alignment). This latter mechanism was motivated by the different rates of evolution across genomes.

Matrix manipulation, site detection and phylogenetic footprinting

For matrix manipulation, TFBS detection and some other actions (such as sequence 'logo' drawing) we intensively used the 'TFBS software', a set of object-oriented Perl modules (with extensions in C and C++) developed for the acceleration of promoter analysis scripting [38].

The phylogenetic footprinting TFBS reference collection

An initial set of annotated binding sites was identified from TRANSFAC (version 4.0) [20,21] for human (662 sites) and mouse (376 sites). Each binding site was extended with 50 bp of flanking sequence in both directions from the respective promoter to allow unambiguous mapping onto the corresponding genome assembly (human version hg13 and mouse version mm2 [39,40]). Only sites bound by a TF with a corresponding matrix model in the JASPAR collection were kept.

In order to define orthology without regard to the sequences flanking the binding sites (which would introduce circularity problems), we defined human-mouse pairings on the basis of cDNA sequences. The mappings of GenBank [41] and RefSeq [42,43] cDNAs to the assemblies were obtained from the UCSC Genome Browser Database [39,40]. In addition 50,821 mouse cDNAs from the RIKEN project [44] were mapped to the mouse genome assembly using the client/server version of BLAT [45] with default settings. In brief, for all mappings of a given cDNA, we consider only those with cDNA coverage > 75% and with > 99% sequence identity to the genomic sequence, then sort the set by (number of matches)*(cDNA coverage), and finally take the first mapping in the sorted set.

Each promoter fragment was mapped to its corresponding genome assembly using BLAT, as above. Extended site sequences that unambiguously mapped to the promoter region of the TRANSFAC annotated gene were kept. For each mapped TRANSFAC binding site, the nearest downstream cDNA mapping was located and the GeneLynx record containing that cDNA retrieved. cDNAs with mouse-human ortholog pairs defined in the GeneLynx Mouse [46] database were retained.

For a pair of cDNA sequences thus identified, the genomic sequences spanning representative mappings were extracted and aligned, using BLASTZ [47] (default settings). For each aligned sequence pair, the alignment coverage and the similarities in gene structure as indicated by the mappings were manually evaluated to select not more than one orthologous region per initial TFBS-cDNA-GeneLynx identifier 'triplet'. Promoter-region pairs corresponding to 1,000 bp upstream of the binding site and 100 bp into the first exon were extracted, using the BLASTZ alignment as reference.

Acknowledgements

This project was supported by funds from the Karolinska Institute and the Pharmacia Corporation.

References

1. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
2. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266:**231-245.
3. Fickett JW: **Quantitative discrimination of MEF2 sites.** *Mol Cell Biol* 1996, **16:**437-441.
4. Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, Collins FS: **Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes.** *Mol Cell Biol* 1992, **12:**4919-4929.
5. Pennacchio LA, Rubin EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2:**100-109.
6. Fickett JW, Wasserman WW: **Discovery and modeling of transcriptional regulatory regions.** *Curr Opin Biotechnol* 2000, **11:**19-24.
7. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
8. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES: **Human and mouse gene structure: comparative analysis and application to exon prediction.** *Genome Res* 2000, **10:**950-958.
9. Jareborg N, Durbin R: **Alfresco - a workbench for comparative genomic sequence analysis.** *Genome Res* 2000, **10:**1148-1157.
10. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.
11. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:**1559-1566
12. Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA: **Comparative analysis of regulatory patterns in bacterial genomes.** *Brief Bioinform* 2000, **1:**357-371.
13. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29:**774-782.

14. Pollock R, Treisman R: **A sensitive method for the determination of protein-DNA binding specificities.** *Nucleic Acids Res* 1990, **18:**6197-6204.
15. **JASPAR database** [http://www.phylofoot.org/consite/download]
16. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.** *Nature* 1977, **267:**275-276.
17. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS: **Selective constraint in intergenic regions of human and mouse genomes.** *Trends Genet* 2001, **17:**373-376.
18. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7:**399-406.
19. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
20. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, *et al.*: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
21. **TRANSFAC - The Transcription Factor Database** [http://transfac.gbf.de/TRANSFAC/]
22. **Extended TFBS test set** [http://www.phylofoot.org/consite/testset]
23. **Phylofoot.org tools for phylogenetic footprinting** [http://www.phylofoot.org/]
24. **TESS: Transcription Element Search System** [http://www.cbil.upenn.edu/tess/]
25. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
26. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
27. Lenhard B, Hayes WS, Wasserman WW: **GeneLynx: a gene-centric portal to the human genome.** *Genome Res* 2001, **11:**2151-2157.
28. Bulyk ML, Huang X, Choo Y, Church GM: **Exploring the DNA-binding specificities of zinc fingers with DNA microarrays.** *Proc Natl Acad Sci USA* 2001, **98:**7158-7163.
29. Levy S, Hannenhalli S: **Identification of transcription factor binding sites in the human genome sequence.** *Mamm Genome* 2002, **13:**510-514.
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
31. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9:**211-223.
32. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-1394.
33. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.
34. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
35. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
36. Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2000, **5:**467-478.
37. Schneider TD: **Measuring molecular information.** *J Theor Biol* 1999, **201:**87-92.
38. Lenhard B, Wasserman WW: **TFBS: Computational framework for transcription factor binding site analysis.** *Bioinformatics* 2002, **18:**1135-1136.
39. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, *et al.*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.
40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
41. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/index.html]
42. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
43. **RefSeq** [http://www.ncbi.nlm.nih.gov/RefSeq/]
44. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al.*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
45. Kent WJ: **BLAT - the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
46. Lenhard B, Wahlestedt C, Wasserman W: **GeneLynx Mouse: integrated portal to the mouse genome.** *Genome Res*, in press.
47. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103-107.
48. **MEDLINE** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi]
49. Cao A, Moi P: **Regulation of the globin genes.** *Pediatr Res* 2002, **51:**415-421.