Human-mouse genome comparisons to locate regulatory sites

Wyeth W. Wasserman[1,3], Michael Palumbo[2], William Thompson[2], James W. Fickett[1] & Charles E. Lawrence[2]

Elucidating the human transcriptional regulatory network[1] is a challenge of the post-genomic era. Technical progress so far is impressive, including detailed understanding of regulatory mechanisms for at least a few genes in multicellular organisms[2–4], rapid and precise localization of regulatory regions within extensive regions of DNA by means of cross-species comparison[5–7], and *de novo* determination of transcription-factor binding specificities from large-scale yeast expression data[8]. Here we address two problems involved in extending these results to the human genome: first, it has been unclear how many model organism genomes will be needed to delineate most regulatory regions; and second, the discovery of transcription-factor binding sites (response elements) from expression data has not yet been generalized from single-celled organisms to multicellular organisms. We found that 98% (74/75) of experimentally defined sequence-specific binding sites of skeletal-muscle-specific transcription factors are confined to the 19% of human sequences that are most conserved in the orthologous rodent sequences. Also we found that in using this restriction, the binding specificities of all three major muscle-specific transcription factors (MYF, SRF and MEF2) can be computationally identified.

Regulatory regions of individual genes can be located using transient transfection with deletion mutants, but the large number, diverse distribution and combinatorial interactions of these regions renders unlikely a successful expansion of laboratory studies to a genome-scale. Pattern-based computational approaches[9–11] can suggest possible transcription-factor binding sites and regulatory modules, but false-positive prediction rates are high[12,13]. Because patterns of gene regulation and the corresponding regulatory controls are often conserved across species, cross-species sequence comparison, so called 'phylogenetic footprinting', may identify regulatory sequences[7,14–16]. Due to the high similarity of both biology and sequence between human and mouse, the mouse genome is receiving considerable attention as a tool for cross-species comparisons[17]. This strong similarity, however, has raised doubts regarding the general usefulness of human-mouse sequence comparison for distinguishing functionally conserved features against a background of recently evolved sequence[5,14,15].

Qualitative comparisons using existing algorithms[18,19] indicate that comparison of orthologous human and rodent sequences is useful[7], but quantitative comparisons are lacking. Here we use the recently developed Bayes block aligner (BBA), which focuses on aligning highly conserved, ungapped blocks in which regulatory elements are most likely to reside[20].

We have altered the BBA, originally developed for the analysis of proteins, to permit the analysis of genomic DNA. The alignment algorithm first produces data represented by a two-dimensional histogram reporting the probability that each pair of nucleotides from two subject sequences are located within a conserved block (Fig. 1*a*). Summation of all cells corresponding to each human nucleotide, for example base j (collapsing all bars onto the human axis), produces a histogram (Fig. 1*b*)

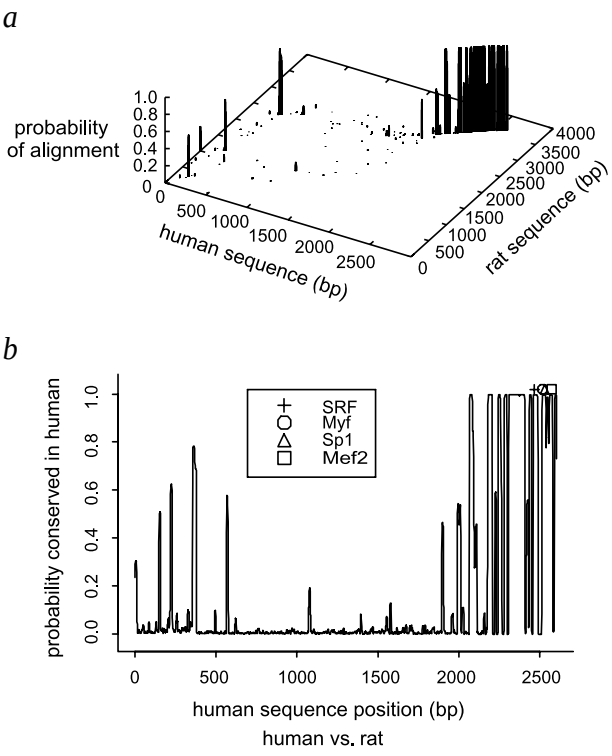

Fig. 1 Probability of alignment for the sequences flanking the 5′ end of the first exon of natriuretic propeptide (NPPA). ***a***, Two-dimensional histogram output of the Bayesian block aligner indicates the probability that any given base j in the human sequence aligns to any given base k in the rat sequence. Probabilities are determined from a set of alignments representative of all possible alignments of the two sequences. ***b***, Probability that a nucleotide j in the human NPPA 5′ flanking sequence is aligned to any nucleotide in the rat sequence in the set of representative alignments reported in (*a*).

Table 1 • Phylogenetic footprinting statistics

Percentage of human genomic sequence aligned to rodent

Per cent aligned	Aligned positions	Length of human sequences
18.6	7,540	40,548

Percentage of identical nucleotides observed in conserved regions

Conserved positions	Identical to rodent	Percentage
7,540	7,029	93.2%

The numbers above are for sequences in which the ratio of rodent length to human length is greater than 0.5. The most conservative estimate of percentage aligned is obtained by comparing the number of conserved positions with the minimum sequence length in each pair. The per cent aligned increases to 22.8.

[1]*Bioinformatics Group, SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania, USA.* [2]*Wadsworth Center, New York State Department of Health, Empire State Plaza, PO Box 509, Albany, New York, USA.* [3]*Present address: Center for Genomics Research, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to C.E.L. (e-mail: lawrence@wadsworth.org).*

© 2000 Nature America Inc. • http://genetics.nature.com

© 2000 Nature America Inc. • http://genetics.nature.com

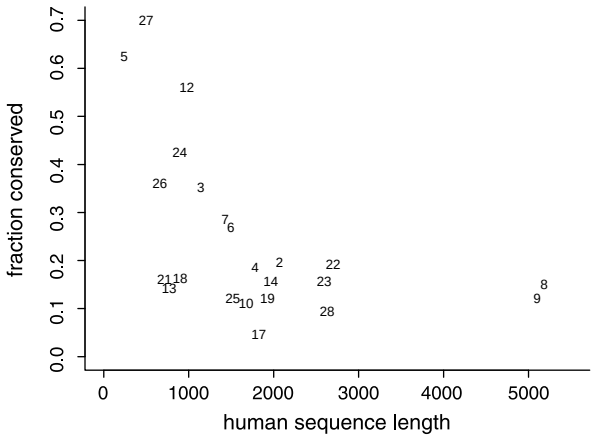

Fig. 2 Conservation of genomic sequence between humans and rodents for alignments where $length_{Rodent}/length_{Human} \geq 0.5$. Fraction of human nucleotides identified as conserved in the comparison of human and rodent sequences. Point labels correspond to the following genes: 1, *DES*; 2, *SLC2A4*; 3, *MYOG*; 4, *MYL4*; 5, *MYL4* (intron); 6, *TNNC1* (intron); 7, *TNNI1*; 8, *MYH7*; 9, *MYH6*; 10, *ACTA1*; 11, *CHRNB1*; 12, *CRYAB*; 13, *COX6A2*; 14, *MYL3*; 15, *MYL2*; 16, *MB*; 17, *PGAM2*; 18, *CHRNG*; 19, *RB1*; 20, *TAGLN*; 21, *CHRND*; 22, *ALDOA* (intron); 23, *NPPA*; 24, *DMD*; 25, *CHRNE*; 26, *ENO3* (intron); 27, *ACTC1*; 28, *CKM*.

that depicts the probability that a human base j is contained in some aligned block, $P_{aligned}(j)$.

To explore the usefulness of human-rodent sequence comparison using this approach, we compared a set of 28 orthologous gene pairs that are specifically upregulated in skeletal muscle, and for which there is considerable genomic sequence available. A total of 99 experimentally defined binding sites exist in the data set, including 24 Sp1 sites (G/C-rich sites) and 75 sequence-specific sites, categorized as follows: (i) myogenin-family (Myf) binding sites (E-boxes); (ii) Mef2 sites (A/T-rich sequences); (iii) SRF sites (CArG boxes); (iv) Tef sites (MCAT boxes); and (v) other experimentally defined, but incompletely characterized, sites.

Our human-rodent comparisons, using the new footprinting algorithm, indicate that only 19% of the human bases have greater than a 50% chance of being placed into an aligned block with a rodent base (Table 1). In other words, 81% of the non-coding genomic sequence is outside the footprinted blocks, a substantial reduction in the sequence 'space'. Even greater reduction may be achieved in the context of the complete genome, as the longer genomic regions, where available, have a somewhat lower percentage of nucleotides in aligned blocks (Fig. 2). Within blocks identified as conserved, 93% of human nucleotides match with identity to the corresponding rodent sequence (Table 1). Excluding binding sites for Sp1, 74 of 75 (98%) binding sites are within the regions shared between humans and rodents (Table 2).

Table 2 • Localization of binding sites to conserved blocks

Factor name	Sites within conserved blocks	Total number of sites
MEF2	20	20
MYF	23	23
SRF	15	15
Tef	7	8
other	9	9
total	74	75

We have pooled nine experimentally defined transcription-factor binding sites for which the mediating transcription factors have not been conclusively identified, therefore this is a heterologous category. Excluded are the binding sites for the Sp1 TF that binds preferentially to G/C-rich sequences and has limited sequence specificity[16]. Of the 24 reported Sp1 sites, 18 (75%) occur within conserved blocks.

As a result, sequence-specific regulatory sites are over 320 times more likely to occur within footprinted blocks. Sp1 sites, which are G/C-rich patches of sequence, have a lower level of conservation (18/24 in footprinted blocks), consistent with the binding characteristics of Sp1 (ref. 21).

Discovery of new transcription-factor binding sites from large-scale expression data can result from the alignment of regulatory regions of co-expressed genes. Such discovery has been feasible in the case of single-celled organisms because most regulatory elements are located within 200–500 bp of the 5′ end of ORFs. In multicellular organisms, however, regulatory elements may be found upstream or downstream of the gene, as well as in introns, and may be spread over tens or even hundreds of kilobases. Experimental evidence suggests that regulatory elements are more probable within a few thousand base pairs 5′ of the transcription start site, but current algorithms are able to locate only some of these regions[12,22]. Phylogenetic footprinting with the BBA provides a way forward by reducing the amount of sequence to be searched.

An alignment algorithm based on a Gibbs sampling approach[23,24] has successfully been used to identify regulatory sites in sets of experimentally determined vertebrate regulatory regions[16], as well as in yeast presumptive promoters[8,25]. We applied this algorithm to the muscle gene set described above. When applied to contextual non-coding sequence of many kilobases around the human genes, the algorithm produces biologi-

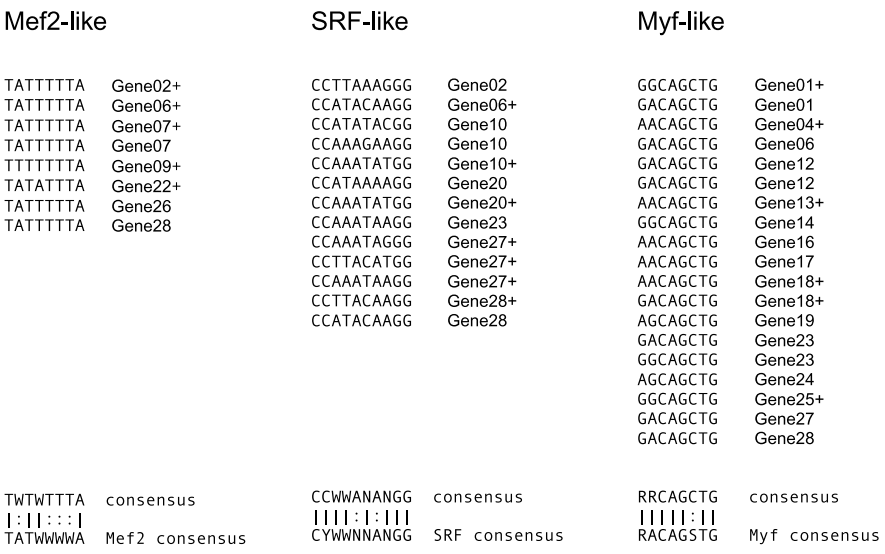

Fig. 3 Three patterns identified in the 5′ flanking sequences of genes selectively expressed in skeletal muscle. Patterns were determined with a Gibbs sampling algorithm that took into consideration the probability of alignment output from the phylogenetic footprinting analysis and the heterogeneous background model. A label (+) indicates the sites with experimental evidence of function. Gene labels correspond to those given Fig. 2. Mef2-Cons, SRF-Cons and Myf-Consensus sequences were taken from ref. 16.

© 2000 Nature America Inc. • http://genetics.nature.com

cally meaningless patterns (data not shown). The regulatory patterns emerge, however, when the algorithm is extended to take conservation into account (that is, when the algorithm is applied to the 19% of sequence found by phylogenetic footprinting; Fig. 3). The consensus patterns of SRF, Mef2 and Myf binding sites, common to a large number of muscle genes, were identified. Excluding a single site, all Tef and incompletely characterized sites were positioned within conserved blocks of sequence, but were not consistently detected by the Gibbs sampling. As anticipated, the GC-rich patches to which Sp1 binds were also not detected.

These results indicate that comparison of human and rodent genomic sequence for a set of co-regulated genes can substantially reduce the size of the sequence space to be searched for functional sites, and make it possible to computationally deduce the binding specificities of critical transcription factors. Our initial data set included only muscle-specific genes, so, although we know of no reason why these results should not apply to other types of genes, broad applicability of this approach awaits further study of additional expression conditions and experimental validation of any new site predictions. Additionally, the tolerance for noise (genes incorrectly classified) in the data set of coordinately regulated genes must be investigated. Previous reports[5,14] indicated that greater evolutionary distances from humans may be required for phylogenetic footprinting. The correct identification of orthologous genes can be difficult for more distantly related species, however, and the biological roles and expression patterns are more likely to be altered. Fortunately, mouse-human comparisons reduce these difficulties and our results indicate that such comparisons may be of considerable value in deciphering the regulatory specifications encoded in the human genome.

Methods

Muscle regulatory regions and sites data set. We assembled a collection of experimentally defined regulatory regions from the literature. Each included sequence has been demonstrated to direct transcription of a promoter in a selective manner in skeletal muscle or a suitable cell-culture model system, and each contained experimentally reported transcription-factor binding sites that were sequence specific. The term 'selective' indicates that transcription does not occur in most tissues or cell models, although expression is observed in one or two additional tissues for many of the genes analysed (most commonly cardiac muscle or brain). An earlier version of this collection was used in an analysis of muscle regulatory regions[16], and an annotated skeletal muscle regulatory region database is available (http://www.cbil.upenn.edu/MTIR/HomePage.html). For every such human regulatory region contained in genomic sequence of over 500 bp and for which orthologous rodent genomic sequence over 500 bp was available, the longest available syntenic sequence pair was included in our study set. Subsequent removal of exon sequence truncated 2 of these sequences to under 500 bp: region 5 is a full-length intron and region 27 is a 485-bp 5′ flanking segment.

Sets of genes expressed in muscle may have little else in common. Our data set includes, for example, genes that determine cell fate in development (for example *MYOG*), energy metabolism genes (for example *ENO3*) and genes specifying the contractile apparatus (for example *MYH7*).

BBA algorithm adapted for phylogenetic footprinting. The BBA has been described for the analysis of protein sequences[20]. The probability of a conserved base is the sum of the probabilities of all alignments containing the base. Individual alignments are constructed using a Bayesian sampling algorithm whose underlying recursion is based on the alignment method[26], which uses a maximum number of gaps, rather than the more commonly used word hashing methods of BLAST (ref. 27) or match/mismatch/gap penalty minimization[19]. This focuses the alignment on conserved blocks and is well suited to the Bayesian formalism. The recursive sampling algorithm yields a representative sample of alignments which can be used to calculate the probability that any given base in the first sequence aligns with a specific base in the second sequence. The program was modified to accept a DNA-similarity matrix. To bias the output towards alignments with short blocks of high similarity, we used the PAM1 similarity matrix[28] (Table 3), that is, a substitution matrix based on the assumption of only one accepted mutation per 100 bp. In this matrix, the probability of a transition is three times that of a transversion.

Table 3 • PAM1 matrix for DNA comparisons

	A	C	G	T
A	0.99	0.002	0.006	0.002
C	0.002	0.99	0.002	0.006
G	0.006	0.002	0.99	0.002
T	0.002	0.006	0.002	0.99

Gibbs sampling procedure for binding-site detection. We applied a pattern detection algorithm to the conserved sequences to identify motifs likely to represent the binding sites of transcription factors contributing to skeletal-muscle-specific gene expression. The Motif Sampler[23,24] iterates between refining a description of the motif and aligning sites in the sequences that may represent instances of the motif. Variation in local base composition adversely affects sequence alignment[27]. Because such variation can be complex in untranscribed sequence[29], and because binding motifs are often AT- or GC-rich, these adverse effects can be difficult to control using existing masking algorithms[30]. The motif alignment algorithm used here uses an alternative approach that uses a heterogeneous background model of sequence composition to account for these variations. The individual input sequence is analysed for heterogeneity in composition using a recently developed Bayesian segmentation algorithm[29]. This algorithm returns the probabilities of observing each of the four bases for each position in the sequence $p0_{i,b}$ $i=1,\ldots, I$ $b=\{A,T,C,G\}$, where I is the length of the sequence. These probabilities are based on compositional heterogeneity of the sequence and the uncertainty in this heterogeneity. The extended Gibbs sampling algorithm used here incorporates this information as a local background model. Specifically, the probability that the sequence R_a, $R_{a+1},\ldots,R_{a+w}$ where R_v is the base at position v in the sequence, is sampled as a binding site is proportional to the ratio of the probabilities of the segment under the site model, pm_v, verses the background, that is,

$$\frac{\prod_{k=a}^{a+w} pm_{k-a,R_k}}{\prod_{k=a}^{a+w} p0_{k,R_k}}$$

where w is the width of the site model.

Phylogenetic footprinting results are incorporated into the sampling algorithm by forbidding a site from being sampled whenever $p_{aligned}$ for the centre location of the site is less than 0.5.

A web server for phylogenetic footprinting analysis and the Gibbs sampling algorithm is available (http://www.wadsworth.org/res&res/bioinfo/). Summaries of the public literature on most of the genes in the data set are available on the muscle regulation site (http://www.cbil.upenn.edu/MTIR/HomePage.html). Accession numbers for the analysed sequences, and the compilation of experimentally defined binding sites, are also available (http://www.cgr.ki.se/cgr/groups/wasserman/muscle).

Acknowledgements

We thank our colleagues at SmithKline Beecham and the Wadsworth Center for input, and the Computational Molecular Biology Core at the Wadsworth Center and I. Auger for assistance. This work was supported by grants from the NIH to J.W.F. (NHGRI R01 HG00981-03) and C.E.L. (NHGRI R01 HG01257).

Received 11 August 1999; accepted 15 July 2000.

© 2000 Nature America Inc. • http://genetics.nature.com

© 2000 Nature America Inc. • http://genetics.nature.com

1. Kadonaga, J.T. Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell* **92**, 307–313 (1998).
2. Orkin, S.H. Regulation of globin gene expression in erythroid cells. *Eur. J. Biochem*. **231**, 271–281 (1995).
3. Qin, W. *et al*. Molecular characterization of the creatine kinases and some historical perspectives. *Mol. Cell. Biochem.* **184**, 153–167 (1998).
4. Yuh, C.H., Bolouri, H. & Davidson, E.H. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
5. Aparicio, S. *et al*. Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nature Genet*. **16**, 79–83 (1997).
6. Brickner, A.G., Koop, B.F., Aronow, B.J. & Wiginton, D.A. Genomic sequence comparison of the human and mouse adenosine deaminase gene regions. *Mamm. Genome* **10**, 95–101 (1999).
7. Hardison, R.C., Oeltjen, J. & Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **8**, 959–966 (1997).
8. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
9. Fickett, J.W. & Wasserman, W.W. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**, 19–24 (2000).
10. Stormo, G.D. & Fields, D.S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **23**, 109–113 (1998).
11. Werner, T. Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome* **10**, 168–175 (1999).
12. Fickett, J.W. & Hatzigeorgiou, A.G. Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878 (1997).
13. 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.* **266**, 231–245 (1997).
14. Duret, L. & Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399–406 (1997).
15. Koop, B.F. Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.* **11**, 367–371 (1995).
16. Wasserman, W.W. & Fickett, J.W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
17. Battey, J., Jordan, E., Cox, D. & Dove, W. An action plan for mouse genomics. *Nature Genet*. **21**, 73–75 (1999).
18. Sonnhammer, E.L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **29**, GC1–10 (1995).
19. Huang, X.Q., Hardison, R.C. & Miller, W. A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.* **6**, 373–381 (1990).
20. Zhu, J., Liu, J.S. & Lawrence, C.E. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25–39 (1998).
21. Lania, L., Majello, B. & De Luca, P. Transcriptional regulation by the Sp family proteins. *Int. J. Biochem. Cell Biol.* **29**, 1313–1323 (1997).
22. Scherf, M., Klingenhoff, A. & Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**, 599–606 (2000).
23. Lawrence, C.E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
24. Liu, J.S., Neuwald, A.F. & Lawrence, C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**, 1156–1170 (1995).
25. Spellman, P.T. *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*. **9**, 3273–3297 (1998).
26. Sankoff, D. & Cedergren, R.J. A test for nucleotide sequence homology. *J. Mol. Biol.* **77**, 169–164 (1973).
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410 (1990).
28. Agarwal, P. & States, D.J. A Bayesian evolutionary distance for parametrically aligned sequences. *J. Comput. Biol.* **3**, 1–17 (1996).
29. Liu, J.S. & Lawrence, C.E. Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38–52 (1999).
30. Wootton, J.C. & Federhen, S. Analysis of compositional biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).

© 2000 Nature America Inc. • http://genetics.nature.com