Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover

Emmanouil T. Dermitzakis and Andrew G. Clark

Department of Biology, Institute of Molecular Evolutionary Genetics, Pennsylvania State University

Comparisons between human and rodent DNA sequences are widely used for the identification of regulatory regions (phylogenetic footprinting), and the importance of such intergenomic comparisons for promoter annotation is expanding. The efficacy of such comparisons for the identification of functional regulatory elements hinges on the evolutionary dynamics of promoter sequences. Although it is widely appreciated that conservation of sequence motifs may provide a suggestion of function, it is not known as to what proportion of the functional binding sites in humans is conserved in distant species. In this report, we present an analysis of the evolutionary dynamics of transcription factor binding sites whose function had been experimentally verified in promoters of 51 human genes and compare their sequence to homologous sequences in other primate species and rodents. Our results show that there is extensive divergence within the nucleotide sequence of transcription factor binding sites. Using direct experimental data from functional studies in both human and rodents for 20 of the regulatory regions, we estimate that 32%–40% of the human functional sites are not functional in rodents. This is evidence that there is widespread turnover of transcription factor binding sites. These results have important implications for the efficacy of phylogenetic footprinting and the interpretation of the pattern of evolution in regulatory sequences.

Introduction

Although regulatory regions are not under the same constraints as coding sequences, alignments of regulatory regions of human and rodent genes often reveal blocks of highly conserved sequences (Hardison, Oeltjen, and Miller 1997; Jareborg, Birney, and Durbin 1999; Leung et al. 2000; Wasserman et al. 2000). Observation of such strong sequence conservation suggests conserved function, thereby generating testable hypotheses that have often been confirmed (Leung et al. 2000; Wasserman et al. 2000). However, studies in *Drosophila* have revealed compensatory changes in gene enhancers (Ludwig et al. 2000), illustrating that conservation of function can be maintained in the face of fluidity in the exact composition of regulatory regions. Compensatory changes are also possible in coding regions, but they do not usually lead to evolution beyond recognition (Mateu and Fersht 1999). Individual binding sites may exhibit relatively little conservation, either because of the degeneracy of the transcription factor binding requirements or because their small size makes it relatively likely that a new functional site will arise by chance (Florea et al. 2000; Ludwig et al. 2000). A new site may relax the selective constraint acting on another already present site, allowing for transcription factor binding site turnover. Nucleotide variation in regulatory regions is considered an important component for disease risk (Risch and Merikangas 1996; Collins, Guyer, and Chakravarti 1997) because variation in binding sites may alter gene expression level and likely contribute to variation in human disease risk (Picketts, Mueller, and Lillicrap 1994; McDermott et al. 1998; Wei and Hemmings 2000; Werth et al. 2000). Understanding the evolutionary processes that binding sites undergo would prove valuable for the inference of potential phenotypic effects and for the interpretation of likely function from human-rodent sequence comparisons. Knowledge of the distribution of divergence within functional binding sites will provide useful information for the calibration of phylogenetic footprinting methods.

In the present study, we analyzed the evolution of human functional binding site sequence in 51 regulatory regions by contrasting the sequences with those of nonhuman primates and rodents. The sequence analysis is rooted by the direct experimental confirmation that the sites under study are functional sequences in the human promoters. For a subset of 20 of the regulatory regions, we obtained comparative functional data from the primary literature for both human and rodents. By comparing regulatory regions from a series of species across a range of divergence times from humans, we capture binding sites at varying degrees of sequence divergence. On the basis of the functional information, this analysis suggests attributes of the manner in which regulatory regions undergo evolutionary turnover.

Materials and Methods

Sequence Data

Human genes were selected for analysis based on the completeness of experimental assessment of identification of functional binding sites in promoter regions (see subsequently). Sequence data were obtained from the NCBI GenBank. We used a combination of keyword and BLAST searches to identify the homologous sequences in non-human primate species and rodents. Some of the rodent sequences were also retrieved from the MGI database (www.informatics.jax.org). A summary of the relevant data is presented in table 1. Species are indicated with the common or genus name. For the analysis, species within the Old World monkey lineage were pooled together, and species from within the New World monkey lineage were separately pooled. Divergence was calcu-

Key words: regulatory evolution, binding site turnover, mammals.

Address for correspondence and reprints: Emmanouil T. Dermitzakis, 1 Rue Michel-Servet, Division of Medical Genetics, Medical School, University of Geneva, 1211 Switzerland. E-mail: Emmanouil.Dermitzakis@medecine.unige.ch.

Mol. Biol. Evol. 19(7):1114–1121. 2002

Table 1
Summary of the Data Used in this Study

Gene	Species	Human Binding Sites (in bp)[a]/ Rodent Binding Sites[b]
acid-labile subunit	Human, mouse	9
acute regulatory protein	Human, macaque	6, 9, 10, 8
adenosine deaminase	Human, mouse	13, 16, 31
adh1	Human, baboon, macaque, mouse	19, 26, 25, 44, 17, 13
aldolase A	Human, mouse	49, 25, 19
Apolipoprotein AI	Human, mouse	27, 29, 16
Apolipoprotein AII	Human, mouse	7, 21, 14, 22, 17, 25, 8
Apolipoprotein CIII.	Human, macaque, mouse	15, 52, 17, 22, 21, 21
Apolipoprotein E	Human, chimp, mouse	21, 12, 32, 21, 7, 6, 13, 19
App	Human, macaque	12, 12
atrial natriuretic factor	Human, mouse	15
c/ebpα	Human, mouse	16, 16, 17, 30, 16/21, 24
ccr5	Human, chimp, gorilla, orangutan, macaque, *aotus*, mouse	9, 10, 9, 18, 9, 16, 20, 10
CD68	Human, mouse	9, 13, 14
Cftr	Human, gibbon, macaque, *saimiri*, mouse	7, 18, 7, 7, 36
c-myb	Human, mouse	22
COL1A1	Human, mouse	28/15
c-reactive protein	Human, mouse	8, 11, 34
CYP1A1	Human, mouse	16
dio2	Human, rat	8, 7, 4
Erythropoietin	Human, mouse	18
factor IX	Human, chimp, macaque, mouse	22, 15, 15
γ-interferon	Human, *callithrix*	13, 61, 13
GnRH	Human, rat	/8, 8, 8, 9
GRP78	Human, rat	33
growth hormone	Human, macaque, *callithrix*, mouse	27, 35, 28
Haptoglobin	Human, macaque, *ateles*, mouse	15, 21, 14, 8, 7, 8/8, 8, 7
Huntingtin	Human, chimp, gorilla, mouse	6, 8
interleukin-3	Human, chimp, macaque, *callithrix*, mouse	7, 9, 29, 22, 28
interleukin-5	Human, mouse	11, 5, 6, 16
L-plastin	Human, mouse	8, 15, 9, 6/6
monoamine oxidase	Human, chimp, gorilla, orangutan	7, 7, 7
msh2	Human, chimp, gorilla, orangutan, *cercopithecus*, *callithrix*	7, 5, 5, 6
mucin 1	Human, gibbon, mouse	7, 28, 14, 12, 6, 17, 8, 32, 6, 13, 14
mucin 2	Human, mouse	8, 10/10
MyoD	Human, mouse	7, 17, 6, 15, 22
Myoglobin	Human, mouse	25
neurofilament M	Human, chimp, gorilla, orangutan, mouse	7, 7, 12
olfactory marker protein	Human, mouse	23, 11, 11
Oxytocin	Human, mouse	6, 6, 6, 6
Plasminogen activator	Human, rat	8, 10
platelet glycoprot. IBα	Human, mouse	16, 16
Preproinsulin	Human, chimp, *cercocebus*, *aotus*, mouse	11, 30, 8, 5, 6, 7, 30, 8
Proglucagon	Human, rat	37/91
p-selectin	Human, mouse	/23, 9, 7
rh50	Human, chimp, gorilla, orangutan, gibbon, macaque, mouse	6, 6, 13, 8, 7, 7
s-cardiac troponin C	Human, mouse	25
Sry	Human, chimp, gorilla, *cercopithecus*, mouse	7, 8
Surf 1–2	Human, mouse	12, 11, 14, 6, 6, 11
Thyroglobulin	Human, rat	21
TNFα	Human, chimp, gorilla, orangutan, gibbon, macaque, baboon, *aotus*, mouse	10, 10, 10, 10, 8, 8, 10, 6, 6, 8/

[a] Size of binding sites mapped in the human sequence and used in the analysis. See also figure 1 for binding factors.
[b] Rodent-specific binding sites.

lated based on the consensus sequence of the lineage. Special attention was paid to the confirmation that the sequences compared were homologous, especially for the human-rodent comparisons. A combination of BLAST searches, with the coding sequence of the genes and gene annotation available in the NCBI GenBank and MGI for human and mouse was used to verify homology. The GenBank accession numbers are provided as supplementary material (see *Supplementary Data* on MBE website: http://www.molbiolevol.org, and web site: http://bio.cse.psu.edu/mousegroup/Reg_annotations/).

Alignments

The primate sequences were aligned with ClustalW and by manual inspection. The divergence among primates was low (<10%), making confidence in the alignments high. For the alignments of human and rodent

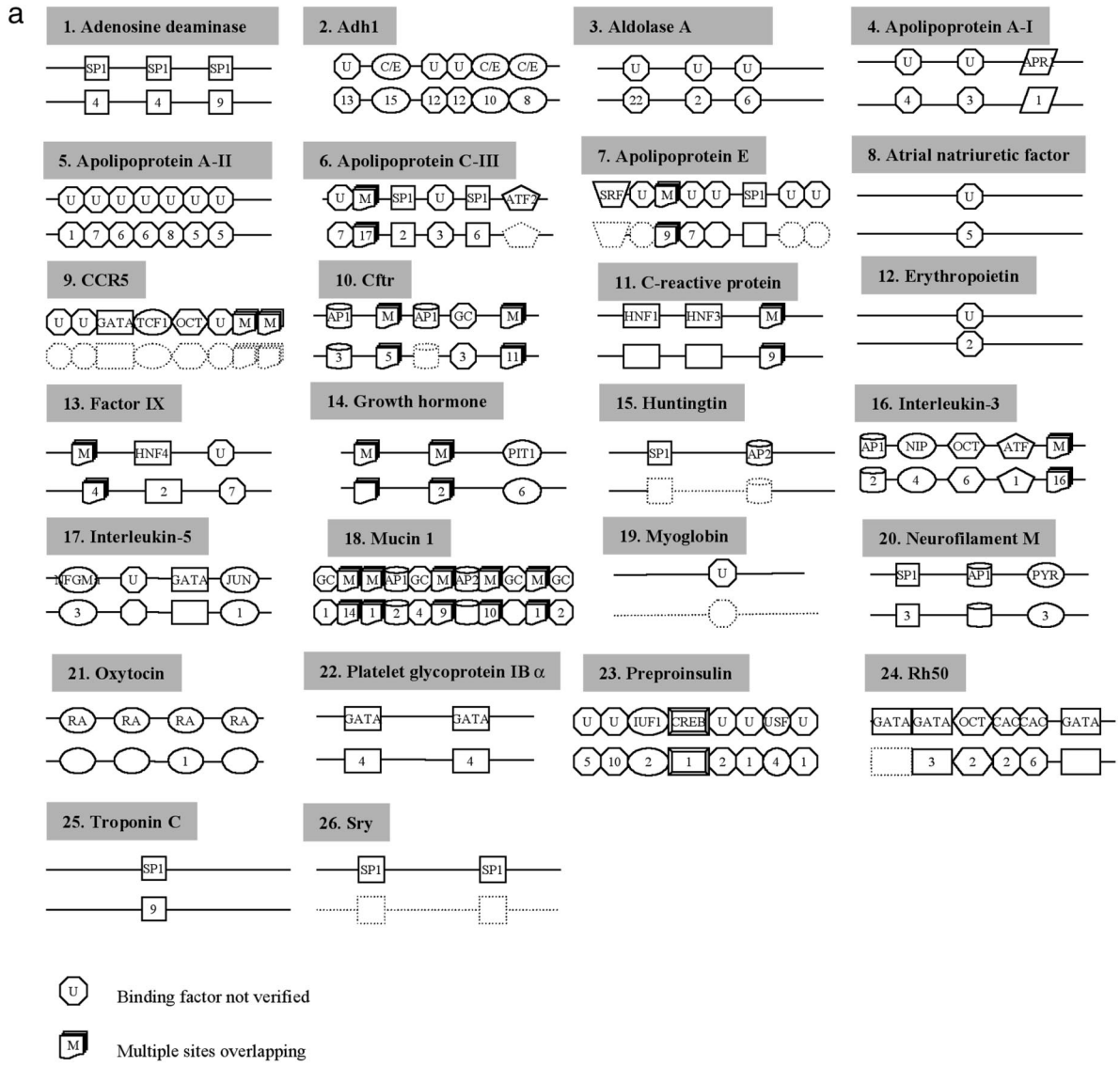

FIG. 1.—Graphical representation of the promoter data used for human (top) and rodents (bottom). Shapes indicate binding sites, and symbols inside the human sequence indicate the binding factor. Some sites do not have a definitive binding factor, but they are verified to be functional (indicated with U), and some others have multiple binding factors with overlapping sequences (indicated with M). In the case where no sequence data were available, the space is left blank for the respective species. Dashed lines and dashed shapes indicate that the regulatory sequence was available, but no significant alignment was found for the respective region. ''Del'' indicates deletion of the binding site because of a larger deletion of the sequence in the respective position. Numbers inside the shapes for the sequences of rodent or human (for the rodent-specific sites) indicate the number of nucleotides that are different in this species' sequence from the human reference functional sequence (including gaps). For the sizes of the binding sites refer to table 1: *a*, Regulatory regions with available functional data only for human.

sequences, we used the web-based software PipMaker to obtain significant local alignments (Schwartz et al. 2000). PipMaker alignments were subsequently manually optimized to obtain the best possible alignment for the binding site sequences. In addition, we used the Bayes Aligner (BA) developed by Zhu, Liu, and Lawrence (1998) to compare with some of the PipMaker alignments within the binding site sequences. Alignments with BA produced essentially the same result. In the rare cases where the alignment was not the same, PipMaker alignments were uniformly better (lower divergence). Therefore, we used the manually optimized PipMaker alignments for our analysis.

Human Functional Transcription Factor Binding Sites

The transcription factor binding sites, used in the analysis, were selected on the basis of direct experimental confirmation of binding ability (footprinting, gel shift assays) and function (promoter deletion experiments, directed mutagenesis, expression of reporter genes) in previous studies. We identified the location of these binding sites in the human sequence by searching the primary literature and the TRANSFAC database (Wingender et al. 2000) (see *Supplementary Data* for references used for the identification of the binding sites). Divergence of binding site sequences for all the

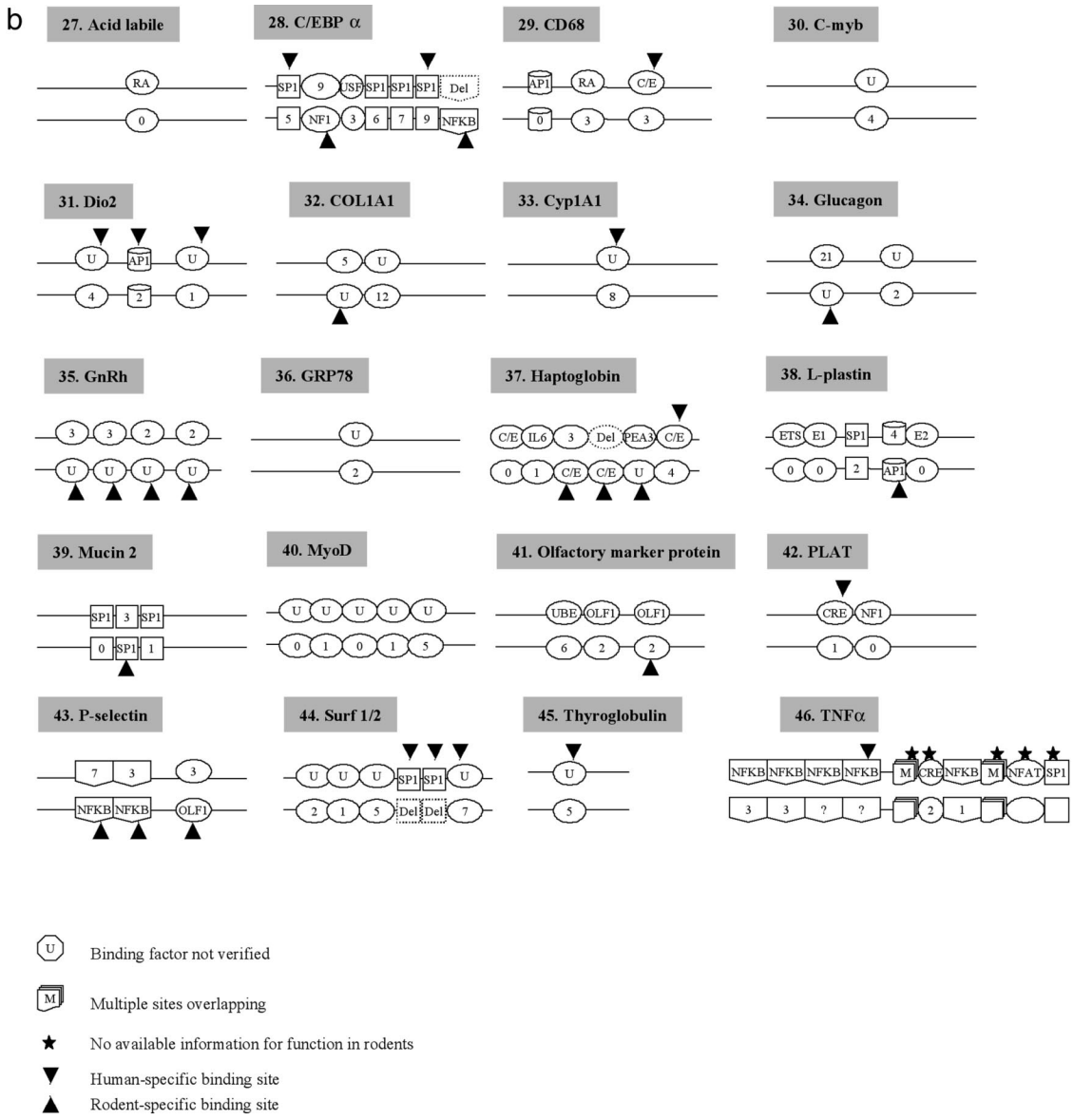

FIG. 1. (*Continued*)—*b*, Regulatory regions with available functional data for both human and rodent. Arrows indicate the species in which the binding site is functional.

human-rodent analysis was done including alignment gaps because we are interested in how different the sequences are in the species compared and not how the substitutions occurred.

Comparative Functional Analysis for Human and Rodents

Data were collected from the primary literature. We restricted the analysis to studies that tested the function and binding ability of binding sites with the same criteria and methods. The criteria for the validity of the function of transcription factor binding sites were as strict as that for the human collection of binding sites. From 20 genes we collected data on 64 binding sites that align between human and rodent, 33 of which share function between human and rodents, 14 that are functional in humans only (human specific), and 17 that are rodent specific (see *Supplementary Data* for references and GenBank accession numbers of the regulatory region sequences).

Results

We analyzed 51 gene regulatory regions in which sequence data are available for human and at least one other primate species or rodent. We used a set of binding sites in these 51 human gene regulatory regions that had strong experimental evidence for a functional role, derived from footprinting, gel-shift assays accompanied by at least one other functional confirmation from either promoter deletion experiments, directed mutagenesis assays, or ability to drive expression in reporter genes. For each regulatory region we used interspecific sequence alignments produced by ClustalW (for primates) or PipMaker (for rodents) followed by manual optimiza-

tion. Binding sites were mapped on the sequences by using reports in the primary literature or by using data available in the database TRANSFAC (see *Supplementary Data*). Summary of the genes analyzed is shown in table 1 and figure 1.

Analysis of Divergence Within Regulatory Regions Between Human and Other Primates

Nucleotide divergence within binding sites between the human sequence and the homologous sequence in other primates suggests that there is a slow process of accumulation of substitutions within binding site sequences. In particular, it appears that the divergence of binding sites between human and macaque is concentrated only in a few sites rather than being distributed homogeneously across sites (fig. 2*a*). We tested this hypothesis by simulating the same average level of divergence in a sample of short sequences equal in length and number to the one aligned between human and macaque. We then computed the variance of divergence between the initial and the derived sequences for each of the 1,000 simulated data sets and compared it with the distribution of variance values obtained from the simulated sets. The observed variance fell in the right tail at $P = 0.015$ (fig. 2*b*), indicating that the substitution pattern within binding site sequences between human and macaque has significantly greater dispersion than the neutral Poisson expectation. The excess dispersion suggests heterogeneity in rates of substitution across binding sites, either because of higher flexibility of the binding properties of some of the transcription factors or because of more relaxed constraints in some binding sites.

Analysis of Regulatory Sequence Divergence Between Human and Rodents

Human-rodent sequence comparisons are widely used to identify regulatory elements in humans (Hardison, Oeltjen, and Miller 1997; Wasserman et al. 2000). However, it is not known as to what proportion of the embedded functional binding sites in human regulatory regions is conserved in rodents. This is a relevant question because nonconserved elements will not produce a strong signal of conservation; therefore, they will not be identified by sequence comparisons. Among comparisons of 46 regulatory regions of human-rodent homologs, 43 produced at least some significant PipMaker local alignments within the region (*sry*, *ccr5*, and *myoglobin* were not successfully aligned).

Average divergence of sequence in the human-rodent comparison within binding sites (*p*-distance: $d = 0.229$, standard deviation = 0.177; Kimura 2-parameter: $d = 0.273$, SD = 0.182) is lower than that of the average synonymous human-mouse divergence (Kimura 2-parameter: $d = 0.468$, SD = 0.169; Makalowski and Boguski 1998) but much higher than that of the nonsynonymous human-mouse divergence (Kimura 2-parameter: $d = 0.090$, SD = 0.102; Makalowski and Boguski 1998), and the divergence of the background sequence (*p*-distance: $d = 0.310$, SD = 0.175; Kimura 2-parameter: $d = 0.399$, SD = 0.178) is very similar to the synonymous divergence. It is possible that other binding sites reside in the aligned regions and are not yet identified as functional. However, the fact that the Kimura 2-parameter estimate of divergence is not very different from the synonymous rate of substitution implies that the density of such potentially unidentified binding sites is low. Additionally, there is no correlation between amino acid sequence divergence of the genes and binding site sequence divergence ($P = 0.680$), and the amino acid divergence in the genes compared is generally low, averaging $d = 0.269$ (SD = 0.139). Therefore, the relatively high binding site divergence we observe cannot be explained by rapid overall gene divergence. In addition, there is no correlation between divergence in individual binding sites in human-rodent and human-macaque comparisons ($r = 0.001$, $P = 0.909$), suggesting that constraints for each site are generally independent in the two different lineages and not a property of the importance of the site for the expression of the gene. Manual inspection of expression profiles from public databases (Unigene, LocusLink, MGI, NCBI) does not suggest any major differences in expression pattern of the genes between human and rodents, but we cannot exclude the possibility that such changes have occurred. Unfortunately, data on tissue- and temporal-specific expression patterns are not unified sufficiently to allow a formal comparison of human versus rodent expression patterns.

Proportion of Species-Specific Transcription Factor Binding Sites

In order to estimate how many binding sites exhibit species-specificity in function we need experimental data for both species. Such data were available for 20 of the 43 alignable regulatory regions compared between human and rodents. A total of 64 alignable binding sites have been identified in these 20 regions, out of which 33 have shared function between human and rodents (mouse or rat), 14 are human specific and 17 are rodent specific. First we tested whether the subset of the data for which there is functional information for both species is representative of the original sample of 43 genes (fig. 3). The nonparametric Mann-Whitney *U*-test (Sokal and Rohlf 1997, pp. 440–447) shows that there is no significant difference between the divergence values obtained from the sample of 20 genes and the divergence values from the remainder of the data (W = 7,746, $P = 0.1948$). In addition, there is no difference between the divergence values of the human-specific versus rodent-specific binding sites (Mann-Whitney: W = 151, $P = 0.9173$), so they can be pooled in one class of species-specific binding sites. There was a highly significant difference, as expected, in the divergence values in binding sites with shared function versus the species-specific binding sites (Mann-Whitney: W = 628, $P = 0.000$). Finally, there was no difference between the divergence values in binding sites compared between human-mouse versus the values in binding sites compared between human-rat (Mann-Whitney: W = 468, $P = 0.930$).

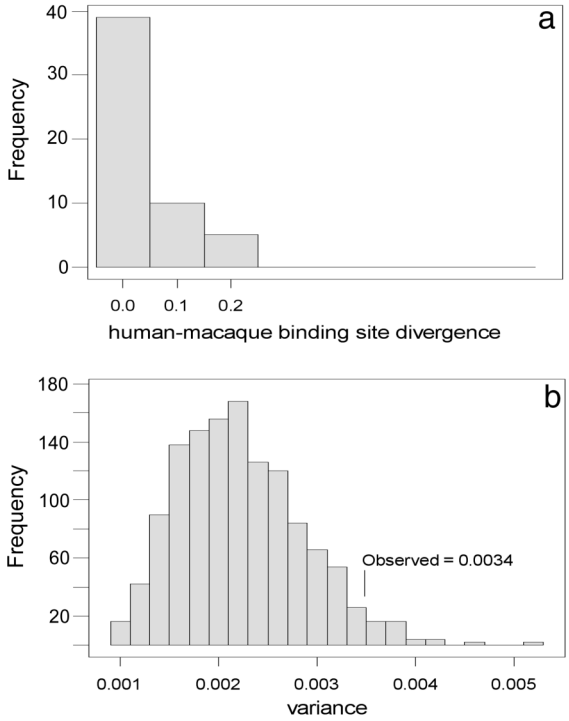

FIG. 2.—*a*, Distribution of divergence within binding sites for human-macaque; *b*, Distribution of variance from 1,000 simulations of a random Poisson process of substitution within binding site sequence for the human-macaque divergence level; the observed value is indicated with a vertical line.

Our data collection method was not biased with respect to functional conservation. Assuming that the comparative studies available in the primary literature are not biased either, we can estimate the proportion of binding sites that do not have shared function between human and rodents. An average of 15.5 sites are species specific (average of 14 human specific and 17 rodent specific) in a total of 33 + 15.5 = 48.5 functional sites present in each species. From this we can calculate that 32% (15.5/48.5) of the functional sites in either human or rodents are not functional in the other species. This is probably an underestimate because observation of the primary literature suggests that most studies consider the conservation in the mechanisms of regulation between human and rodents as null hypothesis; therefore, a strong pattern of functional divergence has to be present so that it is observed and reported.

In order to bypass this bias, we used another method to estimate the proportion of species-specific binding sites, this time taking into account the distribution of divergence of each of the two functional classes of the 64 binding sites (shared function vs. species-specific function). We used these distributions to define the probability of shared function of a binding site between species, given a value of divergence of the functional sequence from the other species sequence. For each functional class we counted the number of occurrences for each interval of divergence equal to 0.1 (e.g., 0.00–0.10, 0.11–0.2, 0.21–0.3 etc) and calculated the proportion of values that fall within this interval for each class. We then estimated the probability that a site does not share

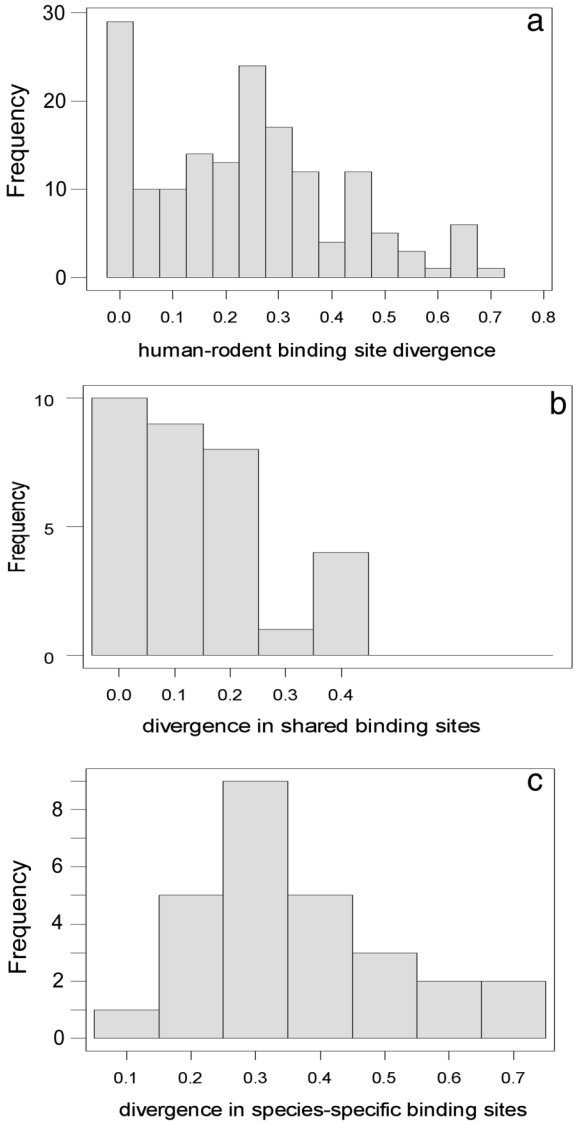

FIG. 3.—Distribution of divergence within binding sites: *a*, for all the data between human-rodents; *c*, for the binding sites with shared function between human-rodents; *d*, for the binding sites with species-specific function in human and rodents.

function in the two species compared, by dividing, for each interval, the proportion of the species-specific values in this interval with the sum of proportions of species-specific and shared values for the same interval. We then used the data from the other subset of the data for which there was functional information only for the human binding sites and computed the predicted number of sites with species-specific function by multiplying the probability defined above with the number of binding sites observed within the same interval of divergence. A total of 38 out of 96 binding sites were estimated to be human specific (40%), similar to the experimental estimate.

Discussion

The results of the present study shed light on long-standing questions about the processes of evolution of

transcription factor binding sites. The pattern of conservation of transcription factor binding sites suggests independent gain and loss in different phylogenetic lineages. The striking variation in the degree of sequence conservation across sites indicates that selective constraints are not always shared among phylogenetic lineages. Comparisons between human and rodents remain informative for the identification of many essential regulatory regions and binding sites (Hardison, Oeltjen, and Miller 1997; Wasserman et al. 2000). However, based on our analysis, a proportion between 32% and 40% of the functional human binding sites are not functional in rodents. It is possible that new binding sites have emerged in the rodent regulatory sequences that replace the function of the lost sites (Florea et al. 2000; Ludwig et al. 2000). This is very likely, given the short length of binding sites and the degeneracy of sequence requirements of the binding factor. In addition, new functions or expression patterns may arise by the independent loss or gain of regulatory elements (Shasikan et al. 1998). These data indicate that the conserved fraction of the genome may be substantially smaller than the functional fraction.

This pattern of evolution has important implications for the use of phylogenetic methods to identify functional regulatory elements for basic and medical research. Distant interspecific comparisons will reveal mainly highly conserved binding sites, and focusing only on those imposes an unfortunate bias in our understanding of regulatory variation. The highly conserved binding sites are those likely to have a radical effect on the expression of the gene, and nucleotide variation in these sites is likely to be associated with rare monogenic disorders. Complex disorders are likely to be mediated by common variants in less constrained binding sites (Risch and Merikangas 1996), precisely those sites that are missed in distant comparisons. On the other hand, comparisons of more closely related species are confounded by the low divergence even in nonfunctional sequences, which will produce many false positives. The positive aspect of our results is that 60%–68% of the transcription factor binding sites are functionally conserved between human and rodents. Therefore, their nucleotide sequence is functionally constrained, and by using the appropriate parameters for calibration, which our data and analysis provides, several methods will be able to identify them within human-rodent alignments of regulatory regions.

The small size of transcription factor binding sites and the degeneracy of binding requirements allows not only for the accumulation of conservative substitutions within binding sites but also for the independent emergence of new binding sites because many different nucleotide combinations will satisfy the binding requirements of a DNA-binding protein (Berg and von Hippel 1987). These new sites may relax the evolutionary constraint in previously essential sites and lead to loss of some of them without serious phenotypic consequences (Ludwig et al. 2000). This pattern of evolution will make it difficult to identify regulatory elements that have undergone turnover. Thus, a tight combination of probabilistic methods for binding site prediction, such as Hidden Markov Models (Durbin et al. 1998, pp. 46–132; Eddy 1998), study of polymorphism in promoter sequences, and extensive functional (Ren et al. 2000) and computational studies (Bussemaker, Li, and Siggia 2001) will be able to detect nonconserved binding sites. Detailed studies of regulatory sequence function combined with more sophisticated comparative genomics (Dubchak et al. 2000; Sumiyama, Kim, and Ruddle 2001), including comparison across multiple species of varying degrees of divergence (such as dog and rabbit) and polymorphism analysis will be informative in capturing the fluid regulatory landscape of mammalian genomes. Finally, these results may lay the foundation for studying how species are different from each other, enabling the identification of genomic segments that are responsible for these differences.

Acknowledgments

We thank Douglas Cavener, Ross Hardison, Brian Lazzaro, Webb Miller, Laura Elnitski, Jim Marden, Kristi Montooth, and Kenneth Weiss for constructive discussions and comments on earlier versions of the manuscript. This work was supported by a Penn State Life Sciences Consortium Innovative Research fund and an NSF dissertation improvement grant to E.T.D.

LITERATURE CITED

BERG, O. G., and P. H. VON HIPPEL. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. **193**:723–750.

BUSSEMAKER, H. J., H. LI, and E. D. SIGGIA. 2001. Regulatory element detection using correlation with expression. Nature Genet. **27**:167–174.

COLLINS, F. S., M. S. GUYER, and A. CHAKRAVARTI. 1997. Variations on a theme: cataloging human DNA sequence variation. Science **278**:1580–1581.

DUBCHAK, I., M. BRUDNO, G. G. LOOTS, L. PACHTER, C. MAYOR, E. M. RUBIN, and K. A. FRAZER. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. **10**:1304–1306.

DURBIN, R., S. EDDY, A. KROGH, and G. MITCHISON. 1998. Biological sequence analysis. Cambridge University Press, Cambridge.

EDDY, S. 1998. Profile hidden markov models. Bioinformatics **14**:755–763.

FLOREA, L., M. LI, C. RIEMER, B. GIARDINE, W. MILLER et al. 2000. Validating computer programs for functional genomics in gene regulatory regions. Curr. Genomics **1**:11–27.

HARDISON, R. C., J. OELTJEN, and W. MILLER. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. **7**:959–966.

JAREBORG, N., E. BIRNEY, and R. DURBIN. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**:815–824.

LEUNG, J. Y., F. E. MCKENZIE, A. M. UGLIALORO, P. O. FLORES-VILLANUEVA, and B. C. SORKIN. 2000. Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. Proc. Natl. Acad. Sci. USA **97**:6614–6618.

LUDWIG, M., C. BERGMAN, N. H. PATEL, and M. KREITMAN. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**:564–567.

MAKALOWSKI, W., and M. BOGUSKI. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95**:9407–9412.

MATEU, M. G., and A. R. FERSHT. 1999. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization. Proc. Natl. Acad. Sci. USA **96**:3595–3599.

MCDERMOTT, D. H., P. A. ZIMMERMAN, F. GUIGNARD, C. A. KLEEBERGER, S. F. LEITMAN, and P. M. MURPHY. 1998. CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). Lancet **352**: 866–870.

PICKETTS D. J., C. R. MUELLER, and D. LILLICRAP. 1994. Transcriptional control of the factor IX gene: analysis of five *cis*-acting elements and the deleterious effects of naturally occurring hemophilia B Leyden mutations. Blood **84**:2992–3000.

REN, B., F. ROBERT, J. J. WYRICK, et al. (11 co-authors). 2000. Genome-wide location and function of DNA binding proteins. Science **290**:2306–2309.

RISCH, N., and K. MERIKANGAS. 1996. The future of genetic studies of complex human diseases. Science **273**:1516–1517.

SCHWARTZ, S., Z. ZHANG, K. A. FRAZER, A. SMIT, C. RIEMER, J. BOUCK, R. GIBBS, R. HARDISON, and W. MILLER. 2000. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res. **10**:577–586.

SHASIKAN, C. S., C. B. KIM, M. A. BORBELY, W. C. H. WANG, and F. H. RUDDLE. 1998. Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a *cis*-acting element. Proc. Natl. Acad. Sci. USA **95**:15446–15451.

SOKAL, R. R., and F. J. ROHLF. 1997. Biometry. 3rd edition, W. H. Freeman and Co.

SUMIYAMA, K., C. B. KIM, and F. H. RUDDLE. 2001. An efficient *cis*-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics **71**:260–266.

WASSERMAN, W., M. PALUMBO, W. THOMPSON, J. W. FICKETT, and C. E. LAWRENCE. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. **26**:225–228.

WEI, J., and G. P. HEMMINGS. 2000. The NOTCH4 locus is associated with susceptibility to schizophrenia. Nat. Genet. **25**:376–377.

WERTH, V. P., W. ZHANG, K. DORTZBACH, and K. SULLIVAN. 2000. Association of a promoter polymorphism of tumor necrosis factor-alpha with subacute cutaneous lupus erythematosus and distinct photoregulation of transcription. J. Investig. Dermatol. **115**:726–730.

WINGENDER, E., X. CHEN, R. HEHL, H. KARAS, I. LIEBICH, V. MATYS, T. MEINHARDT, M. PRUSS, I. REUTER, and F. SCHACHERER. 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28**:316–319.

ZHU, J., J. S. LIU, and C. E. LAWRENCE. 1998. Bayesian adaptive alignment and inference. Bioinformatics **14**:25–39.

THOMAS EICKBUSH, reviewing editor

Accepted February 25, 2002