Mammalian Genomes Ease the Location of Human Transcription Factor Binding Sites but Do Not Ease Their Description

Lee A. Newberg*†‡ Charles E. Lawrence*†

In *Systems Biology: Genomic Approaches to Transcriptional Regulation*
Cold Spring Harbor Laboratory, New York
March 4–7, 2004

Abstract

Comparisons of multiple related genomes have already produced a number of interesting findings, and sequencing resources are available to obtain the genomes of many more species. For studies of human disease, there is naturally a strong interest in the genomes of vertebrates, especially mammals. Decisions concerning the particular species to sequence depend on a number of important factors. While much useful and constructive discussion about these choices has ensued, there have been few quantitative analyses addressing this issue.

Here we consider two of these factors: 1) pattern discovery of functional elements, such as transcription factor binding site models, and 2) identification of unusually conserved sequence fragments. To address these issues, we examined data from seven mammals (dog, cow, pig, rat, cat, baboon, and chimpanzee) which are being sequenced in the NISC Comparative Sequencing Program. We find that, taken together, the data from human, mouse, and the seven additional mammals are only 1.5 times as effective for pattern identification as the data from human and mouse alone. Contrastingly, they are 3.5 times as effective for identification of conserved fragments.

For many reasons, the sequencing of these mammalian genomes is, and will continue to be, a valuable endeavor, but our results suggest that its contribution to the identification of the patterns of functional sites in DNA sequence will be limited. Interestingly, our results are less pessimistic about its contribution to the identification of sequence conservation, and they suggest that the availability of additional sequences will contribute significantly to such an endeavor.

*The Bioinformatics Center, New York State Department of Health, Albany, New York 12208-3425, USA

†Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180-3590, USA.

‡Corresponding author. Corresponding Address: The Bioinformatics Center, New York State Department of Health, 150 New Scotland Ave., Albany, NY 12208-3425, USA.

Problem Statement

While considerable progress has been made in the identification of the genes comprising the human genome, and in the identification of the products of these genes, much less progress has been made in the identification of non-coding functional elements, such as transcription factor binding sites. To address this issue, we focus on two major approaches that have been employed to computationally characterize such functional elements: the characterization of DNA sequence patterns for functional sites, and the identification of unusually conserved DNA fragments.

Identifying Sequence Patterns

Pattern recognition is a crucial part of many DNA analyses, including the identifications of transcription start sites, translation start sites, splice junctions, and *cis* regulatory sites. Approaches to these tasks include those employing hidden Markov models for sequence alignment, [Liu et al., 1999], Markov models for gene finding, [Burge and Karlin, 1997, Reese et al., 1997, Delcher et al., 1999], Gibbs sampling and EM algorithms for motif finding, [Lawrence and Reilly, 1990, Lawrence et al., 1993], and multiple sequence database mining algorithms, [Johansson et al., 2003]. The advantages of using multiple sequences stem from the averaging over the noise present in individual observations, to discern common models of nucleotide or residue frequencies, or to detect conservation. These advantages are reflected in the shrinkage of standard error bars with increasing amounts of sequence data. However, these advantages decrease when the data in the input sequences are correlated, such as with the sequences of phylogenetically closely related species. This loss of advantage, which decreases error-bar shrinkage, and thus the effective sample size of the sequence data, can be substantial. We present an approach to estimate this loss.

Each basepair position in a multiple alignment is characterized by the set of nucleotides observed in a column of the multiple alignment. At a basepair position within a transcription factor binding site, for example, these data need

to be combined appropriately to yield a specification for the pattern recognized by a transcription factor. Nearly all of the methods for these pattern identification problems characterize each basepair of the pattern through the specification, by a vector, of the probabilities over the four standard nucleotide pairs, $\vec{\theta} = (\theta_A, \theta_T, \theta_C, \theta_G)$, where by convention a nucleotide pair is labeled by the name of the nucleotide on the same DNA strand as the gene. Because our goal is to detect and measure the proportion of preference for each possible nucleotide pair at each position within a functional site, we measure the contribution to pattern recognition by a collection of species, as its ability to estimate $\vec{\theta}$ with tight error bars. Specifically, the contribution of a collection of species, $Q_{\vec{\theta}}$(species), is large when the sum of the estimator variances, $\sum_{b\in\{A,T,C,G\}} \mathrm{Var}[\theta_b]$, is small.

For the estimates given here, we use the historically important phylogenetic model for nucleotide substitution described by Neyman [1971] and Felsenstein [1981], with the evolutionary distance between any two species calibrated by the expected number of mutations per sequence position between those species, [Tajima and Nei, 1982, Lanave et al., 1984, Rodríguez et al., 1990]. In Section 1 of the Supplementary Information we give more details on these choices, and we show that the results have only a limited sensitivity to these assumptions; a discussion of alternatives is presented in Section 2.

We estimate the variances using the most widely employed statistical approach, maximum likelihood estimations. Briefly, in this approach, the matrix of estimates of variances and covariances of $\vec{\theta}$ is obtained by taking the inverse of the Fisher information matrix, the matrix of second derivatives of the expected log-likelihood, [Kendall and Stuart, 1998]. Specifics of this process can be found in Section 3 of the Supplementary Information.

When sequences from several species are statistically independent, (*i.e.*, each genome is evolutionarily distant from every other genome), the sum of the estimator variances will be inversely proportional to the number of species. Thus, it is natural to measure $Q_{\vec{\theta}}$(species), for a collection of species related by a phylogenetic tree, by the "effective number of independent species," or simply the "effective sample size," as the ratio of the sum of the variances computed for any single species divided by this sum for the tree.

Identifying Sequence Conservation

Under the assumption that functional constraints on mutations will limit the number of accepted point mutations, several approaches have been developed to identify fragments of genomes that are more highly conserved, [Schwartz et al., 2000]; applications of these methodologies have been successful, [Slightom et al., 1997, Oeltjen et al., 1997, Jang et al., 1999, Rijnkels et al., 2003]. For example, recent analyses of this kind identified vast numbers of previously undiscovered

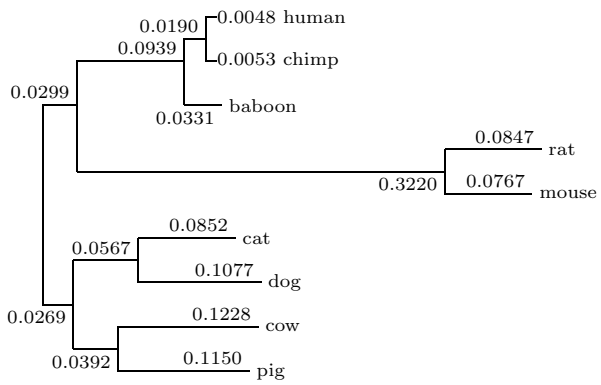

Figure 1: **Phylogenetic Tree of Mammals** from Siepel and Haussler [2003], showing the expected number of mutations per sequence position for each tree edge.

regions of conserved DNA of unknown function, [Dermitzakis et al., 2003, Margulies et al., 2003]. The hope is that the simultaneous use of multiple sequences will refine this approach, to enable these regions to be more precisely defined. Thus, it is natural that we consider the extent to which additional genomes will facilitate this approach.

In our model, the identification of unusually conserved DNA fragments is captured through the specification of a rate-of-mutation parameter, γ, which is normalized to be 1 in the absence of selective pressures, and is reduced to a value between 0 and 1 as a result of selective pressure. Again, we employ a maximum likelihood approach and we assess the contribution of a collection of species through its effect on the height of the error bars of γ. (See Section 1 of the Supplementary Information for more information.)

In this case, we do not calibrate the estimator variance in terms of an effective sample size of independent species; it is nonsensical to speak of scaling the evolutionary distance between two species when that distance is statistically infinite; however we can measure improvement relative to the estimator variance of the human-mouse combination. We will define Q_γ(species), the contribution for a collection of species related by a phylogenetic tree, as the ratio of the estimator variance for the human-mouse combination divided by the estimator variance for that phylogenetic tree.

Results

Using ancestral repeats data from the NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program, [Thomas et al., 2003], a phylogenetic tree for nine mammals was determined, [Siepel and Haussler, 2003], which we depict in Figure 1. The estimated overall distribution of nucleotides for this tree is: $\theta_A = 0.2967$, $\theta_T = 0.3122$, $\theta_C = 0.1949$, and

	Species	$Q_{\vec{\theta}}$(species)
1	*Homo sapiens* / human	1.000
2	*Mus musculus* / mouse	1.403
3	*Canis familiaris* / dog	1.614
4	*Bos taurus* / cow	1.758
5	*Sus scrofa* / pig	1.863
6	*Rattus norvegicus* / rat	1.952
7	*Felis catus* / cat	2.036
8	*Papio cynocephalus anubis* / baboon	2.068
9	*Pan troglodytes* / chimpanzee	2.074

Table 1: **Identification Patterns in DNA Sequence**: the optimal greedy order in which to sequence seven remaining mammalian genomes (after human and mouse), along with an estimate of the resulting effective number of independent species, $Q_{\vec{\theta}}$(species). These values are plotted in Figure 2.

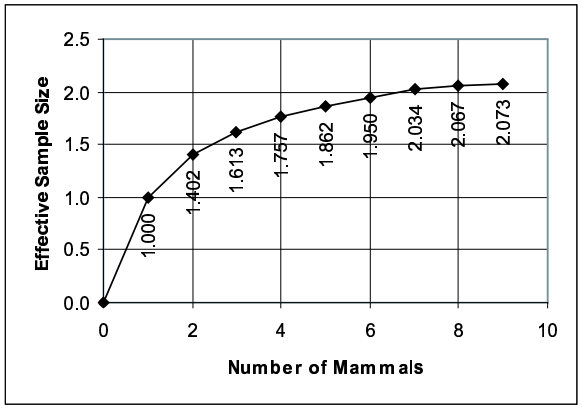

Figure 2: **Identification of Patterns in DNA Sequence**: the effective number of independent species, $Q_{\vec{\theta}}$(species), starting with human and mouse, and adding seven remaining mammals in the order: dog, cow, pig, rat, cat, baboon, and chimpanzee.

$\theta_G = 0.1962$.

Under the assumption that human and mouse are already sequenced, we determine that the subsequent species that will most increase the effective number of independent species, $Q_{\vec{\theta}}$(species), is dog. That is, the subtree of Figure 1 containing just human, mouse, and dog has the highest effective number of independent species, among all subtrees that are composed of human, mouse, and exactly one other species.

Once dog has been sequenced, the most useful subsequent species is cow. Proceeding in this greedy fashion we determine the order in which to sequence the remaining mammals (depicted in Table 1, and with the efficiency plotted in Figure 2).

The sequencing of dog will gain us only about 5/9 as much as did the sequencing of mouse, with returns diminishing yet further for subsequent species. Even with sequence data for all nine of these genomes, we will barely double our ability to characterize the sequence patterns of nucleotides, and the $Q_{\vec{\theta}}$(species) values are worse with a less uniform $\vec{\theta}$ or with $\gamma < 1$.

Note that for the tree of Figure 1, rat would have been a slightly better choice than mouse for the second species to sequence (after human). The effect of such a scenario on Table 1 would be to exchange the order of rat and mouse, and to very slightly perturb the second through fifth $Q_{\vec{\theta}}$(species) values.

The results, with respect to the rate-of-mutation parameter γ, are plotted along side the results for $\vec{\theta}$ estimation, in Figure 3. The recommended order for sequencing the species is the same as with pattern identification. For conservation identification, the sequencing of dog would gain us a variance reduction factor of about 1.7. With sequence data for all nine of these genomes, we would nearly quadruple our ability to identify DNA conservation, over the situation with human and mouse alone.

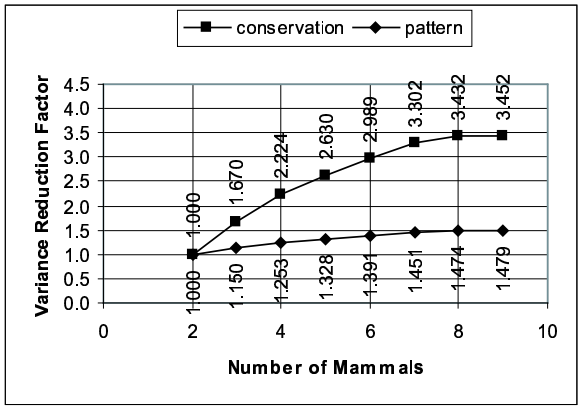

Figure 3: **Comparison of Variance Reduction Factors** in the identification of patterns in DNA sequence *vs.* the identification of DNA conservation. The mammals are added in the optimal greedy order: human, mouse, dog, cow, pig, rat, cat, baboon, and chimpanzee. Both variances are normalized, so that the variance of the human and mouse pair is equal to 1.

Discussion

In searching for the patterns describing transcription factor binding sites in prokaryotes, McCue and colleagues were able to identify transcription factor binding sites and sets of co-regulated genes, or regulons, using only the genome sequences of multiple related species, [McCue et al., 2001, 2002]. They achieved these genomic-scale results because they could identify patterns likely to be transcription factor binding sites, using a single gene and its orthologs in related species. The results of Table 1 do not encourage a similar approach among the mammals. It seems that significant progress in transcription factor binding site pattern identification for mammals will require additional data that identify multiple genes, for which multiple observations of each will be necessary if we are to discern a pattern of interest. For example, data from a carefully controlled expression array study may be appropriate.

The addition of non-mammalian species would likely improve the effective sample size substantially, but with the tradeoff of a loss of specificity. Thus, DNA patterns associated with broadly conserved functions (such as the biosynthesis of fundamental cellular components) may be identifiable with such data sets, but the addition of non-mammal species will not aid in the identification of those patterns that are specific to specialized mammalian functions.

Contrastingly, the results for the identification of conserved regions of DNA are encouraging. With sequences from additional mammals, we will likely attain significant improvement in our ability to locate regions of conserved DNA. In particular, researchers have used conservation as a step in locating transcription factor binding sites in γ-proteobacteria, [Rajewsky et al., 2002], metazoans, [Lenhard et al., 2003], and monkeys, [Boffelli et al., 2003], and we can expect this technique to become more useful in human studies with the increasing in the availability of mammalian genomes. Unfortunately, as indicated above, it appears that subsequent computational identification of the sequence patterns of functional sites will not be as easy.

Acknowledgements

The authors gratefully acknowledge the use of the Computational Molecular Biology Core of the Wadsworth Center. C.E.L. is supported in part by NIH (NHGRI) grant 5R01HG-00125707 and DOE grant DEFG0201ER63204.

Appendix 1: Nucleotide Substitution Model

In the model described by Neyman [1971] and Felsenstein [1981], calibrated via the technique of Tajima and Nei [1982], Lanave et al. [1984], and Rodríguez et al. [1990], and slightly expanded for presentation here, $\Pr[b_{\rm des}|b_{\rm anc}]$, the chance that a descendant will show nucleotide $b_{\rm des}$ when an ancestor shows nucleotide $b_{\rm anc}$ is given by a matrix

$$\begin{aligned} M_x &= \begin{pmatrix} \Pr[A|A] & \Pr[T|A] & \Pr[C|A] & \Pr[G|A] \\ \Pr[A|T] & \Pr[T|T] & \Pr[C|T] & \Pr[G|T] \\ \Pr[A|C] & \Pr[T|C] & \Pr[C|C] & \Pr[G|C] \\ \Pr[A|G] & \Pr[T|G] & \Pr[C|G] & \Pr[G|G] \end{pmatrix} \\ &= e^{-\gamma k x} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ &\quad + (1 - e^{-\gamma k x}) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} x &= \text{the evolutionary ``distance'' between the ancestral and descendant individuals,} \\ m &= \text{the expected number of nucleotide mismatches between the two individuals,} \\ x &= -\frac{\ln(1 - km)}{k}\ , \\ \theta_b &= \text{the equilibrium probability of nucleotide } b, \\ \gamma &= \text{the relative rate of mutation, and} \\ k &= \frac{1}{1 - (\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2)}\ . \end{aligned}$$

This model has the necessary features that as $x \to 0^+$, the transition matrix is the identity matrix; that as $x \to +\infty$, the transition matrix gives an equilibrium distribution independent of which nucleotide we started with (*i.e.*, all of the rows are equal); and that $M_{a+b} = M_a M_b$, correctly modeling that the transition resulting from evolution described by an evolutionary distance a, followed by evolution described by an evolutionary distance b, is equal to the evolution described by the sum of the evolutionary distances.

For non-functional alignment positions (*i.e.*, those which are not coding, not regulatory, etc.), the equilibrium distribution $\vec{\theta}$ is chosen to be $\theta_A = 0.2967$, $\theta_T = 0.3122$, $\theta_C = 0.1949$, $\theta_G = 0.1962$ (or something similar that is representative of the non-functional DNA in the genomes in question) and can be chosen to be different to indicate functional positions. The rate of mutation γ is 1 for non-functional alignment positions and can be chosen to be a value between 0 and 1 to indicate functional positions.

The scale factor k is chosen to calibrate the evolutionary distances in the tree, so that when x and m are small, they are nearly the same. That is, k is chosen so that $\lim_{m \to 0^+} x/m = 1$ when $\gamma = 1$.

Appendix 2: Alternate Background Models

There is a strong indication in many settings that other nucleotide substitution models, such as those that recognize the difference between transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions (other changes in the nucleotides), are more powerful than is the equilibrium-based model that we have chosen, [Kimura, 1980, Hasegawa et al., 1985, Tavaré, 1986, Yang, 1994, Felsenstein and Churchill, 1996]. The use of our chosen model with an equilibrium distribution on nucleotide pairs other than the near-uniform distribution that we have employed here, is also an alternative possibility that proves important in other settings.

Our analyses using models other than the Felsenstein model with the chosen distribution did result in changes in the individual qualities of the various possible subtrees to some extent. However, in no case did they yield a change in the ordering of these qualities, and hence, none of them indicates a different priority order of species to sequence.

Appendix 3: Fisher information matrix approach

We measure the confidence intervals for $\vec{\theta}$ via the Fisher information matrix. Specifically, we suppose that a multiple-alignment sequence position's data set D is drawn randomly according to the model of Felsenstein, as parameterized by a distribution $\vec{\theta}^*$, and we measure the expected log likelihood of the estimator $\vec{\theta}$ via the formula

$$\log L(\vec{\theta}) \quad = \quad \sum_D \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\theta}^*] \ .$$

Intuitively, our confidence limits are tight if this function falls off quickly as $\vec{\theta}$ deviates from $\vec{\theta}^*$. This rate of decline is measured by the Hessian of $\log L(\vec{\theta})$, with respect to $\vec{\theta}$ and evaluated at $\vec{\theta} = \vec{\theta}^*$, and the matrix inverse of the negative of the Hessian is the covariance matrix for the $\vec{\theta}$ estimator. The matrix trace of the covariance matrix is the sum of the estimator variances for the parameters of the sought-for probability distribution on nucleotide pairs.

Because the components of $\vec{\theta}$ are constrained to sum to 1, there are three degrees of freedom in $\log L(\vec{\theta})$, rather than four. A possible choice for the degrees of freedom is given by the set of equations:

$$\theta_A \quad = \quad \psi_1 \ , \quad \theta_T = \psi_2 \ , \quad \theta_C = \psi_3 \ , \quad \theta_G = 1 - \psi_1 - \psi_2 - \psi_3$$

where each ψ_i value is nonnegative, and their sum is not more than 1. We have used this set of non-degenerate parameters, although any set of three linearly independent (not necessarily orthogonal) parameters will do. With this choice, the matrix of variances and covariances of the θ_b estimators is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \left(-\frac{\partial^2 \log L}{\partial \psi_{i'} \, \partial \psi_{i''}} \Big|_{\vec{\psi}=\vec{\psi}^*} \right)^{-1} \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \ .$$

The trace of this matrix, *i.e.*, the sum of the variances, can be computed as the sum of all of the off-diagonal elements of $\left(-\frac{\partial^2 \log L}{\partial \psi_{i'} \, \partial \psi_{i''}} \Big|_{\vec{\psi}=\vec{\psi}^*} \right)^{-1}$, plus twice the sum of its diagonal elements.

The confidence interval for the γ parameter was evaluated in a similar fashion. We defined the expected log likelihood of the estimator γ via the formula

$$\log L(\gamma) \quad = \quad \sum_D \log(\Pr[D|\gamma]) \Pr[D|\gamma^*] \ .$$

We evaluated the second derivative of $\log L(\gamma)$, with respect to γ and evaluated it at $\gamma = \gamma^*$. The reciprocal of the negative of this second derivative is the variance for the γ estimator.

References

D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, February 28 2003.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, April 25 1997.

A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27(23):4636–4641, December 1 1999.

E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucia, E. Kirkness, C. Rossier, and S. E. Antonarakis. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, 302(5647):1033–1035, November 7 2003.

J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*, 13(1):93–104, January 1996.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

W. Jang, A. Hua, S. V. Spilson, W. Miller, B. A. Roe, and M. H. Meisler. Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res*, 9(1):53–61, January 1999.

Ö. Johansson, W. Alkema, W. W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: The MSCAN algorithm. *Bioinformatics*, 19 Suppl. 1:i169–i176, July 2003.

M. G. Kendall and A. Stuart. *Kendall's Advanced Theory of Statistics*. Edward Arnold, 1998.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, December 1980.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):86–93, 1984.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, October 8 1993.

C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

B. Lenhard, A. Sandelin, L. Mendoza, P. Engström, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, May 22 2003.

J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Markovian structures in biological sequence alignments. *J Am Stat Assoc*, 94(445):1–15, March 1999.

E. H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res*, 13(12):2507–2518, December 2003.

L. A. McCue, W. Thompson, C. S. Carmack, and C. E. Lawrence. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res*, 12(10):1523–1532, October 2002.

L. A. McCue, W. Thompson, C. S. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, 29(3):774–782, February 1 2001.

J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S. S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, NY, 1971.

J. C. Oeltjen, T. M. Malley, D. M. Muzny, W. Miller, R. A. Gibbs, and J. W. Belmont. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res*, 7(4):315–329, April 1997.

N. Rajewsky, N. D. Socci, M. Zapotocky, and E. D. Siggia. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*, 12(2): 298–308, February 2002.

M. G. Reese, F. H. Eeckman, D. Kulp, and D. Haussler. Improved splice site detection in Genie. *J Comput Biol*, 4(3): 311–323, Fall 1997.

M. Rijnkels, L. Elnitski, W. Miller, and J. M. Rosen. Multispecies comparative analysis of a mammalian-specific genome domain encoding secretory proteins. *Genomics*, 82(4):417–432, October 2003.

F. Rodríguez, J. L. Oliver, A. Marín, and J. R. Medina. The general stochastic model of nucleotide substitution. *J Theor Biol*, 142(4):485–501, February 22 1990.

S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. Pipmaker—a web server for aligning two genomic DNA sequences. *Genome Res*, 10(4):577–586, April 2000.

A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, December 5 2003. [Epub ahead of print].

J. L. Slightom, J. H. Bock, D. A. Tagle, D. L. Gumucio, M. Goodman, N. Stojanovic, J. Jackson, W. Miller, and R. Hardison. The complete sequences of the galago and rabbit β-globin locus control regions: Extended sequence and functional conservation outside the cores of DNase hypersensitive sites. *Genomics*, 39(1):90–94, January 1 1997.

F. Tajima and M. Nei. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J Mol Evol*, 18(2):115–120, 1982.

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

J. W. Thomas, J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell, B. Maskeri, N. F. Hansen, M. S. Schwartz, R. J.

Weber, W. J. Kent, D. Karolchik, T. C. Bruen, R. Bevan, D. J. Cutler, S. Schwartz, L. Elnitski, J. R. Idol, A. B. Prasad, S. Q. Lee-Lin, V. V. Maduro, T. J. Summers, M. E. Portnoy, N. L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C. P. Brinkley, S. Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S. L. Ho, M. C. Huang, E. Karlins, P. L. Laric, R. Legaspi, M. J. Lim, Q. L. Maduro, C. A. Masiello, S. D. Mastrian, J. C. McCloskey, R. Pearson, S. Stantripop, E. E. Tiongson, J. T. Tran, C. Tsurgeon, J. L. Vogt, M. A. Walker, K. D. Wetherby, L. S. Wiggins, A. C. Young, L. H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C. L. Shu, P. J. De Jong, C. E. Lawrence, A. F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E. D. Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793, August 14 2003.

Z. Yang. Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39(1):105–111, July 1994.