How Much Is Another Mammalian Genome Worth?

Lee A. Newberg and Charles E. Lawrence

Bioinformatics Center
Wadsworth Center
NYS Department of Health

Department of Computer Science
Rensselaer Polytechnic Institute

RECOMB Workshop on Regulatory Genomics, March 26, 2004

Observations

Multiple transcription factor binding sites within a species	→	a position-weight matrix (*a.k.a.*, motif) describing the sites.
Add sites from the same species	→	better motif confidence.
Add sites from other species	→	better motif confidence, but ... evolutionarily close species are highly correlated.

Overview

The Questions

How much do the binding sites from other species help

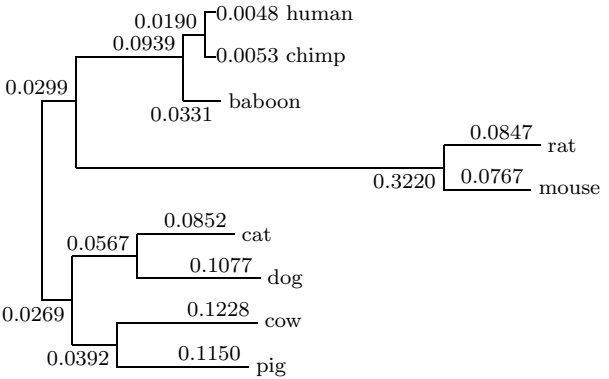
1. ... with motif confidence?
2. ... with finding cross-species conserved sequence?

The Answers

1. Not as much as hoped.
2. Quite a bit.

The Phylogenetic Model

For aligned sequences



$$\theta_A = 0.2967, \theta_T = 0.3122, \theta_C = 0.1949, \theta_G = 0.1962.$$

Adam Siepel and David Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21(3):468–488, March 2004.

J. W. Thomas *et al.*. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793, August 14 2003.

Nucleotide Substitution Model (Historical)

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

where x = the evolutionary distance between the ancestral and descendant individuals, and

θ_b = the equilibrium probability of nucleotide b.

It behaves well: M_0, $M_{+\infty}$, M_{x+y}, reversible.

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

Probability of Aligned Sequence Data

Bottom-up tree algorithm. Two alternating general steps:

$$\text{Edge traversal:} \quad \Pr[D_c|b_p] = \sum_{b_c \in \{A,T,C,G\}} \Pr[D_c|b_c] \overbrace{\Pr[b_c|b_p]}^{M_x}$$

$$\text{Vertex traversal:} \quad \Pr[D_p|b_p] = \prod_{c \in \{p\text{'s children}\}} \Pr[D_c|b_p]$$

where

p = parent/ancestral node of an edge

c = child/descendant node of an edge

D_x = observed data in the leaves of the subtree rooted at x

b_x = nucleotide/base for node x

Jerzy 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.

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

Nucleotide Substitution Model (Updated)

$$M_x = e^{-\gamma k x} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (1 - e^{-\gamma k x}) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix}$$

$$\text{where} \quad k = \frac{1}{1 - (\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2)}$$

$$\gamma = \text{the relative rate of mutation}$$

$$\gamma \in (0, 1] \; .$$

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

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

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.

Consensus Distribution Estimator (Overview)

Maximum Likelihood Estimate

For each motif column choose $\vec{\theta}$ to maximize the probability of the n observed aligned sites' data, $\vec{D} = (D_1, \ldots, D_n)$. (Each D_i is a tuple of nucleotides, one nucleotide from each species.)

$$\begin{aligned} \Pr[\vec{D}|\vec{\theta}] &= \prod_{i=1}^{n} \Pr[D_i|\vec{\theta}] \\ \mathrm{LL}(\vec{\theta}) &= \log \Pr[\vec{D}|\vec{\theta}] = \sum_{i=1}^{n} \log \Pr[D_i|\vec{\theta}] \\ \hat{\theta} &= \operatorname*{argmax}_{\vec{\theta}} \Big(\mathrm{LL}(\vec{\theta}) \Big) \end{aligned}$$

Confidence Ellipsoid

How fast does

$$\mathrm{LL}(\vec{\theta})$$

fall off as $\vec{\theta}$ deviates from $\hat{\theta}$?

Consensus Distribution Estimator (Detail)

Maximum Expected Log Likelihood Estimate

Integrate out the data. Asymptotic analysis.

$$\log \Pr[D_i|\vec{\theta}] \;\rightarrow\; \sum_D \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\theta}^*]$$

$$\mathrm{LL}(\vec{\theta}) = \sum_{i=1}^{n} \log \Pr[D_i|\vec{\theta}] \;\rightarrow\; \mathrm{LL}(\vec{\theta}) = n \sum_D \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\theta}^*]$$

Confidence Ellipsoid

$\mathrm{LL}(\vec{\theta})$ shape at maximum. Asymptotic normality / Fisher information matrix:

$$\mathrm{LL}(\vec{\theta}) \sim \log \frac{\exp\left[-\frac{(\vec{\theta}-\vec{\theta}^*)^T V^{-1} (\vec{\theta}-\vec{\theta}^*)}{2}\right]}{\sqrt{\det(2\pi V)}} \;\rightarrow\; \mathrm{Var}[\hat{\theta}] \sim \left(-\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\mathrm{LL}(\vec{\theta})\Big|_{\vec{\theta}=\vec{\theta}^*} \right)^{-1}$$

Maurice G. Kendall and Alan Stuart. *Kendall's Advanced Theory of Statistics*, Volume 1: Distribution Theory. Edward Arnold, 1994.

Consensus Distribution Estimator (Detail)

Confidence Ellipsoid

$$\begin{aligned} \mathrm{LL}(\vec{\theta}) &= n \sum_{D} \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\theta^*}] \\ \mathrm{Var}[\hat{\theta}] &= \left(-\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\mathrm{LL}(\vec{\theta})\Big|_{\vec{\theta}=\vec{\theta^*}} \right)^{-1} \end{aligned}$$

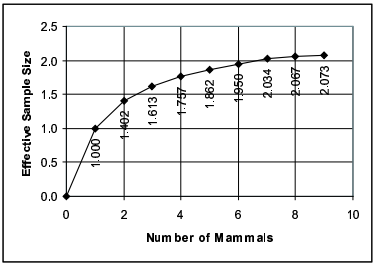
Efficiency = Effective Number of Observations

Proportional to:

$$\frac{1}{\mathrm{trace}\left(\mathrm{Var}[\hat{\theta}]\right)}$$

S. D. Silvey. *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Chapman and Hall, 1980.

Effective Number of Independent Species



	Species	# 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

Question #2: Conservation

γ-Proteobacteria

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

Metazoans

Boris Lenhard, Albin Sandelin, Luis Mendoza, Pär Engström, Niclas Jareborg, and Wyeth W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, May 22 2003.

Monkeys

Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D. Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611): 1391–1394, February 28 2003.

Yeast

- Tong Ihn Lee *et al.*. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, October 25 2002.

- Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 15 2003.

- Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A. Cohen, and Mark Johnston. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, July 4 2003.

Etc.

Conservation Estimator

Maximum Expected Log Likelihood Estimate

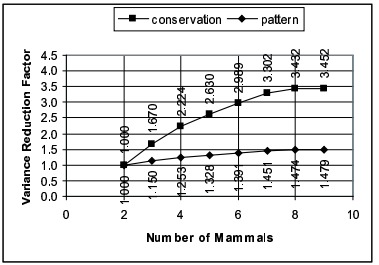
$$
\begin{aligned}
M_x &= e^{-\gamma k x} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (1 - e^{-\gamma k x}) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \\
\mathrm{LL}(\gamma) &= n \sum_D \log(\Pr[D|\gamma]) \Pr[D|\gamma^*] \\
\mathrm{Var}[\hat{\gamma}] &= 1 / \left(-\frac{\partial^2}{\partial \gamma^2} \mathrm{LL}(\gamma) \Big|_{\gamma=\gamma^*} \right)
\end{aligned}
$$

Efficiency

Proportional to:

$$
-\frac{\partial^2}{\partial \gamma^2} \mathrm{LL}(\gamma) \Big|_{\gamma=\gamma^*}
$$

Efficiency Relative to Human with Mouse



	Species	Pattern	Conservation
1	*Homo sapiens* / human		
2	*Mus musculus* / mouse	1.000	1.000
3	*Canis familiaris* / dog	1.150	1.670
4	*Bos taurus* / cow	1.253	2.224
5	*Sus scrofa* / pig	1.328	2.630
6	*Rattus norvegicus* / rat	1.391	2.989
7	*Felis catus* / cat	1.451	3.302
8	*Papio cynocephalus anubis* / baboon	1.474	3.432
9	*Pan troglodytes* / chimpanzee	1.479	3.452

Conclusion

With more mammalian genomes

- ... the locating of transcription factors via their conservation will be significantly enhanced,
- ... but it appears that subsequent computational identification of the sequence patterns of functional sites will not be as easy.

For more information

`http://www.rpi.edu/~newbel/publications/NewbergLawrenceCSHL2004.pdf`

Only Three Degrees of Freedom

With

$$\begin{aligned} \theta_A &= \psi_A \\ \theta_T &= \psi_T \\ \theta_C &= \psi_C \\ \theta_G &= 1 - \psi_A - \psi_T - \psi_C \end{aligned}$$

we compute

$$\mathrm{Var}[\hat{\theta}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \left(-\frac{\partial}{\partial \psi_i} \frac{\partial}{\partial \psi_j} \mathrm{LL}(\vec{\psi}) \Big|_{\vec{\psi}=\vec{\psi}^*} \right)^{-1} \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$