Phylogenetic Motif Detection by Expectation-Maximization on Evolutionary Mixtures

A.M. Moses, D.Y. Chiang, and M.B. Eisen

PHYLOGENETIC MOTIF DETECTION BY EXPECTATION-MAXIMIZATION ON EVOLUTIONARY MIXTURES

A.M. MOSES

Graduate group in Biophysics and Center for Integrative Genomics, University of California, Berkeley Email: amoses@ocf.berkeley.edu

D.Y. CHIANG

Department of Molecular and Cell Biology, University of California, Berkeley Email: dchiang@ocf.berkeley.edu

M.B. EISEN

Department of Genome Sciences, Lawrence Berkeley Lab and Department of Molecular and Cell Biology, University of California, Berkeley Email: mbeisen@lbl.gov

1 Abstract

The preferential conservation of transcription factor binding sites implies that non-coding sequence data from related species will prove a powerful asset to motif discovery. We present a unified probabilistic framework for motif discovery that incorporates of evolutionary information. We treat aligned DNA sequence as a mixture of evolutionary models, for motif and background, and, following the example of the MEME program, provide an algorithm to estimate the parameters by Expectation-Maximization. We examine a variety of evolutionary models and show that our approach can take advantage of phylogenic information to avoid false positives and discover motifs upstream of groups of characterized target genes. We compare our method to traditional motif finding on only conserved regions. An implementation will be made available http://rana.lbl.gov.

2 Introduction

A wide range of biological processes involve the activity of sequence-specific DNA binding proteins, and an understanding of these processes requires the accurate elucidation of these proteins' binding specificities. The functional binding sites for a given protein are rarely identical, with most proteins binding to families of related sequences collectively referred to as their 'motif' [1]. Although experimental methods exist to identify sequences bound by a specific protein, they have not been widely applied, and computational approaches [2,3,4] to 'motif discovery' have proven to be a useful alternative. For example, the program MEME [5], models a collection of

sequences as a mixture of multinomial models for motif and background and uses an Expectation-Maximization (EM) algorithm to estimate the parameters.

Because functional binding sites are evolutionarily constrained, their preferential conservation relative to background sequence has proven a useful approach for their identification [6]. With the availability of complete genomes for closely related species e.g., [7], it is possible to incorporate an understanding of binding site evolution into motif discovery as well. At present, few motif discovery methods simultaneously take advantage of both the statistical enrichment of motifs and the preferential conservation of the sequences that match them. One recent study [7] enumerated spaced hexamers that were both preferentially conserved (in multiple sequence alignments) and statistically enriched. Another method, FootPrinter, [8] identifies sequences (with mismatches) with few changes over an evolutionary tree. Neither of these methods, however, makes use of an explicit probabilistic model.

Here we present a unified probabilistic framework that combines the mixture models of MEME with probabilistic models of evolution, and can thus be viewed as an evolutionary extension of MEME. These evolutionary models (used in the maximum likelihood estimation of phylogeny [9]) consider observed sequences to have been generated by a continuous time Markov substitution process from unobserved ancestral sequences, and can accurately model the complicated statistical relationship between sequences that have diverged along a tree from a common ancestor. Our approach considers observed sequences to have been generated from ancestral sequences that are two component mixtures of motif and background, each with their own evolutionary model. The value of varying evolutionary models has been realized in other contexts as well, e.g., [10] and such models have been successfully trained using EM [11]. A mixture of evolutionary models has been used previously to identify slowly evolving non-coding sequences [12], and this work can equally be regarded as an extension of that approach. Given a set of aligned sequences, we use an EM algorithm to obtain the maximum likelihood estimates of the motif matrix and a corresponding evolutionary model.

3   Methods

3.1   Probabilistic model

We first describe the probabilistic framework used to model aligned non-coding sequences. We employ a mixture model, which can be written generically as

$$p(data) = \sum_{models} p(model)p(data|model)$$

where $p(x)$ is the probability density function for the random variable x. The sum over models indicates that the data is distributed as some mixture of component models, where the prior, $p(model)$, is the mixing proportion. For simplicity, we first address the case of pair-wise sequence alignments.

Given some motif size, w, we treat the entire alignment as a series of alignments of length w, each of which may be an instance of the motif or a piece of background sequence. We denote the pair of aligned sequences as X and Y, where the ith position in the sequence as a vector of length 4, (for each of ACGT), where $X_{ib}=1$ if the bth base is observed, and 0 otherwise. We denote the unobserved ancestral sequence, A, similarly, except that the values of A_{ib} are not observed. For a series of alignments of total length N, the likelihood, L, is given by

$$L = \prod_{i=0}^{N-w} \sum_{m_i} p(m_i) \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k \mid A_{kb}, m_i)p(A_{kb} \mid m_i)$$

where the m_i are unobserved indicator variables indexing the component models; in our case m is either motif or background. Generically, we let

$$p(m_i) = \pi_m,$$

the prior probability for each component.

We incorporate the sequence specificity of the motif by letting the prior probabilities of observing each base in the ancestral sequence, $p(A_{kb}|m_i)$, be the frequency of each base at each position in the motif (the frequency matrix). We write

$$p(A_{kb}|m_i) = f_{mkb},$$

such that if m is motif, f_{mkb} gives the probability of observing the bth base at the k-i^{th} position. For the background model we use the average base frequencies for each alignment, and assume that they are independent of position. This allows us to run our algorithm on several alignments simultaneously [15] and the densities are therefore conditioned on the alignment as well, but omit this here for notational clarity.

Finally, noting that because the two sequences descended independently from the ancestor, we can write $p(X_k, Y_k|A_{kb}, m_i) = p(X_k|A_{kb}, m_i) p(Y_k|A_{kb}, m_i)$, where $p(X_k|A_{kb}, m_i)$ is the probability of the residue X_k, given that the ancestral sequence, A, was base b at that position – a substitution matrix for each component model. For simplicity we use the Jukes-Cantor [16] substitution matrix, which is, in our notation,

$$p(X_k \mid A_{kb}, m_i) = \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}\alpha_{mk}}\right)^{X_{kb}} \left(\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}\alpha_{mk}}\right)^{1-X_{kb}}$$

where α_{mk} is the rate parameter at position k.

It is here that we incorporate differences in evolution between the motif and background by specifying different substitution matrices for each component. For example, if we set α_m smaller for the motif than for background, the motif evolves at a slower rate than the background – it is conserved. We test a variety of different substitution models for the motif and summarize the implications for motif discovery in the Gcn4p targets. (See results) Unfortunately, as the dependence of these models on the equilibrium frequencies becomes more complicated, deriving ML estimators for the parameters becomes more difficult, and more general optimization methods may be necessary. Once again, we can allow each alignment its own background rate, [15] and express the motif rate as a proportion of background.

3.2 An EM algorithm to train parameters

Following the example of the MEME program [5] which uses an EM (an iterative optimization scheme guaranteed to find local maxima in the likelihood) algorithm to fit mixtures to unrelated sequences, we now derive an EM algorithm to train the parameters of the model described above. We write the 'expected complete log likelihood' [17]

$$\langle \ln L_c \rangle = \sum_{i=0}^{N-w} \sum_{m_i} \langle m_i \rangle \left[\ln \pi_m + \sum_{k=i}^{i+w-1} \sum_{b=0}^{3} \langle A_{kb} \rangle \left(\ln p(X_k, Y_k \mid A_{kb}, m_i) + \ln f_{mkb} \right) \right]$$

where ln denotes the natural logarithm, and maximize by setting the derivatives with respect to the parameters to zero at each iteration. Setting

$$\frac{\partial \langle \ln L_c \rangle}{\partial \pi_m} = 0, \ \frac{\partial \langle \ln L_c \rangle}{\partial f_{mkb}} = 0 \text{ and } \frac{\partial \langle \ln L_c \rangle}{\partial \alpha_{mk}} = 0$$

and solving gives

$$\pi_m = \frac{1}{N-w} \sum_i \langle m_i \rangle, \ f_{mkb} = \frac{\sum_i \langle m_i \rangle \langle A_{kb} \rangle}{\sum_i \langle m_i \rangle} \text{ and } \alpha_{mk} = -\frac{3}{4} \ln \left(\frac{1 - \frac{4}{3} R_{mk}}{1 + R_{mk}} \right)$$

where R_{km} is the ratio of expected changed to identical residues under each model, and is given by

$$R_m = \frac{\sum_{i=0}^{N-w} \langle m_i \rangle \sum_{k=i}^{i+w-1} \sum_{b=0}^{3} \langle A_{kb} \rangle (2 - Y_{kb} - X_{kb})}{\sum_{i=0}^{N-w} \langle m_i \rangle \sum_{k=i}^{i+w-1} \sum_{b=0}^{3} \langle A_{kb} \rangle (X_{kb} + Y_{kb})}$$

for all k in the case of a constant rate across the motif. The sufficient statistics $\langle A_{kb} \rangle$ and $\langle m_i \rangle$, are derived by applying Bayes' theorem and are computed using the values of the parameters from the previous iteration. We have

$$\langle m_i \rangle = p(m_i | X, Y) = \frac{p(m_i) p(X, Y | m_i)}{p(X, Y)}$$

where

$$p(X, Y | m_i) = \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k | A_{kb}, m_i) p(A_{kb} | m_i)$$

and

$$p(X, Y) = \sum_{m_i} p(X, Y | m_i) p(m_i)$$

Similarly,

$$\langle A_{ib} \rangle = p(A_{ib} | X_i, Y_i) = \sum_{m_i} p(A_{ib} | X_i, Y_i, m_i) p(m_i) = \sum_{m_i} \frac{p(A_{ib}) p(X_i, Y_i | A_{ib}, m_i)}{p(X_i, Y_i | m_i)} p(m_i)$$

In order to extend these results beyond pair-wise alignments, we can simply replace the two sequences X and Y with the probability of the entire tree below conditioned on having observed base b in the ancestral sequence. The likelihood becomes

$$L = \prod_{i=0}^{N-W} \sum_{m_i} p(m_i) \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(tree | A_{kb}) p(A_{kb} | m_i)$$

where $p(tree|A_{kb})$ are computed using the 'pruning' algorithm [9]. Of course, a tree topology is needed in these cases and we used the accepted topology for the *sensu stricto Saccharomyces* [7] and computed for each alignment the maximum likelihood branch lengths using the paml package [18].

3.3   Implementation

We implemented a C++ program (EMnEM: Expectation-Maximization on Evolutionary Mixtures) to execute the algorithm described above, with the following extensions. Because instances of a motif may occur on either strand of DNA sequence, we also treat the strand of each occurrence as a hidden variable, and sum over the two possible

orientations. In addition, because the mixture model treats each position in the alignment independently, we down-weight overlapping matches by limiting the total expected number of matches in any window of $2w$ to be less than one. Finally, because EM is guaranteed only to converge to a local optimum in the likelihood, we need to initialize the model in the region of the likelihood space where we believe the global optimum lies. Similar to the strategy used in the MEME program [5], we initialize the motif matrix with the reconstructed ancestral sequence of length w at each position in the alignments, and perform the full EM starting with the sequence at the position that had the greatest likelihood. EMnEM will be made available at http://rana.lbl.gov.

3.4 Time complexity

The time complexity of the EM algorithm is linear with total length of the data, and the initialization heuristic we have implemented is quadratic with the length. Interestingly, because our algorithm runs on aligned sequences, relative to MEME, which that treats sequences independently, the total length is reduced by a factor of $^1/_S$, where S is the number of sequences in the alignment. Usually, we lose this factor in each iteration when calculating $p(tree|A_{kb})$ using the 'pruning' algorithm [9], as it is linear in S. We note, however, that for evolutionary models (e.g., Juckes-Cantor) where $p(tree|A_{kb})$ is independent of $p(A_{kb}|m_i)$, we may learn the PSPM without re estimating the sufficient statistics $\langle A_{kb} \rangle$ (the reconstructed ancestral sequence) at each iteration. In these cases the complexity of EMnEM will indeed be linear in the length of the *aligned* sequence, a considerable speedup, especially in the quadratic initialization step.

4 Results and Discussion

4.1 A test case from the budding yeasts

In order to compare our algorithm under various evolutionary models as well as to other motif discovery strategies, we chose to compare all methods on a single test case: the upstream regions from 5 *sensu stricto Saccharomyces* (*S. bayanus*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, and *S. paradoxus*) of 9 known Gcn4p targets that are listed in SCPD [19]. In order to control for variability in alignment quality at different evolutionary distances, we made multiple alignments of all available upstream regions using T-coffee [20] and then extracted the appropriate sequences for any subset of the species. The Gcn4p targets from SCPD are a good set on which to test our method because there are a relatively high number of characterized sites in these promoters. In addition, the upstream regions of these genes contain stretches of poly T, which are not

known to be binding sites. As a result, MEME ("tcm" model, w 10) assigns a lower (better) evalue to a 'polyT' motif (e=2.7e-03) than to the known Gcn4p motif (e=1.6e06) when run on the *S. cerevisiae* upstream regions. Because this is typical of the types of false positives that motif finding algorithms produce, we use as an indicator of the success of our method the log ratio of the likelihood of the evolutionary mixture model using the real Gcn4p matrix, to that using the polyT matrix. If this indicator is greater than zero, i.e.,

$$log \left[\frac{p(data|Gcn4p)}{p(data|polyT)} \right] > 0$$

the real motif has a greater likelihood than the false positive, and should be returned as the top motif.

4.2 *Incorporating a model of motif evolution can eliminate false positives*

In order to explore the effects of incorporating models of motif evolution into motif detection, we tested several evolutionary models. In particular we were interested in the effect of incorporating evolutionary rate, as real motifs evolve slower than surrounding sequences. Using alignments of *S. cerevisiae* and *S. mikatae*, we calculated the log ratio of the likelihood using the real Gcn4p matrix to the likelihood using the polyT matrix with Jukes-Cantor substitution under several assumptions about the rate of evolution in the motif (Figure 1). Interestingly, slower evolution in the motif, either ¼ or 0.03 (the ML estimate) times background rate, is enough to assign a higher likelihood to the Gcn4p motif and thus eliminate the false positive. We tried two additional evolutionary models, in which the rate of substitution at each position depends on the frequency matrix. In the Felsenstein '81 model (F81) the different types of changes occur at different rates, but the overall rate at each position is constant, while the Halpern-Bruno model (HB) assumes there is purifying selection at each position and can account for positional variation in overall rate [21,22]. In each case, these more realistic models further favored the Gcn4p matrix over the polyT.
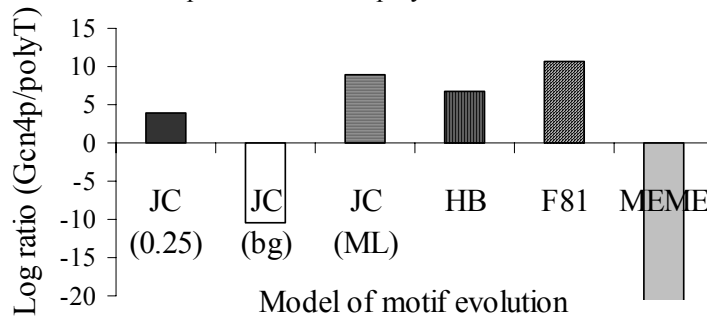
Figure 1. Effect of models for motif evolution on motif detection Plotted is the log ratio of the likelihood using the Gcn4p PSPM to the likelihood using polyT PSPM under various evolutionary models in alignments of *S. cerevisiae* to *S. mikatae*. Models that allow the motif to evolve more slowly than background, JC (0.25), JC (ML) and JC (HB), and models in which the rates of evolution take into account the deviation from equilibrium base frequencies, F81 and JC (HB), assign higher likelihood to the Gcn4p PSPM. Also plotted is the negative log ratio of the e-values from MEME ('tcm' model, w 10). JC are Jukes-Cantor models with rate parameter equal to background (bg), ¼ of background (0.25) or set to the maximum-likelihood estimate below background (ML).

4.3   Success of motif discovery is dependent on evolutionary distance

In order to test the generality of the results achieved for the *S. cerevisiae S. mikatae* alignments, we calculated the log ratio of the likelihood of the evolutionary mixture using the real Gcn4p matrix to the polyT matrix over a range of evolutionary distances and rates of evolution (figure 2, filled symbols). At closer distances, more of the data is redundant, while over longer comparisons, conserved sequences should stand out more against the background. Indeed, at the distance of *S. cerevisiae* to *S. paradoxus* (~0.13 substitutions per site), the likelihood of polyT is greater, while at the distance of *S. cerevisiae*, *S. mikatae*, and *S. paradoxus* (~0.31 subs. per site) the Gcn4p matrix is favored. Interestingly, this is true regardless of the rate of evolution assumed for the motif. While at all evolutionary distances slow evolution favors the Gcn4p matrix *more* than when the motif evolves at the background rate, the effect of including slower evolution is smaller than the effect of the varying evolutionary distance. Only at the borderline distance of *S. cerevisiae* to *S. mikatae* (~0.25 subs. per site), do the models perform differently. We also ran MEME (with the "tcm" model, w set at 10) on the all sequences (from all genes and all species) and calculated the negative log ratio of the MEME e-values for the two motifs (figure 2, heavy trace). MEME treats all the sequences independently, and continues to assign the polyT matrix a lower e-value over all the evolutionary distances. At least for this case, it seems more important to accurately model the phylogenetic relationships between the sequences (i.e., using a tree) than to accurately model the evolution within the motif.
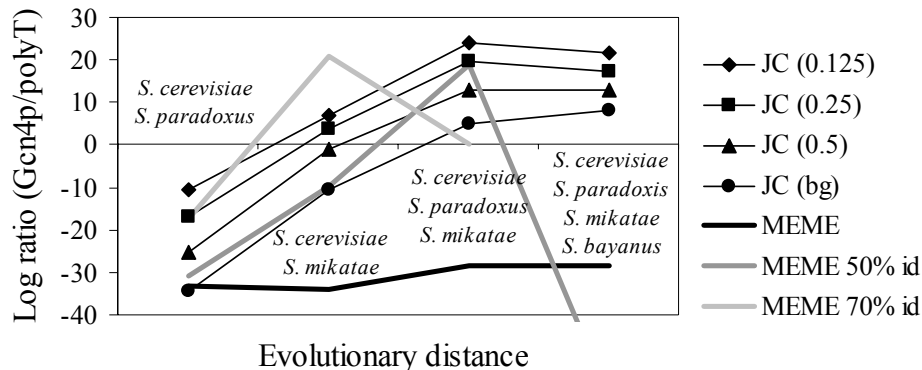
Figure 2. Effect of evolutionary distance on motif detection. Log ratio of the likelihood using the Gcn4p matrix to the likelihood using polyT matrix and alignments that span increasing evolutionary distance. At distances greater than *S. cerevisiae* to *S. mikatae* the evolutionary mixture assigns the Gcn4p matrix a greater likelihood whether the rate of evolution in the motif is equal to, ½, ¼ or ⅛ of the background rate, (diamonds, squares, triangles and circles, respectively). Also plotted are negative log ratios of the MEME evalues for the Gcn4p to polyT, using the entire sequences, or pre-filtering alignments for 20 base pair windows of at least 70% or 50% identity to a reference genome (heavy, lighter and lightest traces, respectively.)

4.4 The unified framework is preferable to using evolutionary information separately

In order to compare our method, which incorporates evolutionary information directly into motif discovery, to approaches that use such information separately, we scanned the alignments at each evolutionary distance and removed regions than were less than 50 or 70 % identical to a reference genome in a 20 base pair window. This allows MEME, which does take into account phylogenic information, to focus on the conserved regions. We ran MEME and computed the negative log ratio of the e-values for the Gcn4p matrix and the polyT matrix. While in both cases there were distances where the real motif was favored (figure 2, lighter traces), the effect of the filtering was not consistent. At distances too close, not enough is filtered out, and the polyT is still preferred, while at distances too far, real instances of the motif will no longer pass the cutoff and the real motif is no longer recovered (figure 2, lighter traces). Thus, while incorporating evolutionary information separately can help recover the real motif, it depends critically on the choice of percent identity cutoff.

4.5 Examples of other discovered motifs

We ran both our program and MEME on the upstream regions of target genes of some transcription factors with few characterized targets and/or poorly defined motifs In several cases, for a given motif size, our algorithm ranked a plausible motif first, and MEME ranked a polyT motif first (see Table 1).

Genes	Binding factor	EMnEM rank	MEME rank	Motif
HEM13, RTT101, ROX1	Rox1p +	1	2	TATTGTTC
ERG2, ERG3, ERG9, UPC2	Upc2p ++	1	2	TCTAAACGAA
RNR2, RNR3, RNR4, RFX1	Rfx1p ++	1	2	GTTGCCAGAC
CDC19, PGK1, TPI1, ENO1, ENO2, ADH1	Gcr1p +	1	2	CTTCCACTA
ARO80, ARO9, ARO10	Aro80p ++	1	1	
TRR1, TRX2, GSH1, SSA1, AHP1	Yap1p ++	-	-	
ZRG17, ZRC1, FET4	Zap1p ++	3	2	

Table 1. Motif discovery using EMnEM and MEME. The EMnEM program was run using the Jukes Cantor model for motif evolution with the rate set to ¼ background (JC 0.25) on *S. cerevisiae S. mikatae* alignments in each case. For cases where EMnEM ranked the motif higher, the consensus sequence and a plot of the information content is shown. MEME was run on the unaligned sequences from both species simultaneously. Target genes are from SCPD[20] (+) or YPD [23] (++). – indicates that a plausible motif was not found.

5   Conclusions and future directions

We have provided an evolutionary mixture model for transcription factor binding sites in aligned sequences, and a motif finding algorithm based on this framework. We believe that our approach has many advantages over current methods; it produces probabilistic models of motifs, can be applied directly to multiple or pair-wise alignments, and can be applied simultaneously at multiple loci. Our method should be applicable to any group of species whose intergenic regions can be aligned, though because alignments may not be possible at large evolutionary distances, our reliance on them is a disadvantage of our method relative to FootPrinter [18]. It is not difficult to conceive of extending this framework to unaligned sequences by treating the alignment as a hidden variable as well; unfortunately, the space of multiple alignments is large, and improved optimization methods would certainly be needed.

In addition to motif discovery, our probabilistic framework is also applicable to binding site identification. Current methods that search genome sequence for matches to

motifs are also plagued by false positives, but optimally combining sequence specificity and evolutionary constraint may lead to considerable improvement.

6   Acknowledgements

7   References

1. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. (2000) Jan;**16**(1):16-23.
2. Stormo GD, Hartzell GW 3rd. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A*. (1989) Feb;**86**(4):1183-7.
3. Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*. 1990;**7**(1):41-51.
4. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*. (1993) Oct 8;**262**(5131):208-14.
5. Bailey TL, Elkan C, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, (1994.)
6. Hardison, Conserved noncoding sequences are reliable guides to regulatory elements, *Trends in Genetics*, (2000) Sep;**16**(9):369-372
7. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. (2003) May 15;423(6937):241-54.
8. Blanchette M, Schwikowski B, Tompa M. Algorithms for phylogenetic footprinting. *J Comput Biol.* (2002);**9**(2):211-23.
9. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. (1981);**17**(6):368-76.
10. Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*. 2000 Sep;**16**(9):760-6. Erratum in: *Bioinformatics* 2001 Mar;**17**(3):290

11. Holmes I, Rubin GM. An expectation maximization algorithm for training hidden substitution models. *J Mol Biol*. (2002) Apr 12;**317**(5):753-64.
12. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*. (2003) Feb 28;**299**(5611):1391-4.
13. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* (1997) Oct;**13**(5):555-6.
14. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning* Springer-verlag NY, (2001)
15. Yang, Z. Maximum likelihood models for combined analyses of multiple sequence data. *J Mol Evol.* **42**:587-596 (1996.)
16. Yang, Z., N. Goldman, and A. E. Friday. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol Biol Evol.* **11**:316-324 (1994)
17. M. I. Jordan, An Introduction to Probabilistic Graphical Models, in preparation.
18. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*. (1997) **13**(5):555-556
19. Zhu J, Zhang MQ. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*. (1999) Jul-Aug;**15**(7-8):607-611.
20. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. (2000) Sep 8;**302**(1):205-17.
21. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 1998 Jul;**15**(7):910-917.
22. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol.* (2003) **3**:18
23. Hodges PE, Payne WE, Garrels JI. The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae. *Nucleic Acids Res*. (1998) Jan 1;**26**(1):68-72.