Pacific Symposium on Biocomputing 13:489-500(2008)

USE OF AN EVOLUTIONARY MODEL TO PROVIDE EVIDENCE FOR A WIDE HETEROGENEITY OF REQUIRED AFFINITIES BETWEEN TRANSCRIPTION FACTORS AND THEIR BINDING SITES IN YEAST

RICHARD W. LUSK

Department of Molecular and Cell Biology, University of California, Berkeley
Berkeley, California 94720, USA
E-mail: lusk@berkeley.edu
www.berkeley.edu

MICHAEL B. EISEN

Genomics Division, Lawrence Berkeley National Laboratory, Department of Molecular and Cell Biology, University of California, Berkeley
Berkeley, California 94720, USA
E-mail: mbeisen@lbl.gov
www.lbl.gov

Keywords: binding sites, evolution, PWM, ChIP-chip, affinity

1. Abstract

The identification of transcription factor binding sites commonly relies on the interpretation of scores generated by a position weight matrix. These scores are presumed to reflect on the affinity of the transcription factor for the bound sequence. In almost all applications, a cutoff score is chosen to distinguish between functional and non-functional binding sites. This cutoff is generally based on statistical rather than biological criteria. Furthermore, given the variety of transcription factors, it is unlikely that the use of a common statistical threshold for all transcription factors is appropriate. In order to incorporate biological information into the choice of cutoff score, we developed a simple evolutionary model that assumes that transcription factor binding sites evolve to maintain an affinity greater than some factor-specific threshold. We then compared patterns of substitution in binding sites predicted by this model at different thresholds to patterns of substitution observed at sites bound *in vivo* by transcription factors in *S. cerevisiae.* Assuming that the cutoff value that gives the best fit between the observed and predicted values will optimally distinguish functional and non-functional sites, we discovered substantial heterogeneity for appropriate cutoff values among factors. While commonly used thresholds seem appropriate for many factors, some factors appear to function at cutoffs satisfied commonly in the genome. This evidence was corroborated by local patterns of rate variation for examples of stringent and lenient p-value cutoffs. Our analysis further highlights the necessity of taking a factor-specific approach to binding site identification.

2. Introduction

A gene's expression is governed largely by the differential recruitment of the basal transcription machinery by bound transcription factors.[1,2] In this way, transcription factor binding sites are fundamental components of the regulatory code, and this code's decipherment is partially a problem of recognizing their location and affinity.[3] These are usually determined using position weight matrices, although a number of more recently developed methods are beginning to become adopted.[4] We use position weight matrices here due to their ease of use with evolutionary analysis and their established theoretical ties with biochemistry. A position weight matrix generates a score comprising the log odds of a given subsequence being drawn from a binding site distribution of nucleotide frequencies vs. an analogous background distribution.[5] The score's p-value is used to determine the location of binding sites: subsequence scores above a predetermined cutoff designate that subsequence to be a binding site, and subsequence scores below the cutoff designate the subsequence to be ignored.

The interpretation of regulatory regions is thus dependent on the choice of the p-value cutoff. However, this choice is not straightforward, although it is commonly made to conform to established but biologically arbitrary statistical standards, e.g. $p < .001$. In addition to assuming that this particular p-value is appropriate, the user here also assumes that a single p-value is appropriate for all transcription factors. Being that score shares an approximately monotonic relationship with affinity,[6,7] this implies that the nature of the interaction between different transcription factors and their binding sites is the same. This may not be the case. For example, some transcription factors may require a stronger binding site to compensate for weaker interactions with other transcription machinery, and so a lenient cutoff would be inappropriate. Conversely, the choice of a stringent cutoff could eliminate viable sites of factors that commonly rely on cooperative interactions with other proteins to be recruited to the DNA. A single common standard of significance is a compromise that may not be reasonable.

Ideally, biological information should inform the choice of a p-value and its consequent ramifications in the determination of function. Several recent approaches have well used expression[8] and ChIP-chip[9] data towards understanding binding specificity. Here we take advantage of selective pressure as a third source of information. Tracking selective pressure has the advantage of directly interpreting sequence in terms of its value to the organism in its environment; to a degree, function can be inferred by observing the impact of selection. To this end, we propose a simple selective model of binding site evolution. Selection prevents the fixation of low affinity sites that may not affect expression to a satisfactory level and does not maintain unnecessary high affinity sites. We train the model on the ChIP-chip data available in yeast, and we find evidence for a wide heterogeneity in required binding site affinity between factors. Supporting recent work by Tanay,[10] many factors appear to require only weak affinity for function, and we find some evidence that these may rely on cooperative binding to achieve specificity.

3. Results and Discussion

3.1. *Definition and training of the affinity-threshold model*

In order to use selection as a means to investigate function, a model must be defined to describe how selection acts on functional and non-functional binding site sequence. Our model was created to be the simplest possible for our purposes. We assume that binding sites evolve independently from other sites in their promoter, but that all sites that bind the same factor evolve equivalently. We interpret a binding site's function in a binary manner: our model supposes that there exists a satisfactory level of expression and that binding site polymorphisms that are able to drive this expression level or greater have equal fitness, while binding site polymorphisms that cannot are deleterious. By assuming that this deleterious effect is large enough to preclude fixation in *S. cerevisiae*, our model imposes an effective threshold on permitted affinity: it does not allow a substitution to occur if it drops the position weight matrix score beneath a given boundary. Analogous reasoning lets us treat repressors identically. By imposing a threshold on permitted affinity and by relying on the assumption that position weight matrix score shares a monotonic relationship with affinity,[6] we impose a threshold weight matrix score.

Our purpose in training the model is to find where that threshold lies for each factor, which we accomplish using simulation. For any given threshold and matrix, we simulate the relative rates of substitution that would be expected, and then we compare these rates to empirically determined rates to choose the most appropriate threshold. The simulation is run as follows: we start with the matrix's consensus sequence, and make one mutation according to the neutral HKY[11] model. The sequence's score is evaluated: if it exceeds the threshold, the mutation is considered fixed and the count of substitutions at that position is incremented, and if not, no increment is made and the sequence reverts back to the original sequence. This mutate-select process is repeated. Assuming that the impact of polymorphism is negligible, removing a given fraction of mutations by selection will reduce the substitution rate by that fraction. Thus, the proportion of accepted over total mutations at each position is evaluated to be the rate of mutation relative to the neutral rate.

We use sum-of-squares as a distance metric to compare each affinity-threshold rate distribution to the empirical distribution, and we considered the best-fitting affinity threshold to be the affinity threshold that generates the distribution with the smallest distance to the empirical relative rates.

3.2. *The affinity-threshold model well describes binding site substitution rates*

The Halpern-Bruno model[12] has been incorporated into effective tools for motif discovery[13] and identification,[14] and it has been shown to well describe yeast binding site relative rates of substitution.[15] These rates are also generated by our model, and so we judged our model's accuracy by comparing its performance to the Halpern-Bruno model's performance (fig. 1). We aligned ChIP-chip bound regions and computed summed position-specific rates of substitution for the aggregate binding sites of the 111 transcription factors that met our conservation requirements. We were able to find a threshold at which the affinity-threshold model better resembled the empirical data than the Halpern-Bruno model did for 42 of the 49 factors with adequate training data (see Methods). The affinity-threshold model well approximates the position-specific substitution rates of most factors.

The best-fitting score threshold for a transcription factor's binding sites may correspond to their minimum non-deleterious affinity for that transcription factor. If this minimum is variable and can be found through our evolutionary analysis, then we should be able to detect that variability robustly. To this end, we used a bootstrap to assess the reliability of our

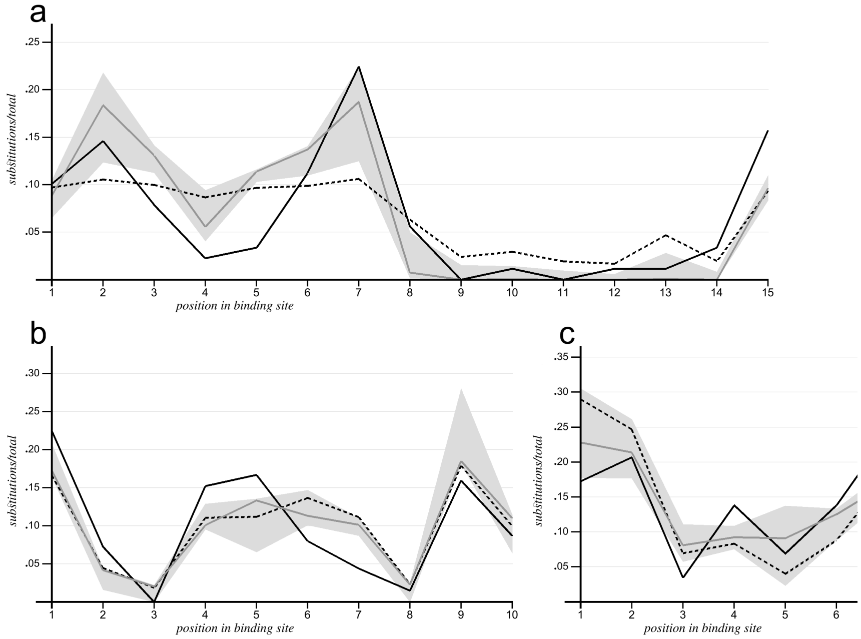

Fig. 1. Position specific rate variation and model predictions for (*a*) Fkh2, (*b*) Fhl1, and (*c*) Aft2: relative rate ($subst/subst_{tot}$) vs position in site. The black line marks the empirical rates, the dashed line marks the Halpern-Bruno predicted rates, and grey line marks the best-fitting affinity-threshold. The grey bar contains the set of rates predicted by all affinity thresholds within the factor's 95% confidence interval

predictions, resampling the the aligned sites. Although most transcription factors had large confidence intervals, they were dispersed over sufficiently wide intervals such that we could form three distinct sets (table 1). We grouped factors with lower bounds greater than 5.9 into a "stringent threshold" set, factors with upper bounds lower than 5.1 into a "lenient threshold" set, and factors with upper bounds lower than 12 and lower bounds greater than -2 into a "medium threshold" set; transcription factors appear to have variable site affinity requirements. We use these sets in all further analysis.

3.3. *The affinity-threshold model predicts extant score distributions for most factors*

If the affinity-threshold model is a reasonable approximation of the evolution of the system, then it should describe other properties of the system beyond the position-specific rate variation of binding sites. One additional prediction of the model is the distribution of binding site scores. For each

Table 1. Affinity threshold confidence intervals and corresponding site prevalence for transcription factors in the stringent (left), medium (middle), and lenient (right) threshold groups

	CI^a	$Prev.^b$		CI^a	$Prev.^b$		CI^a	$Prev.^b$
Reb1p	8.3-11.1	.226-.117	Cin5p	−0.4-8.5	.997-.294	Sut1p	−9.9-4.2	.988-.845
Bas1p	5.8-13.6	.566-.005	Mbp1p	2.7-11.7	.793-.059	Aft2p	−9.8-4.2	.988-.794
Fkh2p	8.1-15.2	.497-.003	Fhl1p	4.2-11.3	.702-.048	Phd1p	−9.8-5.1	.998-.867
Cbf1p	6.2-12.0	.219-.028	Gcn4p	4.0-10.6	.682-.080	Ace2p	−9.9-−0.8	.999-.999
Abf1p	11.0-12.9	.108-.075	Swi6p	3.8-9.9	.854-.166	Yap6p	−9.9-4.2	.993-.909
Sum1p	6.2-14.5	.484-.009	Ste12p	1.0-6.5	.997-.705	Adr1p	−9.5-2.3	.991-.856
Tye7p	8.6-11.3	.183-.037	Nrg1p	−1.3-7.0	.968-.388	Hap5p	−9.4-−2.1	.993-.993
Mcm1p	8.7-19.5	.133-.002				Mot3p	−2.9-5.1	.996-.595
Hap4p	11.0-14.9	.059-.003						

Note: [a] 95% confidence interval, log base two scores
[b] Prevalence: first and second quantities are the fraction of all promoters containing a site meeting the lower and upper bounds of the CI, respectively

factor in the groups determined above we sampled the Markov chain and computed the mean binding site score under the affinity-threshold model. We compared this to the average maximum score for that transcription factor in ChIP-chip bound regions (fig. 2). Although it had a downward bias, the affinity-threshold model predicted the extant distribution of stringent- and medium- threshold transcription factor binding sites. However, it fared worse with the lenient-threshold binding sites, suggesting that the evolution of these sites may not operate within the simplifying bounds of the model, i.e. perhaps their evolution is governed by a more complex fitness landscape instead of our stepwise plateau. Nevertheless, average maximum scores in bound regions for these factors are still found commonly in the genome.

3.4. *Stringent- and lenient-threshold binding sites have distinct patterns of local evolution*

The lenient set of transcription factors allows for binding sites that would be found often by chance in the genome. If this lenient affinity is truly sufficient, these transcription factors may rely on other bound proteins to separate desired from undesired binding sites. In contrast, sites meeting the affinity threshold for stringent-threshold transcription factors should be high-occupancy sites without a need for additional information due to their strong predicted affinity.

To investigate this hypothesis, we counted the average number of different transcription factors bound at each promoter for each of the factors used in the Harbison et al ChIP-chip experiments. Let “lenient-group sites”

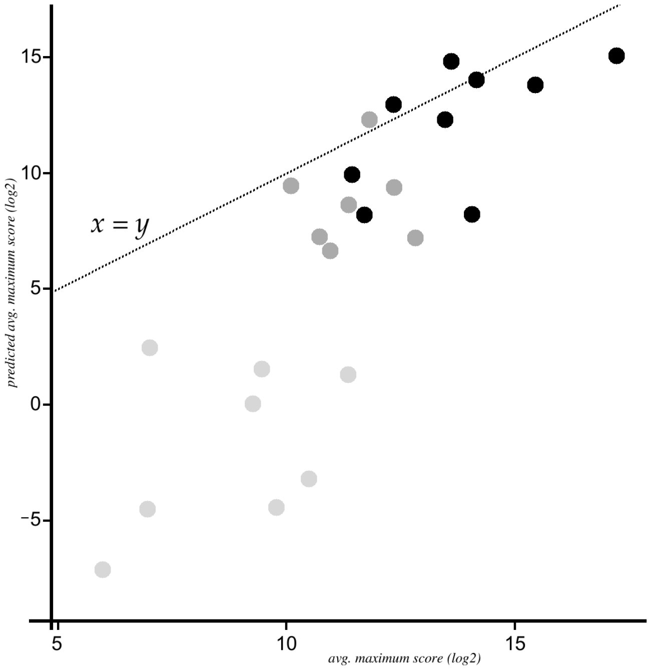

Fig. 2. Predicted average score at best-fitting affinity threshold vs. average maximum score in ChIP-chip bound regions (log base two scores). Stringent-, medium-, and lenient-threshold transcription factors presented as black, dark grey, and light grey dots, respectively.

refer to sites bound by lenient-threshold transcription factors (e.g. Sut1p, table 1), and let "medium-group" and "stringent group" sites be defined similarly. As expected, the stringent and lenient groups were separated, the lenient group promoters having just under three more unique bound factors per promoter for each of three binding significance cutoffs. However, the medium and lenient groups were not well separated.

We used the variation in local substitution patterns to determine whether medium and lenient group factors could be distinguished by an enrichment of local binding events. While medium and lenient group sites have similar numbers of different transcription factors bound to promoters that they also bind, lenient group sites will have a higher density of other binding sites immediately surrounding theirs if recruitment by other proteins is necessary for their function. This density should be reflected in the local pattern of evolution, as the sequence will be comparatively restrained.

We calculated rates of substitution surrounding the binding sites of

Table 2. Average number of binding sites per promoter, grouped by best-fit affinity threshold and ChIP-chip binding p-value

Group	$p < X$		
	.005	.001	.0001
Stringent	7.78	4.74	3.33
Medium	10.30	7.09	5.13
Lenient	10.73	7.59	6.25

stringent-, medium-, and lenient-threshold transcription factors. All transcription factors in each set were pooled and the rate of substitution was calculated and summed by distance to the transcription factor edge. All three sets have a reduced rate of substitution at the position adjacent to the binding site (fig. 3a), suggesting that some of these weight matrices do not describe the entire factor. Lenient group sites have a depressed rate of substitution relative to the areas surrounding the medium and stringent group sites (fig. 3b, $p \cong 0, \chi^2 = 160.8$, 1df), consistent with a hypothesis of increased local binding. In contrast, the regions surrounding stringent group sites are marked by a shoulder of increased substitution rate (fig. 3a). This shoulder suggests a model in which high-affinity sites sterically inhibit transcription factors from binding to adjacent regions, preventing them from being used as regulatory material. The stringent and lenient group sites are distinguished by their expected patterns of local substitution rate variation.

Transcription factors may best interact if they are on the same side of the DNA,[16–18] suggesting that binding sites of interacting factors should be phased at approximately 10.4 base pairs to match the periodicity of the double helix, although this will vary according to the particular nature of interaction between the two proteins. If binding sites coordinated in this manner, the substitution rate should match this periodicity. We evaluated the fit of a model that allowed for a 10.4 base pair periodicity in the rate, although the noted variability between interacting factors will reduce the quality of this match. We fit the twenty base region ten bases from the edge of the transcription factor, allowing for two turns of the DNA while avoiding possible occluding effects of the original bound factor. The regions local to lenient group sites fit this model significantly better than they fit a uniform rate model (fig. 3c, $p = .0053, \chi^2 = 10.53$, 2df), while the regions surrounding medium and stringent group sites did not.

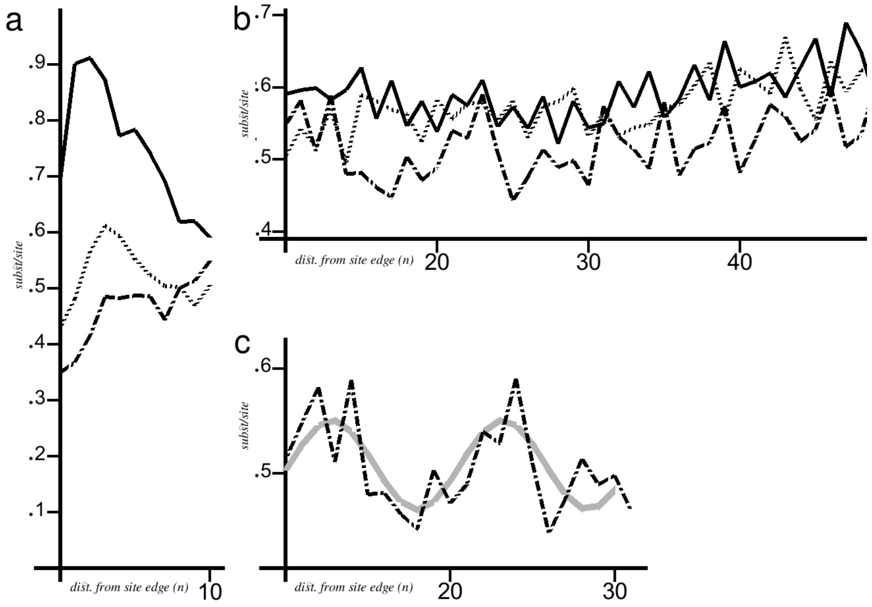

Fig. 3. Local rate of substitution ($subst/site$) vs distance to binding site edge (bp). The solid, dotted, and dot-dashed lines mark the local rates surrounding stringent-, medium-, and lenient-affinity group transcription factor binding sites. In (c), the grey line marks the predicted periodic rate of evolution near lenient-affinity group sites

4. Conclusion

We developed a simple model of binding site evolution to investigate the possibility of differences in transcriptions factors' requirements for binding site affinity. Unlike other models of binding site evolution, the affinity-threshold model is geared toward understanding the transcription factor itself rather than its binding sites. The model was used to create three groups of transcription factors with stringent, lenient, and intermediate requirements for binding site affinity, and these groups were supported by the extant distribution of binding sites and their distinctive patterns of localized substitution rate. We note that some factors appear to evolve and exist at thresholds that poorly distinguish their binding sites from background sequence, perhaps making consideration of context essential for their accurate identification.

5. Methods

5.1. *Rate of binding site evolution*

We downloaded the *S. cerevisiae* sequences used in the Harbison et al[19] study and used bi-directional best FASTA[20] hits ($p < 1e^{-5}$) to find the orthologous subsequences in *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* contigs available at SGD.[21] We aligned the sequences using Mlagan.[22]

We obtained ChIP-chip binding data from Harbison et al, using all available conditions for each factor. We used a binding p-value cutoff of .001 to determine binding, but the analysis was fairly robust to using different cutoffs: we also calculated rates of evolution for of transcription factor binding sites for binding p-values of .005 and .0001 and observed similar groups, although some stringent-threshold factors were lowered to the medium-threshold group using the former data set. We downloaded weight matrices for 124 factors,[9] and we used Patser[23] to designate the highest-scoring subsequence(s) within each bound locus to be the subsequence responsible for binding. This choice precludes the inclusion of many functional weak sites, but we wished to minimize the impact of non-functional sites. Alignment errors, binding site turnover, and changes in cis-regulation all will introduce neutral sequence evolution into the model training data, biasing our choice of threshold downward. In particular, Borneman et al[24] highlighted rapid changes in binding for two transcription factors across three yeast species. We hoped to minimize the impact of such by imposing minimal criteria for conservation: we discarded alignments with gaps and alignments containing a sequence with a score beneath zero. We used maximum parsimony for all determinations of substitution rate. Although progress has been made towards determining the neutral mutation processes in *S. cerevisiae* intergenic sequence,[25] we wished to avoid remaining uncertainties and so in all cases we compared relative rates within the binding site instead of absolute rates. We did not further analyze transcription factors for which we were unable to train on at least two mutations per position. We calculated the Halpern-Bruno rates according to the method described in Moses et al.[15]

5.2. *Simulation of the affinity-threshold model*

We simulated the affinity-threshold model for a wide range of thresholds for each of the 124 weight matrices described by MacIsaac et al. We calculated position-specific substitution rates for score thresholds between -10 and the position weight matrix's maximum in increments of 0.1. This process starts with the consensus sequence and is run for eighteen million iterations. We determined 95% bootstrap confidence intervals of the best-fitting threshold by finding the best-fitting affinity threshold for each of 10,000 resamples of the aligned binding sites. Software will be available from http://rana.lbl.gov/∼rlusk/PSB2008/.

5.3. *Predicted equilibrium distribution of scores*

We sampled every 20,000th sequence generated by the Markov chain for the best-fitting affinity threshold model for each transcription factor in the three groups. We compared the mean score of these sequences with the mean maximum score of the sequences meeting a $p < .001$ ChIP-chip binding cutoff.

5.4. *Periodicity testing*

We evaluated two nested models against the $\pm 10 - 30$ base pair region surrounding each binding site. The first supposed a uniform rate α across the region to determine k_p Poisson-distributed mutation events at each position p, and the second added a periodicity of 10.4 to this rate with magnitude β and phase γ. t_p is the number of gapless alignment columns at that position. The maximum likelihood parameters were discovered by direct search.

$$L(k \mid \alpha, \beta, \gamma; t) = \prod_{p=10}^{30} \frac{e^{-f(\alpha,\beta,\gamma)t_p}\Big[f(\alpha,\beta,\gamma)t_p\Big]^{k_p}}{k_p!}$$

$$f(\alpha,\beta,\gamma) = \Big[1 + \beta \sin(2\pi \frac{p-\gamma}{10.4})\Big]\alpha$$

Significance was determined using a likelihood ratio test with β either allowed to fluctuate between zero and one or held to zero.

This work was supported by a National Institutes Grant R01-HG002779 to MBE. RWL was supported by a NSF graduate research fellowship. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, and the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, State, and Community Programs, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. M. Levine and R. Tjian, *Nature* **424**, 147(July 2003).

2. T. I. Lee and R. A. Young, *Annu Rev Genet* **34**, 77 (2000).
3. M. L. Bulyk, *Genome Biol* **5** (2003).
4. E. Sharon and E. Segal, A feature-based approach to modeling protein-dna interactions, in *RECOMB 2007*, eds. T. Speed and H. Huang (Springer-Verlag, Berlin Heidelberg).
5. G. D. Stormo, *Bioinformatics* **16**, 16(January 2000).
6. O. G. Berg and P. H. von Hippel, *J Mol Biol* **193**, 723(February 1987).
7. J. M. Heumann, A. S. Lapedes and G. D. Stormo, *Proc Int Conf Intell Syst Mol Biol* **2**, 188 (1994).
8. E. Segal, Y. Barash, I. Simon, N. Friedman and D. Koller, From promoter sequence to expression: a probabilistic framework, in *RECOMB 2002*, eds. S. Istrail, M. S. Waterman and A. G. Clark
9. K. D. Macisaac, T. Wang, B. D. Gordon, D. K. Gifford, G. D. Stormo and E. Fraenkel, *BMC Bioinformatics* **7**(March 2006).
10. A. Tanay, *Genome Res* (June 2006).
11. M. Hasegawa, H. Kishino and T. Yano, *J Mol Evol* **22**, 160 (1985).
12. A. L. Halpern and W. J. Bruno, *Mol Biol Evol* **15**, 910(July 1998).
13. A. M. Moses, D. Y. Chiang, D. A. Pollard, V. N. Iyer and M. B. Eisen, *Genome Biol* **5** (2004).
14. A. M. Moses, D. Y. Chiang and M. B. Eisen, *Pac Symp Biocomput* , 324 (2004).
15. A. M. Moses, D. Y. Chiang, M. Kellis, E. S. Lander and M. B. Eisen, *BMC Evol Biol* **3**(August 2003).
16. J. Boros, F. L. Lim, Z. Darieva, A. Pic-Taylor, R. Harman, B. A. Morgan and A. D. Sharrocks, *Nucleic Acids Res* **31**, 2279(May 2003).
17. C. Mao, N. G. Carlson and J. W. Little, *J Mol Biol* **235**, 532(January 1994).
18. I. Ioshikhes, E. N. Trifonov and M. Q. Zhang, *Proc Natl Acad Sci U S A* **96**, 2891(March 1999).
19. C. T. Harbison, B. D. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, A. P. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel and R. A. Young, *Nature* **431**, 99 (2004).
20. D. J. Lipman and W. R. Pearson, *Science* **227**, 1435(March 1985).
21. J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng and D. Botstein, *Nucleic Acids Res* **26**, 73(January 1998).
22. M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow and S. a. Batzoglou, *Genome Res* **13**, 721(April 2003).
23. G. Hertz and G. Stormo, *Bioinformatics* **15**, 563(July 1999).
24. A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein and M. Snyder, *Science* **317**, 815(August 2007).
25. C. S. Chin, J. H. Chuang and H. Li, *Genome Res* **15**, 205(February 2005).