Supplement for "A Biophysical Approach to Transcription Factor Binding Site Discovery".

Marko Djordjevic[1], Anirvan M. Sengupta[2], and Boris I. Shraiman[2]
[1] Department of Physics,
Columbia University, New York, NY 10025.
[2] Department of Physics and BioMaPS Institute,
Rutgers University, Piscataway, NJ 08854

July 30, 2003

Appendix A: Distribution of Interaction Energies.

To evaluate the statistical significance of finding a candidate binding site in a genome, it is necessary to have an estimate of the number of sequences with binding energy below a certain threshold in an ensemble of random sequences with the statistical properties similar to those of the genome in question. A closely related useful concept is the "density of states" or "energy distribution" defined as the expected number of sequences with interaction energy $E(S) = \epsilon \cdot S$ falling in the interval E, $E + dE$. This energy distribution $\rho(E)$ is formally defined by

$$\rho(E) =< \delta(E - \epsilon \cdot S) >_s \tag{1}$$

the average of a Dirac $\delta-$ function over the ensemble of S sequences.

To calculate the density of binding energy $\rho(E)$ for a probability distribution of sequences, $P(S)$, let us introduce a Laplace transform

$$\rho(E) = \frac{1}{2\pi i} \sum_S P(S) \int_{-i\infty}^{+i\infty} d\beta e^{\beta E} e^{-\beta S\cdot\epsilon} \approx \min_{\beta \text{ real}} e^{\beta E + \ln Z(\beta)} \tag{2}$$

where we have defined a "partition function"

$$Z(\beta) = \sum_S P(S) e^{-\beta S\cdot\epsilon} \tag{3}$$

and the last approximate equality, accurate up to logs, is the leading term of the steepest descent approximation to the β integral.

We will calculate energy distributions in two sequence ensembles: 1) one-point statistics, which reproduces correct single site probabilities of different nucleotides, but assumes nucleotides at different positions are uncorrelated, and 2) two-point statistics, which reproduces correct joint probability for nearest neighbor nucleotides

For the one-point model the probability of generating the sequence S factorizes

$$P(S) = \prod_{i=1}^{L} \prod_{\alpha} p_{\alpha}^{S_{\alpha i}}$$

and

$$Z(\beta) = \prod_{i=1}^{L} \sum_{\alpha} p_{\alpha} e^{-\beta \epsilon_{\alpha i}} \tag{4}$$

The density of states can be readily calculated in the "thermodynamic limit" $L \to \infty$. For E near its mean, $\rho(E)$ is well approximated by a Gaussian

$$\rho(E) \approx \exp(-(E-\bar{E})^2/2\chi^2)/\sqrt{2\pi\chi^2} \tag{5}$$

with

$$\bar{E} = -\frac{\partial}{\partial\beta} \ln Z(\beta)|_{\beta=0} = \sum_{i=1}^{L} \bar{\epsilon}_i \tag{6}$$

and

$$\chi^2 = \frac{\partial^2}{\partial\beta^2} \ln Z(\beta)|_{\beta=0} = \sum_{i=1}^{L} \sum_{\alpha} p_{\alpha} (\epsilon_{\alpha i} - \bar{\epsilon}_i)^2 \tag{7}$$

where $\bar{\epsilon}_i = \sum\limits_{\alpha} p_{\alpha} \epsilon_{\alpha i}$ is the mean energy of at position i. We note that χ^2 provides a measure of sequence specificity of the factor in question (addition of a position i with small $\sum\limits_{\alpha} p_{\alpha} (\epsilon_{\alpha,i} - \bar{\epsilon}_i)^2$ does not contribute much to χ !). Away from the 'center' the distribution deviations from Gaussianity appear. In fact the support of $\rho(E)$ is finite with the 'bottom of the band' $E_{bb} = \sum\limits_{i} \min_{\alpha} \epsilon_{i\alpha}$ (and the "top" $E_{tb} = \sum\limits_{i} \max_{\alpha} \epsilon_{i\alpha}$).

The estimate of $\rho(E)$ may be improved beyond the thermodynamical limit by including a finite L correction (i.e. approximation correct to the next order in L^{-1}):

$$\log(\rho_{\epsilon}(E)) \approx \log Z(\beta^*) + \beta^* E - \frac{1}{2} \log(2\pi \frac{\partial^2 \log Z}{\partial^2 \beta^*}) \tag{8}$$

with

$$\frac{\partial \log Z(\beta^*)}{\partial \beta^*} = -E \tag{9}$$

These equations are solved numerically, that is, for the given E we first solve second equation to get β^* and then substitute it in the first equation to get $\rho_\epsilon(E)$.

It turns out that the independent nucleotide approximation is not sufficiently accurate for estimation of genomic 'background' (see Fig. 3) and it is necessary to use the two-point statistics model. Hence, for the Model Genomic Background (MGB), we take the sequence ensemble with dinucleotide correlations fixed by the frequencies of the nearest neighbor nucleotide pairs observed in the non-ORF fraction of *E. coli* genome.

$$p^{(2)}_{\alpha\beta} = \begin{pmatrix} 0.0973 & 0.0818 & 0.0523 & 0.0530 \\ 0.0679 & 0.0973 & 0.0560 & 0.0631 \\ 0.0631 & 0.0530 & 0.0490 & 0.0505 \\ 0.0560 & 0.0523 & 0.0584 & 0.0490 \end{pmatrix} \tag{10}$$

It is convenient to define the conditional probability: $p(\alpha|\beta) = p^{(2)}_{\alpha\beta}/p_\beta$ in terms of which:

$$P(S = |\alpha_1, ..., \alpha_L\rangle) = p_{\alpha_1} p(\alpha_1|\alpha_2)...p(\alpha_{L-1}|\alpha_L)$$

so that the "partition function" (Eq. 3) is given by

$$Z(\beta) = \sum_{\alpha_1=1}^{4} ... \sum_{\alpha_L=1}^{4} p_{\alpha_1} e^{-\beta\epsilon_{1\alpha_1}} p(\alpha_1|\alpha_2)...e^{-\beta\epsilon_{L-1\alpha_{L-1}}} p(\alpha_{L-1}|\alpha_L) e^{-\beta\epsilon_{L\alpha_L}} \tag{11}$$

which is readily evaluated numerically as a product of matrices.

The density of states can be computed from $Z(\beta)$ using the same approximations as before: i.e. via Eqs 8, 9.

Appendix B: Maximum Likelihood Determination of Energy Matrix and Chemical Potential.

The energy matrix ϵ, chemical potential μ and the sampling fraction parameter γ are determined by maximizing the likelihood function $\mathcal{L}$:

$$\frac{\delta}{\delta\epsilon^\alpha_i}\mathcal{L} = -\beta \sum_{S\in O} [1 - f(E(S) - \mu)] S_{\alpha,i} - \gamma \int dE f(E - \mu) \frac{\partial}{\partial\epsilon^\alpha_i} \rho_\epsilon(E) = 0 \tag{12}$$

$$\frac{\delta}{\delta\mu}\mathcal{L} = \beta \sum_{S\in O}[1 - f(E(S) - \mu)] - \gamma\beta \int dE\rho_\epsilon(E) f(E-\mu)[1 - f(E-\mu)] = 0 \quad (13)$$

$$\frac{\delta}{\delta\gamma}\mathcal{L} = \frac{n_s}{\gamma} - \int dE\rho_\epsilon(E) f(E-\mu) = 0 \quad (14)$$

where $\beta = 1/k_B T$. In order to make expression for ϵ more explicit we can resort to the Gaussian approximation for ρ_ϵ which is good for low specificity binding and is described in Appendix A:

$$\int dE f(E-\mu)\rho_\epsilon(E) \approx \int \frac{dE}{\sqrt{2\pi}\chi} f(E-\mu)\exp(-E^2/2\chi^2) \quad (15)$$

where we have for simplicity set the average binding energy ($\bar{\epsilon}_i$) to zero. Eliminating γ using Eq. 14 we arrive at the equations which implicitly determine ϵ, μ :

$$\epsilon_{\alpha,i}/\chi = \frac{p_\alpha^{-1}}{n_s}\sum_{S\in O}[1-f(E(S)-\mu)](S_{\alpha,i}-p_\alpha)]\times \frac{\int dE\rho_\epsilon(E) f(E-\mu)}{\int dE\rho_\epsilon(E)\,(E/\chi)\, f(E-\mu)[1-f(E-\mu)]} \quad (16)$$

$$\frac{1}{n_s}\sum_{S\in O}[1 - f(E(S)-\mu)] = \frac{\int dE\rho_\epsilon(E) f(E-\mu)[1-f(E-\mu)]}{\int dE\rho_\epsilon(E) f(E-\mu)} \quad (17)$$

Unconstrained optimization with respect to ϵ and μ is appropriate when we have extensive data on binding sequences obtained under controlled conditions. The likelihood function, and, therefore, the results of the optimization, are sensitive to the distribution of energies of the example binding sites. When the method is applied to the collection of binding sequences derived from the literature, the results could get affected by non-representative sampling. In that case, it is better to fix the magnitude of the energy matrix relative to $k_B T$ at some reasonable value and restrict optimization to the transverse components (in effect, if one thinks of ϵ as a vector, we fix its magnitude but vary the direction). We expect on the physical grounds that $\beta|\epsilon|$ is considerably larger than one. The results of optimization with respect to transverse components of ϵ are insensitive to the assumed value of $\beta\ |\epsilon|$ provided it is large. This justifies our use of zero temperature, or equivalently $\beta|\epsilon| \to \infty$ limit. On the other hand, we have not been able to estimate, on the basis of database derived binding sequences, the magnitude of binding energy. In order to do this, more systematic data are required.