Protein-DNA interaction energetics

Exact (*ab-initio*) calculation of a protein-DNA complex energy is generally a very difficult problem. The energy constituents are, to name a few: the direct electrostatic interaction between charged elements of the protein and the DNA (e.g. phosphate backbone), hydrogen bonds between binding domain aminoacids and DNA bases, effective hydrophobic interactions, water-mediated interactions etc. Though much effort is invested presently in this direction, coherent picture is still missing.

However, a heuristic approach to this seemingly intractable problem originating in the seminal papers by von Hippel and Berg proved to be very successful. The complete theory is described in details elsewhere [1, 2]; here, we only provide the necessary background.

Suppose there are n_s specific sites of length L for a given regulatory protein. In thermal equilibrium with protein solution, the probability of a certain site i to be occupied (or a site binding constant) is proportional to a Boltzmann factor $e^{-\beta E_i}$. Then, by measuring site affinities, it is possible to determine site binding energy. Futhermore, if we assume that each base contributes independently to the binding energy, it is possible to measure individual contributions of the bases by mutating the binding sequence. [7]

The argument of Berg and von Hippel is based on the analogy they draw between thermodynamic picture and evolutionary selection process. This analogy appears reasonable if we assume that during the evolution process only sequences with binding energies in a certain interval $E_s \pm \Delta/2$ are selected. Suppose that the binding domain of the regulatory protein is conserved throughout the evolution process and that there exists some strongest (*consensus*) binding sequence. Then every base-pair mismatch in the sequence will weaken the binding by a certain *discrimination energy*, the value of which depends both on the position and the mutated base-pair. If all positions are equally important and any mutation contributes the same discrimination energy, then specifiyng the required (for selection) sequence energy is equivalent to specifying the number of base-pair mismatches. For a sequence of length L with m mismatches, each contributing the discrimination energy of ϵ,

$$E(m, L) = E_c + m\epsilon, \tag{1}$$

$$\Omega(m, L) = \frac{L!}{m!(L-m)!} \cdot 3^m. \tag{2}$$

The total number of sequences is 4^L. In a genome of a size N, any sequence should be at least

$$L_{min} = \log_4(2N) \tag{3}$$

long so that it won't appear randomly. For *E. coli*, $N = 4.64 \times 10^6$ base pairs, so that $L_{min} \simeq 13$.

Fig. 1shows the results of numerical simulation for 20-bp sequence binding energy for a random "genome" of size 10^7. The logarithm of the density of states $\Omega(E)$ can be quite adequately fitted by a parabola, which is merely a consequence of the Central Limit Theorem (CLT) applied to a sum of 20 random variables. Thus, the genome binding energy spectrum can be described by the Random Energy Model (REM)[3, 4], so that we can define the *evolutionary temperature* T^* as

$$T^* \equiv \lambda^{-1} = \left[\frac{d}{dE}\ln\Omega(E)\right]^{-1}_{E=E_s} = \frac{\Sigma^2}{|E_s - \langle E\rangle|}, \tag{4}$$

where Σ^2 is the variance and $\langle E\rangle$ is the average binding energy. This equation defines the

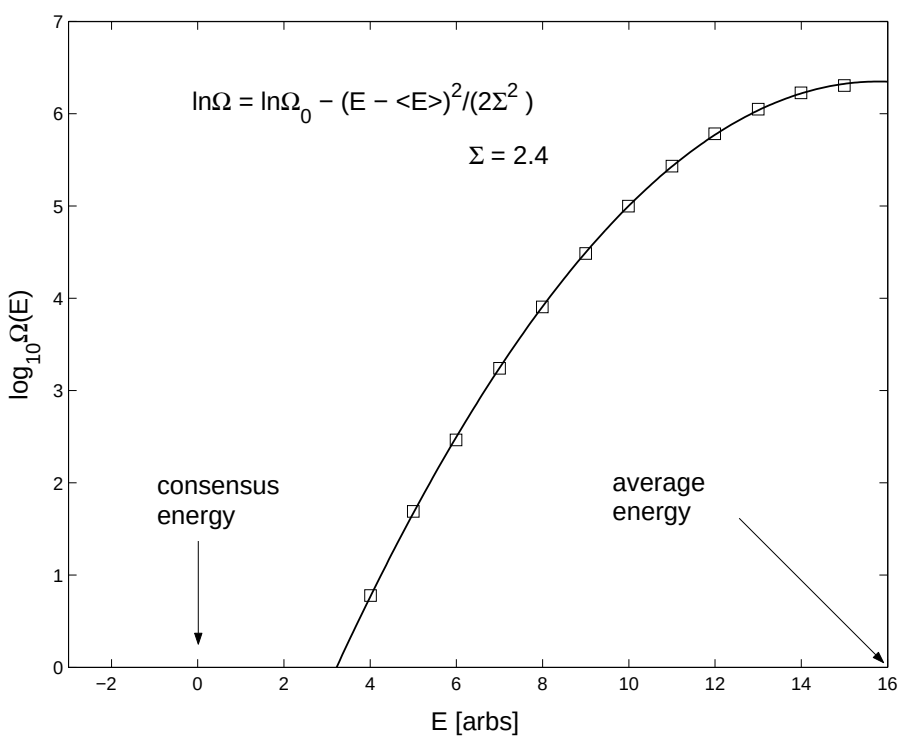

FIG. 1: *Energy spectrum of 20 bp sequence with unit discrimination energy. The squares are the results of computer simulation; the solid line is a quadratic fit.*

transition to the canonical description, which is more appropriate in the general case, when different positions and mutations contribute nonequally. Then, if the entire genome is at "thermal equilibrium" at temperature T^*, the partition function for a set of all possible sequences of length L is

$$Z^* = \prod_{i=1}^{L} \sum_{\alpha=1}^{4} e^{-\lambda \epsilon_{i,\alpha}}, \tag{5}$$

where α counts the possible mutations and $\epsilon_{i,\alpha}$ is the corresponding discrimination energy. Under these conditions, the probability of α-th base to be observed in the selected sequence at the i-th position is

$$p_\alpha(i) = \frac{e^{-\lambda \epsilon_{i,\alpha}}}{\sum_{\gamma=1}^{4} e^{-\lambda \epsilon_{i,\gamma}}}. \tag{6}$$

Thus, if a collection of binding sites for a certain protein is known, it is possible to estimate the binding energies (up to a certain constant factor[8]) by observing the base frequencies at various positions in the sites and taking a logarithm, thus constructing the *weight matrix* [5]

$$-\lambda \epsilon_{i,\alpha} = \ln \left(\frac{p_i(\alpha)}{p_0(\alpha)} \right), \tag{7}$$

where $p_0(\alpha)$ is the frequency of the occurence of the base α inside the genome. The weight matrix is a characteristic of the binding domain of the protein; applying it to any arbitrary DNA sequence produces this sequence binding energy (see Fig. 2). The average binding energy for a set of sites is

$$-\lambda \langle E \rangle = \sum_{i,\alpha} p_i(\alpha) \ln \left(\frac{p_i(\alpha)}{p_0(\alpha)} \right) = I, \tag{8}$$

where I is the information content of the set of sites.

The entire specific binding spectrum is never observed. About $15 - 18\ k_B T$ above the consensus site, which corresponds to 4-5 significant point mutations, a protein switches its binding mode and binds to the DNA *non-specifically*. Non-specific binding has a relatively weak dependence on the underlying sequence, hence its name. It turns out that it is very important in the *kinetics* of the target site location.

---

[1] O. G. Berg and P. H. von Hippel, J. Mol. Biol. **193**, 723 (1987).

[2] P. H. von Hippel and O. G. Berg, Proc. Natl. Acad. Sci. USA **83**, 1608 (1986).

[3] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. USA **99**, 12015 (2002).

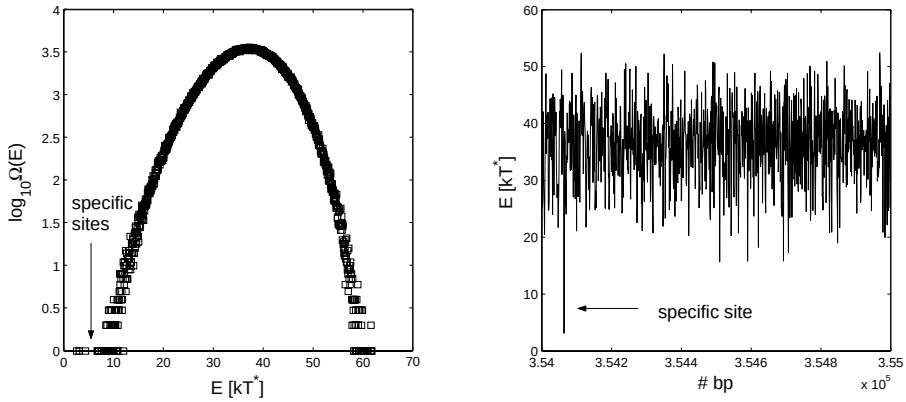

FIG. 2: *Energy spectrum and energy profile for E. coli purine repressor (purR). The weight matrix was built by analyzing 35 known binding sites for purR.*

[4] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[5] G. D. Stormo and D. S. Fields, Trends in Biochem. Sci. **23**, 109 (1998).

[6] Y. Takeda, A. Sarai, and V. M. Rivera, Proc. Natl. Acad. Sci. USA **86**, 439 (1989).

[7] The independence conjecture has been verified experimentally for a very wide class of transcription factors. Here, we ignore correlations between bases inside the binding sequences.

[8] Interestingly enough, most experiments[5, 6] suggest values of $T^*/T \sim 1$.