Energy Fluctuations Shape Free Energy of Nonspecific Biomolecular Interactions

Michael Elkin · Ingemar Andre · David B. Lukatsky

Abstract Understanding design principles of biomolecular recognition is a key question of molecular biology. Yet the enormous complexity and diversity of biological molecules hamper the efforts to gain a predictive ability for the free energy of protein-protein, protein-DNA, and protein-RNA binding. Here, using a variant of the Derrida model, we predict that for a large class of biomolecular interactions, it is possible to accurately estimate the relative free energy of binding based on the fluctuation properties of their energy spectra, even if a finite number of the energy levels is known. We show that the free energy of the system possessing a wider binding energy spectrum is almost surely lower compared with the system possessing a narrower energy spectrum. Our predictions imply that low-affinity binding scores, usually wasted in protein-protein and protein-DNA docking algorithms, can be efficiently utilized to compute the free energy. Using the results of Rosetta docking simulations of protein-protein interactions from Andre et al. (Proc. Natl. Acad. Sci. USA 105:16148, 2008), we demonstrate the power of our predictions.

Keywords Fluctuations · Free energy of biomolecular interactions

1 Introduction

Recent high-throughput experiments demonstrate a high level of multi-specific and non-specific binding in protein-protein [1], protein-DNA [2], and protein-RNA [3] interactions in a living cell. These observations challenge a conventional approach of molecular biology

M. Elkin
Department of Computer Science, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel

I. Andre
Department of Biochemistry and Structural Biology, Lund University, 221 00 Lund, Sweden

D.B. Lukatsky (✉)
Department of Chemistry, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel
e-mail: lukatsky@bgu.ac.il

usually focusing on just a single pathway or function, or a single binding partner for a protein. This suggests that in order to predict correctly the properties of molecular interaction networks, one needs to take into account the effect of multiple binding, essentially computing the free energy of the system rather than the energy of individual states. The latter statement is quite intuitive as any protein in a cell interacts with thousands of proteins (or DNA binding cites), and even if one (or few) of its interaction partners have stronger binding affinity than others, still weaker interactions are not negligible and they may become even dominant. Yet the complexity of biological molecules and a lack of knowledge of accurate inter-molecular interaction potentials hamper computational efforts to predict the free energies of protein-protein, protein-DNA, and protein-RNA binding. A key question is how to estimate the binding free energy based on the partial knowledge of the binding energy spectrum. Each energy in the binding energy spectrum is defined here as the inter-molecular interaction energy of a particular bound state of interacting molecules (e.g., particular binding configuration of protein-protein or protein-DNA complex).

It was recently shown that global symmetry properties of proteins, both on structural and sequence levels, generically define the properties of their binding energy spectrum [4–9]. In particular, it was shown analytically in [4] that the probability distribution for the interaction energies of homodimers, $P(E)$, is always wider as compared to heterodimers, $\sigma_{\text{homo}}/\sigma_{\text{hetero}} = \sqrt{2}$, where σ is the dispersion of $P(E)$. This statistical law was also confirmed computationally, using one of the most advanced methods for computing protein-protein interactions applied to a large dataset of protein complexes from the protein data bank (PDB) [6]. It was also predicted that proteins possessing a higher level of structural correlations (clustering) of amino acids in their interfaces, demonstrate a wider binding energy spectrum, as well [7]. It was shown recently that protein sequences with enhanced strength of diagonal correlations of amino acid positions demonstrate a similar property [8, 9]. Intuitively it means that the clustering of amino acids of the same type statistically enhances the dispersion of the binding energy spectrum [8, 9]. Sequences with a higher level of such clustering will possess a larger dispersion than sequences with a lower level of clustering [8, 9]. We have recently analyzed the properties of the energy spectrum of non-specific protein-DNA binding [10]. Similar to the case of protein-protein interactions, we also observed that the width of the protein-DNA binding energy spectrum depends on the correlation properties of DNA, such as the symmetry and the length-scale of DNA sequence correlations [10].

We emphasize that in all of those examples the average interaction energies of the compared spectra are always identical, and only the dispersions of the energy spectra are different, Fig. 1. The predicted effects are thus essentially governed by the fluctuations of energy, and go beyond the mean field. The case where the average energies are not equal is also discussed below. We assume here that the probability distribution, $P(E)$, is Gaussian. This is an accurate assertion since, practically, the binding energy, E, is a sum of thousands of binary inter-atomic interactions, and this sum is normally distributed according the central limit theorem [4].

Here, we estimate the relative free energy of two interacting systems characterized by the same average binding energies, $\langle E_1 \rangle = \langle E_2 \rangle$, but different dispersions, $\sigma_1 > \sigma_2$, Fig. 1A. We show that the free energy, F, of the system possessing a wider binding energy spectrum, is always shifted towards lower free energies compared to the system possessing a narrower $P(E)$, even if a finite number of energy levels is known, Fig. 1A. In particular, we show that the partition function, Z_1, is almost surely larger than Z_2, with the probability, $P(Z_1 > Z_2) \geq 1 - C/M$, where $C(\sigma_1, \sigma_2)$ is a finite constant, and M is the number of the energy levels used to compute the partition function.
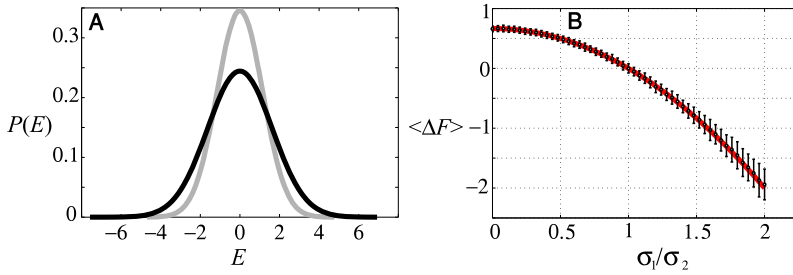
Fig. 1 Calculation of the free energy and fluctuations of the free energy from the energy spectrum. **A.** Example: Gaussian probability distributions for the interaction energy, E, characterized by the identical average energies, $\langle E_1 \rangle = \langle E_2 \rangle$, and different dispersions, $\sigma_1/\sigma_2 = \sqrt{2}$. E is represented in the units of $k_B T$. **B.** Computed average free energy differences, $\langle \Delta F \rangle = \langle F(\sigma_1) - F(\sigma_2) \rangle$, as a function of the ratio of dispersions. *Circles* with error bars represent the simulation results, where the quenched averaging is performed (see the text). We used $M = 1000$ for each computation of the partition function, and the averaging is performed with respect to 200 realizations. $\langle \Delta F \rangle$ is represented in the units of $k_B T$. Error bars represent free energy fluctuations, and show two standard deviations. *Solid curve* represents the analytical result, (7)

We note that in his seminal work [11], Derrida has established that in the random energy model, where the energy spectrum of the system, $P(E)$, is Gaussian, and the partition function, $Z = \sum_{i=1}^{M} \exp(-E^{(i)}/k_B T)$, for each realization of $P(E)$ with M energy states, $E^{(i)}$, the quenched average of the free energy, $\langle F \rangle_q = -k_B T \langle \ln Z \rangle$, is equal to the annealed average, $\langle F \rangle = -k_B T \ln \langle Z \rangle$, in the thermodynamic limit of large M:

$$\langle F \rangle = -k_B T \ln M - \frac{\sigma^2}{2k_B T}, \tag{1}$$

if the temperature T is above some critical temperature, $T > T_c$, where $k_B T_c \sim \sigma/\sqrt{\ln M}$ [11], and σ is the standard deviation of $P(E)$. Using an example from the Rosetta docking simulations of protein-protein interactions, we show below that (1) provides an accurate estimate for the relative free energy, when M reaches only few thousands.

We stress that our model is applicable to interacting systems without a pronounced low-energy (ground) state in their energy spectra. The existence of such a ground state corresponds to a strong, specific binding. On the contrary, a large class of weakly interacting biomolecules in a living cell, such as nonspecific protein-protein, protein-DNA, or protein-RNA binding, represents the systems where our model is operational. Such relatively weak, nonspecific interactions, often called "promiscuous interactions", have been shown to play an important role in different cellular processes, and in many cases, the effect of such weak interactions becomes the dominant factor in a living cell [12].

2 Results

We consider the ensemble of the interaction energies, $\{E^{(i)}\}$, of two interacting biomolecules, where each energy, $E^{(i)}$, corresponds to a given conformation (i.e., a given interaction state), i, of these molecules.

We begin with the definition of the free energy of the system, $F = -\ln Z$, where the partition function, $Z = \sum_{i=1}^{M} e^{-E^{(i)}}$, and we assume for simplicity that $k_B T = 1$, and the energy, E, is represented in the units of $k_B T$; here k_B is the Boltzmann constant, and T is the absolute temperature. We also assume that the energy, E, obeys the Gaussian distribution,

$P(E)$, with zero mean, $\langle E \rangle = 0$, and standard deviation, σ. The set of M energy values, $E^{(i)}$, is obtained as a statistical realization of $P(E)$. In what follows we compare the statistical properties of Z computed based on the realizations drawn from two distributions with different values of standard deviation, $\sigma_1 > \sigma_2$, Fig. 1A.

Since $Y = e^{-E}$ is a lognormal random variable, it is well-known [14] and can be readily verified that the expectation, $\langle Y \rangle$, and the variance, $VAR(Y)$, of Y are given by $\langle Y \rangle = e^{\sigma^2/2}$ and $VAR(Y) = e^{2\sigma^2} - e^{\sigma^2}$. We note that (1) simply follows from $\langle F \rangle = -\ln(M\langle Y \rangle)$.

For a large number M, let $Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)}$ be independent random variables, so that each of them is distributed identically to Y. Since $Z = \sum_{i=1}^{M} Y^{(i)}$, by linearity of the expectation, $\langle Z \rangle = M \cdot e^{\sigma^2/2}$. Also, since $Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)}$ are independent, it follows that

$$VAR(Z) = VAR\left(\sum_{i=1}^{M} Y^{(i)}\right) = \sum_{i=1}^{M} VAR(Y^{(i)}) = M \cdot (e^{2\sigma^2} - e^{\sigma^2}), \tag{2}$$

and the standard deviation of Z, $\sigma(Z) = \sqrt{VAR(Z)}$.

Consider now two normal independent random variables E_1 and E_2, both having zero mean. We assume further that the standard deviation σ_1 of E_1 is greater than the standard deviation σ_2 of E_2, i.e., $\sigma_1 > \sigma_2 > 0$. Let $Y_1 = e^{-E_1}$ and $Y_2 = e^{-E_2}$ be the corresponding lognormal random variables. Next we show that asymptotically almost surely it holds that $Z_1 > Z_2$, where $Z_1 = \sum_{i=1}^{M} Y_1^{(i)}$ and $Z_2 = \sum_{i=1}^{M} Y_2^{(i)}$.

Lemma 1 *Let* $\sigma_1, \sigma_2, \sigma_1 > \sigma_2 > 0$, *be two positive constants. Then*

$$P(Z_1 > Z_2) \geq 1 - \frac{1}{M} \cdot f(\sigma_1, \sigma_2), \tag{3}$$

where

$$f(\sigma_1, \sigma_2) = 4 \cdot \frac{(e^{2\sigma_1^2} - e^{\sigma_1^2}) + (e^{2\sigma_2^2} - e^{\sigma_2^2})}{(e^{\sigma_1^2/2} - e^{\sigma_2^2/2})^2}$$

is a positive constant that depends only on σ_1 *and* σ_2.

Proof Chebyshev's inequality (see, e.g., [13], p. 43) states that for any random variable X with expectation $\langle X \rangle$ and standard deviation σ_X, and any $b > 0$,

$$P(|X - \langle X \rangle| \geq b \cdot \sigma_X) \leq \frac{1}{b^2}. \tag{4}$$

By the preceding argument (see (2)), for $i = 1, 2$,

$$\langle Z_i \rangle = M \cdot e^{\sigma_i^2/2}, \qquad \sigma(Z_i) = \sqrt{M} \cdot \sqrt{e^{2\sigma_i^2} - e^{\sigma_i^2}}.$$

Let

$$A = \frac{\langle Z_1 \rangle + \langle Z_2 \rangle}{2} = M \cdot \frac{e^{\sigma_1^2/2} + e^{\sigma_2^2/2}}{2}.$$

Hence

$$\langle Z_1 \rangle - A = A - \langle Z_2 \rangle = \frac{\langle Z_1 \rangle - \langle Z_2 \rangle}{2} = M \cdot \frac{e^{\sigma_1^2/2} - e^{\sigma_2^2/2}}{2}. \tag{5}$$

Denote $Q = \frac{e^{\sigma_1^2/2} - e^{\sigma_2^2/2}}{2}$ and $D = M \cdot Q$. Observe that since $\sigma_1 > \sigma_2$, both D and Q are positive. Consequently, $\langle Z_1 \rangle - D = \langle Z_2 \rangle + D = A$. Also, by Chebyshev's inequality (see (4)),

$$P(|Z_1 - \langle Z_1 \rangle| \geq D) = P\left(|Z_1 - \langle Z_1 \rangle| \geq \sqrt{M} \cdot \frac{Q}{\sqrt{e^{2\sigma_1^2} - e^{\sigma_1^2}}} \cdot \sigma(Z_1)\right)$$

$$\leq \frac{1}{M} \cdot \frac{e^{2\sigma_1^2} - e^{\sigma_1^2}}{Q^2}.$$

It follows that

$$P(Z_1 \leq A) = P(Z_1 \leq \langle Z_1 \rangle - D) = P((\langle Z_1 \rangle - Z_1) \geq D)$$

$$\leq P(|\langle Z_1 \rangle - Z_1| \geq D) \leq \frac{1}{M} \cdot \frac{e^{2\sigma_1^2} - e^{\sigma_1^2}}{Q^2}.$$

Analogously,

$$P(Z_2 \geq A) = P((Z_2 - \langle Z_2 \rangle) \geq D) \leq P(|Z_2 - \langle Z_2 \rangle| \geq D)$$

$$\leq \frac{1}{M} \cdot \frac{e^{2\sigma_2^2} - e^{\sigma_2^2}}{Q^2}.$$

Hence, by union bound,

$$P((Z_1 > A) \text{ and } (Z_2 < A)) \geq 1 - \frac{1}{M} \cdot \frac{1}{Q^2} \cdot ((e^{2\sigma_1^2} - e^{\sigma_1^2}) + (e^{2\sigma_2^2} - e^{\sigma_2^2})).$$

Finally,

$$P(Z_1 > Z_2) \geq P((Z_1 > A) \text{ and } (Z_2 < A))$$

$$\geq 1 - \frac{1}{M} \cdot \frac{1}{Q^2} \cdot ((e^{2\sigma_1^2} - e^{\sigma_1^2}) + (e^{2\sigma_2^2} - e^{\sigma_2^2})). \tag{6}$$

$\square$

Since the right-hand side of the inequality (3) tends to 1 as M grows, it follows that the event $Z_1 > Z_2$ occurs asymptotically almost surely. This argument generalizes directly to the scenario when $Z_1 = \sum_{i=1}^{M_1} Y_1^{(i)}$ and $Z_2 = \sum_{i=1}^{M_2} Y_2^{(i)}$, where M_1 and M_2 are (not necessarily equal) large integers. The generalized inequality is

$$P(Z_1/M_1 > Z_2/M_2)$$

$$\geq 1 - \frac{4}{(e^{\sigma_1^2/2} - e^{\sigma_2^2/2})^2}\left(\frac{1}{M_1} \cdot (e^{2\sigma_1^2} - e^{\sigma_1^2}) + \frac{1}{M_2} \cdot (e^{2\sigma_2^2} - e^{\sigma_2^2})\right).$$

It is easy to understand the obtained results intuitively. In the calculation of the partition function, $Z = \sum_{i=1}^{M} e^{-E^{(i)}}$, M energies $E^{(i)}$ are drawn from the Gaussian distribution. However, only a subset of *lowest* energies provides the dominant contribution to Z. The contribution from high energies is small, $e^{-|E|} \ll 1$. Since this dominant subset is localized in the low energy tail, the distribution with a larger standard deviation will obviously deliver the larger partition function.

The major practical implication of our result is the ability to estimate the relative free energy of biomolecular interactions without performing the actual calculations of the free energy. We establish that a simple, direct relationship between the average free energy difference and the standard deviations of the energy spectra, (1),

$$\langle \Delta F \rangle = \langle F(\sigma_1) - F(\sigma_2) \rangle = -\frac{\sigma_1^2 - \sigma_2^2}{2}, \tag{7}$$
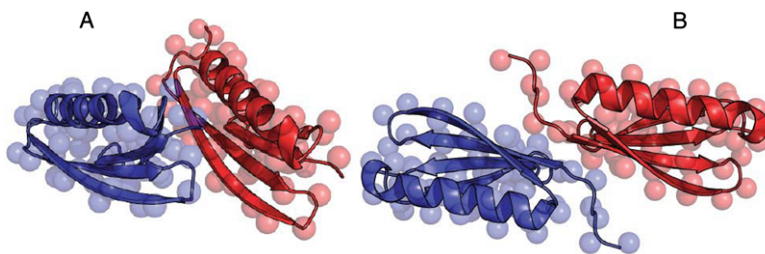
Fig. 2 Snapshot of asymmetric (we use the term "heterodimeric" to describe such symmetry) (**A**), and symmetric (we use the term "homodimeric" to describe such symmetry) (**B**) binding modes from Rosetta docking simulations of protein L (PDB code: 1hz6). The structures represent the lowest energy binding modes (using the same energy term as in Ref. [6], the interchain pair potential in the Rosetta low resolution docking energy function) from an asymmetric or symmetric docking simulation of protein L dimers. Position of centroid atoms, representing the sidechain atoms, is shown as *spheres*

is accurate for a system where only a finite number of energy levels is known. We note that our analytical definition of the average free energy relies on the *annealed* definition of the average, $\langle F \rangle = -\ln\langle Z \rangle$. In systems without frustration the latter definition is known to be in excellent agreement with the *quenched* averaging, $\langle F \rangle_q = -\langle \ln Z \rangle$, unlike the case of highly frustrated systems such as spin glasses [11] or proteins below the glass transition temperature [15]. Indeed, the quenched averaging performed numerically is in excellent agreement with the analytical result, Fig. 1B. The error bars in this plot represent the magnitude of the free energy fluctuations. Yet, our central result in this paper is stronger than the statement described by (7). Here we predict for two systems, that even if a single calculation of the free energy is performed for each system, using a single realization of the probability distributions, $P(E_1)$ and $P(E_2)$, and it is known that $\sigma_1 > \sigma_2$, then we guarantee that $F_1 < F_2$ with the probability approaching one, provided that the number of measured energy levels, M, in each realization is sufficiently large. Finally we note that if one of the distributions, $P(E_1)$, is shifted from zero mean by the energy, $\langle E_1 \rangle = E_0$, the free energy, (7) gets trivially shifted exactly by this magnitude, $\langle \Delta F \rangle = E_0 - (\sigma_1^2 - \sigma_2^2)/2$. This is because the fluctuation contribution to the free energy difference depends exclusively on the widths of the corresponding energy spectra.

3 Example: Free Energy of Nonspecific Protein-Protein Interaction

We now apply our results to the calculation of the free energy of nonspecific protein-protein interactions. We use an example from Andre et al. [6], where the Rosetta docking simulations of self-interacting protein L in homodimeric and heterodimeric conformations were performed [6] (see Fig. 2). In particular, these simulations provide the interaction energies of $\sim 30,000$ homodimeric and $\sim 22,000$ heterodimeric conformations, respectively, Fig. 3A. Each of these conformations is chosen randomly, without any energy optimization. Based on these energies, we computed the free energy difference $\Delta F = F_{\text{homo}} - F_{\text{hetero}}$, as a function of the energy sample size, M, where, $F_{\text{homo}} = \sum_{i=1}^{M} \exp(-E_{\text{homo}}^{(i)})$, and analogously for F_{hetero} (see Fig. 3B). The key result here is that the free energy of homodimeric conformations is always lower then the corresponding free energy of heterodimeric conformations, Fig. 3B. After the sample size, M, reaches only few thousands conformations, the free energy difference reaches its expectation value, $\langle \Delta F \rangle$, with the accuracy reaching 90–95% (see inset in Fig. 3B). The estimate performed above, (6), gives: $P(Z_{\text{homo}} > Z_{\text{hetero}}) \geq 0.95$,
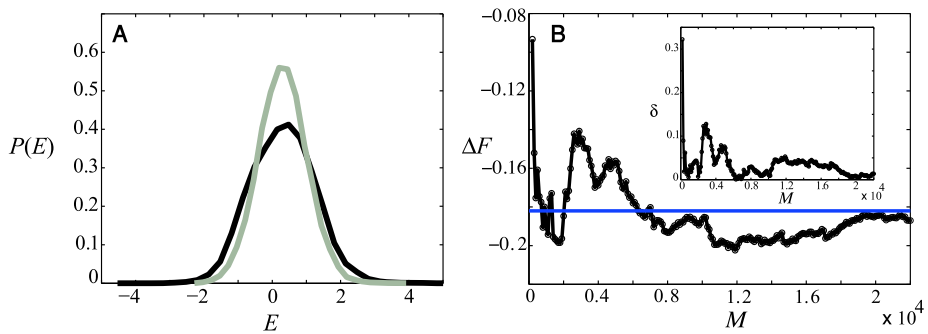
Fig. 3 Example: Calculation of the free energy of nonspecific self-binding for protein L, using the Rosetta docking scores obtained by Andre et al. in Ref. [6]. **A.** Computed probability distributions, $P(E)$, of the Rosetta docking energies for protein L in symmetric (i.e. homodimeric, shown in *black*) and nonsymmetric (i.e. heterodimeric, shown in *gray*) conformations [6]. Snapshots from these Rosetta docking simulations are shown in Fig. 2. The interaction energy, E, is in dimensionless Rosetta score units. There are overall 29,976 homodimeric, $\{E_{\text{homo}}^{(i)}\}$, and 22,038 heterodimeric, $\{E_{\text{hetero}}^{(i)}\}$, conformations sampled, respectively. **B.** Computed free energy difference between homodimers and heterodimers, $\Delta F = F_{\text{homo}} - F_{\text{hetero}}$, as a function of the energy sample size, M, where, $F_{\text{homo}} = \sum_{i=1}^{M} \exp(-E_{\text{homo}}^{(i)})$, and analogously for F_{hetero}. The *horizontal line* represents the expectation value, $\langle \Delta F \rangle = -(\sigma_{\text{homo}}^2 - \sigma_{\text{hetero}}^2)/2$, (7), where σ_{homo} and σ_{hetero} are the standard deviations of the corresponding energy spectra. Inset represents the relative deviation of ΔF from the expectation value, $\delta = |\Delta F - \langle \Delta F \rangle|/|\Delta F + \langle \Delta F \rangle|$, as a function of M

if $M = 5000$, where Z_{homo} and Z_{hetero} are the corresponding partition functions, each obtained based on M energy values. We suggest therefore that our method should provide an efficient way to estimate the free energies of nonspecific biomolecular binding.

4 Conclusion

The majority of macromolecular docking algorithms rejects the lower-affinity binding scores and retain only one or few lowest energy conformations. We suggested here a simple method based on Derrida-type random energy model [11], how those wasted scores can be used in order to estimate the free energy of binding. Our conclusions can be applicable to different biomolecular systems, such as protein-protein, protein-RNA, and protein-DNA complexes. The input energy spectra, $P(E)$, may come from different configurations of two interacting biomolecules, or they can come from a single biomolecule interacting with a set of partner binders. It is important to note that our conclusions hold true even when the probability distribution, $P(E_1)$ with a larger standard deviation than $P(E_2)$, $\sigma_1 > \sigma_2$, is sampled by a smaller number of states, $M_1 < M_2$! The free energy F_1 will be reduced compared with F_2 in the latter case due to the fact that the dominant contribution to the partition function comes from the lower-energy tails of $P(E_1)$ and $P(E_2)$, and thus a wider energy spectrum will always deliver a lower free energy.

In conclusion, we stress that our model is applicable to weakly interacting systems, without a pronounced energy minima in the interaction energy spectrum. We use the term "nonspecific binding" or "promiscuous binding" to describe such systems. Nonspecific binding is widespread in a living cell. In practice, the majority of the interactions are actually nonspecific. Traditionally, such nonspecific interactions are neglected, which leads to significant inaccuracies in the computation of the free energy of the system. Here, we suggested a possible method to estimate the relative free energies of nonspecific binding.

References

1. Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., et al.: High-quality binary protein interaction map of the yeast interactome network. Science **322**, 104–110 (2008). doi:10.1126/science.1158684

2. Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., et al.: *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat. Biotechnol. **28**, 970–975 (2010)

3. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, L., et al.: Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP. Cell **141**, 129–141 (2010). doi:10.1016/j.cell.2010.03.009

4. Lukatsky, D.B., Zeldovich, K.B., Shakhnovich, E.I.: Statistically enhanced self-attraction of random patterns. Phys. Rev. Lett. **97**(17), 178101 (2006). doi:10.1103/PhysRevLett.97.178101

5. Lukatsky, D.B., Shakhnovich, B.E., Mintseris, J., Shakhnovich, E.I.: Structural similarity enhances interaction propensity of proteins. J. Mol. Biol. **365**, 1596–1606 (2007). doi:10.1016/j.jmb.2006.11.020

6. Andre, I., Strauss, C.E.M., Kaplan, D.B., Bradley, P., Baker, D.: Emergence of symmetry in homooligomeric biological assemblies. Proc. Natl. Acad. Sci. USA **105**, 16148–16152 (2008). doi:10.1073/pnas.0807576105

7. Lukatsky, D.B., Shakhnovich, E.I.: Statistically enhanced promiscuity of structurally correlated patterns. Phys. Rev. E **77**, 020901(R) (2008). doi:10.1103/PhysRevE.77.020901

8. Lukatsky, D.B., Afek, A., Shakhnovich, E.I.: Sequence correlations shape protein promiscuity. J. Chem. Phys. **135**(6), 065104 (2011). doi:10.1063/1.3624332

9. Afek, A., Shakhnovich, E.I., Lukatsky, D.B.: Multi-scale sequence correlations increase proteome structural disorder and promiscuity. J. Mol. Biol. **409**(3), 439–449 (2011). doi:10.1016/j.jmb.2011.03.056

10. Sela, I., Lukatsky, D.B.: DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. Biophys. J. **101**(1), 160–166 (2011). doi:10.1016/j.bpj.2011.04.037

11. Derrida, B.: Random-energy model: limit of a family of disordered models. Phys. Rev. Lett. **45**(2), 79–82 (1980). doi:10.1103/PhysRevLett.45.79

12. Khersonsky, O., Tawfik, D.S.: Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. **79**, 471–505 (2010). doi:10.1146/annurev-biochem-030409-143718

13. Alon, N., Spencer, J.: The Probabilistic Method. Wiley-Interscience Series in Discrete Mathematics and Optimization, 3rd edn. (2008)

14. Atchinson, J., Brown, J.A.C.: The Lognormal Distribution, with Special Reference to Its Use in Economics. Cambridge University Press, Cambridge (1957)

15. Shakhnovich, E.I., Gutin, A.M.: Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. Biophys. Chem. **34**(3), 187–199 (1989)