Energy Landscape Theory, Funnels, Specificity, and Optimal Criterion of Biomolecular Binding

Jin Wang[1,2,3,4,*] and Gennady M. Verkhivker[5,†]

[1]*State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130021, People's Republic of China*

[2]*Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York 11794, USA*

[3]*Department of Physics, National University of Singapore 2 Science Drive 3 Singapore, 117542 Singapore*

[4]*Global Strategic Analytics Unit, Citigroup, One Huntington Quadrangle, Suite 1N16, Melville, New York 11747, USA*

[5]*Department of Computational Chemistry, Pfizer Global Research and Development La Jolla, 10777 Science Center Drive, San Diego, California 92121-1111, USA*

(Received 6 June 2002; published 6 May 2003)

We study the nature of biomolecular binding. We found that in general there exists several thermodynamic phases: a native binding phase, a non-native phase, and a glass or local trapping phase. The quantitative optimal criterion for the binding specificity is found to be the maximization of the ratio of the binding transition temperature versus the trapping transition temperature, or equivalently the ratio of the energy gap of binding between the native state and the average non-native states versus the dispersion or variance of the non-native states. This leads to a funneled binding energy landscape.

DOI: 10.1103/PhysRevLett.90.188101 PACS numbers: 87.10.+e, 87.14.–g, 87.15.–v

Understanding biomolecular binding presents us a great intellectual challenge [1]. There are two crucial factors determining the binding process. One is the affinity that measures the stability of associating two molecules together. The other important issue is the specificity. High affinity often cannot guarantee the specificity. This is because many binding modes can have similar affinity, but only a few or none bind discriminately to influence the function [2].

In rational drug design, to reach the affinity and the specificity, accurate estimates for both structure and binding energy are needed but unfortunately are still lacking at present [3]. Recent advances in combinatorial chemistry open up a new way in the drug design industry [4]. Since sequences of small peptides can now be synthesized quite easily and quickly, binding experiments with peptides of many different sequences can be carried out. Specific receptor-ligand complex with good binding properties can be picked out just like the natural evolution selection process. A statistical approach is appropriate and necessary to characterize the sequence ensembles of ligand binding with a particular receptor. It will help to improve the design of the combinatorial synthesis, characterizing the energy landscape and understanding the underlying principles of binding.

The main difference between biomolecular folding and binding is the presence or absence of the chain connectivity [5]. The study of folding involves only one single polypeptide chain, while the binding occurs with at least two chains. It has long been recognized that the hydrophobic interactions are the driving force for both of these processes although it is quantitatively less for the binding process. This is due to the less freedom for the backbone to alter the configuration to obtain a larger extent of hydrophobic interactions in binding. It is also due to the compromises between the stability of preassociated binding biomolecules where less hydrophobic residues are exposed and the stability of the binding complex where more hydrophobic residues contribute to the hydrophobic interfaces of the binding. The long-range electrostatic interactions are believed to play the role of guiding and steering and also provide certain although somewhat smaller contributions to the energetics of the binding process. Biomolecular folding can be viewed as selective self-binding. So the folding and binding can be viewed as a similar process, representing intramolecular and intermolecular recognition.

Indeed, our methodology is based on the analogy between folding and binding. Great progress has been made recently in protein folding using the statistical energy landscape theory [6]. Similar to the biomolecular folding problem, there also exists the problem of the Levinthal paradox in binding [7]. The binding process in nature happens in a biological time scale, while the searching through all the possible configurational states takes cosmological time. To resolve this issue, the resulting binding energy landscape naturally should have a funneled shape towards the native binding state to guarantee the bindability against the fluctuations or wiggles along the binding paths [see panel (a) of Fig. 1] [8]. In fact, it is the purpose of this Letter to give a quantitative measure of binding selectivity or specificity [8] from the energy landscape theory of binding in parallel to the one given in optimal folding [9].

Let us turn to a model Hamiltonian that describes binding between two biomolecules. To first order approximation, we assume that the energetics that favors bringing two or multiple residues close together from two biomolecules is due mainly to the short-range hydrophobic driving force. Here we ignore the preassociation of the

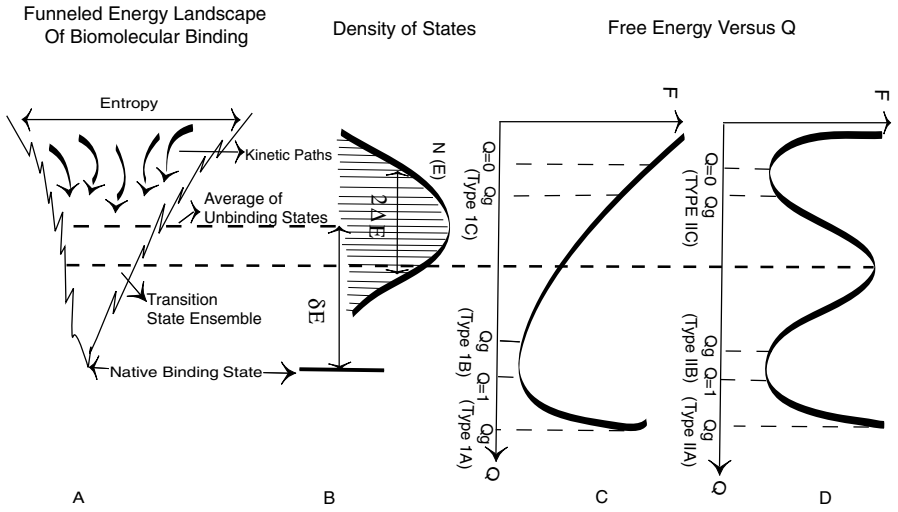

FIG. 1. Funneled energy landscape of the biomolecular binding: panel (a) shows the energy landscape of binding; panel (b) shows the density of states of the corresponding landscape; panel (c) shows the free energy profile for the corresponding landscape with respect to Q, the fraction of native contacts with a barrier between the native and the non-native states; panel (d) shows the free energy profile in Q without a barrier. δE is the gap and ΔE is the spread or the roughness of the energy landscape. Q_g is where the local trapping (local glass transition) occurs.

diffusional encountering process mainly due to the long-range electrostatic steering and guiding because of its limited contribution to the energetics. The form of interaction is $-\epsilon_{ijk\cdots p}(\alpha_i, \alpha_j, \alpha_k, \ldots, \alpha_p, r_i, r_j, r_k, \ldots, r_p)$, where $\epsilon_{ijk\cdots p}$ is the multibody coupling strength. r_i is the position of the ith residue, and α_i represents the physical properties of the residue i, for example, hydrophobic charges, etc. Here we also assume that the environmental solvent effects are already averaged out generating the multibody cooperative hydrophobic interactions among residues upon binding.

We may write down the Hamiltoninan energy function of a binding complex as

$$H = -\sum_{ijk\cdots p} \epsilon_{ijk\cdots p}\sigma_{ijk\cdots p}, \tag{1}$$

where $\sigma_{ijk\cdots p} = 1$ when there is a multibody contact adjacent in space made among monomers $ijk\cdots p$ between the two biomolecules in the binding complex and $\sigma_{ijk\cdots p} = 0$ otherwise. $\sum$ for some indices (for example, i) is up to N_1, and $\sum$ for other indices (for example, j) is up to N_2; N_1 is the sequence length of one biomolecule and N_2 for another in the binding complex. $\epsilon_{ijk\cdots p}$ is a random variable due to the sequence and interaction heterogeneity. Notice that this is mathematically closely related to the random energy model [10].

Suppose there exists a native configurational state n of energy E_n. We can find the probability that configuration a has energy E_a, given that a has an overlap Q with n, where Q is the fraction of native contacts of state a: $Q = \frac{1}{N}\sum_{ij}\sigma_{ij}^a\sigma_{ij}^n$ and N is the total number of native contacts. Q can be used as an order parameter or a reaction coordinate for the physical binding process that measures how close the states are towards native state. Note that for $Q = 1$, the state is in the native binding state and for $Q = 0$, the configurations are in totally non-native unbinding states.

The conditional probability is obtained directly by averaging over the Gaussian distribution of contact energy $\epsilon_{ijk\cdots p}$ ($\frac{\langle\delta[E_a - H(\{\sigma^a_{ijk\cdots p}\})]\delta[E_n - H(\{\sigma^n_{ijk\cdots p}\})]\rangle}{\langle\delta[E_n - H(\{\sigma^n_{ijk\cdots p}\})]\rangle}$). By approximating the cooperative multibody interactions $\sigma_{ijk\cdots p}$ in the Hamiltonian into the factorization of pair interaction terms $\sigma_{ij}\sigma_{jk}\ldots$ through a suitable decomposition law such as in the superposition approximation in the theory of fluid, the expression can be simplified as $\frac{P_{an}[E_a,Q,E_n)}{P_n(E_n)} \sim \exp(-\frac{[(E_a-\bar{E})-Q^{m-1}(E_n-\bar{E})]^2}{2N\Delta\epsilon^2(1-Q^{2(m-1)})}]$, where m is the order of the interactions ($m = 2$ for two body interactions, $m = 3$ for three body interactions..., and $m = p$ for p body interactions), $\bar{E}$ is the average mean energy, and $\Delta\epsilon$ is the effective width of the energy distribution per contact.

The configurational entropy $S_{\rm tot}$ as a function of similarity Q with a given state is treated in detail by the previous studies [11]. For binding, due to the geometrical constraints, the slope of the entropy nearly linear in Q is expected to be reduced, but the functional form remains similar to folding except for some quantitative effects of chain connectivity, translation, and rotation. Given the $S_{\rm tot}(Q)$ and conditional probability distribution obtained earlier, the average numbers of states of energy E and overlap Q with native state n are $\langle n(E, Q, E_n)\rangle = \exp[S_{\rm tot}(Q)]\frac{P(E,Q,E_n)}{P(E_n)}$. This is effectively the microcanonical ensemble description of the thermodynamics. At each stratum of the order parameter or reaction coordinate Q, the set of states is modeled by a random energy model. By the thermodynamic relation of $\frac{\partial\log\langle n(E,Q,E_n)\rangle}{\partial E} = 1/T$, we can obtain the energy and entropy of the biomolecular binding complex as $E(T, Q, E_n) = \bar{E} + Q^{m-1}(E_n - \bar{E}) - \frac{N\Delta\epsilon^2(1-Q^{2(m-1)})}{T}$ and $S(T, Q, E_n) = Ns_{\rm tot}(Q) - \frac{N\Delta\epsilon^2(1-Q^{2(m-1)})}{2T^2}$, where $s_{\rm tot}(Q) = S_{\rm tot}(Q)/N$. The entropy vanishes at a characteristic temperature: $T_g = \Delta\epsilon\sqrt{\frac{(1-Q^{2(m-1)})}{2s_{\rm tot}(Q)}}$ which signals the trapping of the biomolecular binding complex into a low energy conformational state within the stratum characterized by Q. Notice that when $Q = 0$ (non-native unbinding states), $T_g = \Delta\epsilon\sqrt{\frac{1}{2s_{\rm tot}(Q=0)}}$.

From the thermodynamic expression of energy and entropy given above, we can easily obtain the expression for the free energy per contact as

$$\frac{F}{N}(T, Q, E_N) = -Ts_{\text{tot}}(Q) - Q^{m-1}\delta\epsilon_n - \frac{\Delta\epsilon^2}{2T}(1 - Q^{2(m-1)}), \quad (2)$$

where $\delta\epsilon_n = |(E_n - \bar{E})/N|$. The free energy is composed of three terms, the entropy, the native driving force, and the roughness contribution of the energy landscape. In the parameter space in $(\delta\epsilon_n, \Delta\epsilon, T)$, the expression above can have a double minimum structure in the reaction coordinate Q with one minimum at low Q corresponding to the non-native states separated by a barrier from another minimum at high Q corresponding to the native binding state. As the cooperativity measured by multibody interaction order m increases, the free energy minimum of non-native states and native binding state shift towards $Q \sim 0$ and $Q \sim 1$, respectively. To the extent that this approximation is good ($m \to \infty$), we can equate the free energies of the non-native states and native binding state to obtain the the binding transition temperature $[F(Q = 0) = F(Q = 1)]$: $T_b = \frac{\delta\epsilon_n}{2s_{\text{tot}}(Q=0)} \times (1 + \sqrt{1 - \frac{2s_{\text{tot}}(Q=0)\Delta\epsilon^2}{\delta\epsilon_n^2}})$. Notice that when the order of the interaction is finite, the free energy minimum is likely to be at $Q \sim q$ and $Q \sim 1 - q$ (whereeas $0 < q \ll 1$), respectively, by setting $\frac{\partial F}{\partial Q} = 0$ and $\frac{\partial^2 F}{\partial Q^2} > 0$. By further setting $F(Q = q) = F(Q = 1 - q)$, one can show the solution of the resulting equation for T_b approaches the above expression when $q \to 0$, that is $m \to \infty$.

Taking the ratio of binding temperature and trapping temperature, we obtain

$$T_b/T_g(Q = 0) = \Lambda + \sqrt{\Lambda^2 - 1}, \quad (3)$$

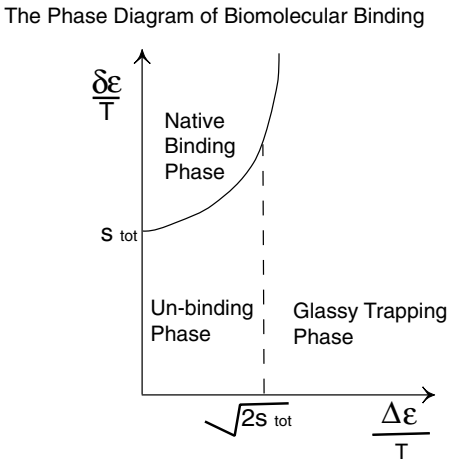

FIG. 2. Phase diagram of biomolecular binding: the vertical axis is the temperature scaled gap or the slope, and the horizontal axis is the temperature scaled roughness of the energy landscape.

where $\Lambda = \frac{\delta\epsilon_n}{\Delta\epsilon\sqrt{2s_{\text{tot}}(Q=0)}}$ is the ratio of the energy gap between the native state and the average of the energy landscape spectrum and the ruggedness or the width (spread) of the distribution of the energy landscape spectrum weighted by entropy per contact $\sqrt{2s_{\text{tot}}(Q = 0)}$ which is on the order of 1.

There are at least three possible thermodynamic phases, the native binding, the non-native, and the glass or trapping phase (see Fig. 2). Clearly, the binding transition temperature should be higher than the trapping temperature in order to guarantee the specificity and to avoid nondiscrimination with traps. In order to assure that, the ratio T_b/T_g should be maximized. From the expression of Eq. (3), this is the equivalent of saying that Λ should also be maximized. Therefore maximizing the ratio of the energy gap (or the slope) versus the roughness of the underlined energy landscape becomes the criterion for the specificity of binding [see panel (b) of Fig. 1]. Only the binding landscape satisfying this criterion will be able to form a thermodynamically stable native complex and bind specifically and, furthermore, survive the natural evolution. This implies a funneled energy landscape of binding analogous to folding where there is a directed steep slope biasing towards the native binding state dominating the fluctuations or wiggles superimposed on the landscape. From this picture, at the initial stage of binding, there are multiple parallel paths leading towards the native binding state. As the binding process progresses, the discrete paths emerge and give dominant contributions only when the roughness of the landscape becomes significant [12].

The optimization criterion above provides a quantitative measure of specificity for biomolecular binding and interactions. It is also potentially important for drug screening. Instead of using only affinity or free energy as the criterion for screening as is often seen in docking studies, one can input the specificity constraints. This is particularly relevant to the combinatorial synthesis [13] searching through sequence space to find the best binding ligand mimicking the natural evolution process.

The binding energy landscape can be divided into at least two major classes in general: type I for the downhill case without the free energy barrier and type II case with the free energy barrier [see panels (c) and (d) in Fig. 1]. Notice that T_g is dependent on Q. There exists a possibility that $T_g(0) < T_b < T_g(Q)$, in particular Q_g. If Q_g is close to zero (non-native states) (type IC and type IIC), then binding is trapped into the local minima at an early stage of the binding process. If Q_g is close to the native state (type IB and type IIB), then at $Q < Q_g$, there are multiple paths leading to binding. Although the barriers are high near $Q > Q_g$, there are only limited numbers of configurational space left to explore at this late stage of binding and kinetic paths to the native state are discrete. The kinetic process is expected to be nonexponential. If $Q_g > 1$ (type IA and type IIA), the landscape is smooth,

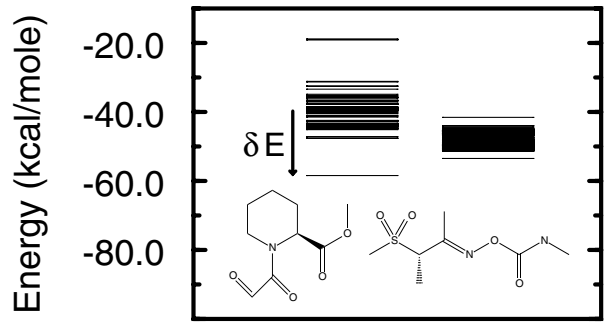

FIG. 3. Energy spectra and chemical structures for the pipecolinyl fragment (left) and the second fragment (right) of size 14 heavy atoms, representative of molecules randomly taken from the commercially available Fine Chemicals Directory (MDL Information Systems, Inc., San Leandro, CA). Energies of structures in the native binding mode, where $R_{\text{native}} = 1.5$ Å are collapsed into a single value. The stability gap δE between the binding mode and the average binding mode for the pipecolinyl fragment is indicated.

multiple paths are funneled towards the native binding state, and the kinetics is expected to be exponential.

As a practical example, we took a model system developed by one of the authors (G.V.) [8] previously. We have performed a Monte Carlo multiple docking studies of the ligand-protein binding process for small molecular fragments in the FK506 inhibitor binding to the protein FKBP12. The energy function is a coarse-grained knowledge based one which includes intramolecular energy terms for the ligand, given by torsional and nonbonded functions and intermolecular ligand-protein steric and hydrogen bond interaction terms calculated from a piecewise linear potential summed over all protein and ligand heavy atoms. Both the common core recognition motif, the pipecolinyl moiety, and another random fragment (the ligand) have been chosen to dock to the FKBP12 protein. The binding energy spectrum was obtained and is shown in Fig. 3. We have calculated the relative stability gap as well as the roughness and the specificity ratio (gap versus roughness) for both the recognition motif spectrum (left spectrum of Fig. 3, $\delta E = 17.3$, $\Delta E = 4.08$, $\delta E/\Delta E = 4.24$) and the randomly chosen fragment spectrum (right spectrum of Fig. 3, $\delta E = 5.7$, $\Delta E = 2.98$, $\delta E/\Delta E = 1.91$). We notice that the spectrum for the recognition motif has a significant energy gap between the native binding mode and the binding states with the average energy (the specificity ratio is significantly larger by a factor of 2) with good structural consnsus, while the one for the randomly chosen fragment does not and has poor structural consensus in multiple docking simulations, indicating a multitude of structurally different binding modes. When including many more different random fragments from the Fine Chemical Directory for the docking studies, it appears that the nativeness (RMSD) is monotonically correlated with the specificity measure $\frac{\delta E}{\Delta E}$. When $\frac{\delta E}{\Delta E}$ is large, then the RMSD is likely to be small (more native), or vice versa. This shows the necessarty of a funneled binding energy landscape with a significant energy gap to roughness ratio to realize the discriminating specificity.

In this study, we have confined ourselves to study the binding dynamics of two already folded biomolecules. In nature, folding and binding process can be dynamically coupled. This is a more complicated and challenging problem currently under active investigation by the Wolynes group at University of California at San Diego [14].

J.W. thanks Professor P. G. Wolynes for constant encouragement and inspiration during the course of this work. J.W. also thanks Dr. Garegin A. Papoian and Professor Terry Hwa for helpful discussions.

*Electronic address: jinwang@sprynet.com
†Electronic address: gennady.verkhivker@pfizer.com

[1] D. E. Jr. Koshland, Proc. Natl. Acad. Sci. U.S.A. **44**, 98 (1958); J. A. McCammon, Curr Opin Struct Biol **8**, 245 (1998).
[2] A. Wlodawer and J.W. Erickson, Annu. Rev. Biochem. **62**, 543 (1993).
[3] J. Cherfils and J. Janin, Curr Opin Struct Biol **3**, 265 (1993).
[4] M. A. Gallop, R.W. Barrett, W. J. Dower, S. P. A. Fodor, and E. M. Gordon, J. Med. Chem. **37**, 1233 (1994); **37**, 1385 (1994).
[5] C. J. Tsai, D. Xu, and R. Nussinov, Folding Des. **3**, R71 (1998).
[6] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995), and references therein.
[7] C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, edited by P. DeBrunner, J. Tsibris, and E. Munck (University of Illinios Press, Urbana, 1969), p. 22.
[8] P. A. Rejto and G. M. Verkhivker, Proc. Natl. Acad. Sci. U.S.A. **93**, 60 (1996); **93**, 8945 (1996); J. Janin, Proteins **25**, 438 (1996); C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov, Protein Sci. **8**, 1181 (1999); A. Tovchigrechko and I. A. Vakser, Protein Sci. **10**, 1572 (2001).U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **99**, 12 105 (2002)
[9] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918 (1992).
[10] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980); Phys. Rev. B **24**, 2613 (1981).
[11] S. S. Plotkin, J. Wang, and P. G. Wolynes, Phys. Rev. E **53**, 6271 (1996); J. Chem. Phys. **106**, 2932 (1997).
[12] J. Wang, J. N. Onuchic, and P. G. Wolynes, Phys. Rev. Lett. **76**, 4861 (1996).
[13] J. G. Saven and P. G. Wolynes, J. Phys. Chem. B **101**, 8375 (1997).
[14] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **97**, 8868 (2000); G. A. Papoian and P. G. Wolynes, in *Proceedings of the 223rd ACS Meeting, Phys. Chem. Abstract 257* (American Chemical Society, Washington, D.C., 2002). G. A. Papoian and P. G. Wolynes, Biopolymers **68**, 333 (2003).