NIH Public Access
Author Manuscript
Chem Rev. Author manuscript; available in PMC 2009 August 31.

Published in final edited form as:
Chem Rev. 2006 May ; 106(5): 1559–1588. doi:10.1021/cr040425u.

Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry and Biology Meet

Eugene Shakhnovich
Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge MA 02138

1. Introduction

As was noted in our recent review [1] the protein folding field underwent a cyclic development. Initially protein folding was viewed as a strictly experimental field belonging to realm of biochemistry where each protein is viewed as a unique system that requires its own detailed characterization – akin to any mechanism in biology. The theoretical thinking at this stage of development of the field was dominated by the quest to solve so-called "Levinthal paradox" that posits that a protein could not find its native conformation by exhaustive random search. Introduction, in the early nineties, of simplified models to the protein folding field and their success in explaining several key aspects of protein folding, such as two-state folding of many proteins, the nucleation mechanism and its relation to native state topology, have pretty much shifted thinking towards views inspired by physics. The "physics"-centered approach focuses on statistical mechanical aspect of the folding problem by emphasizing universality of folding scenarios over the uniqueness of folding pathways for each protein. Its main achievement is a solution of the protein folding problem *in principle*, i.e. demonstration how proteins *could* fold. As a result, a "psychological" solution of the Levinthal paradox was found (i.e. it was generally understood that this is not a paradox. after all). The key success of this stage of the field is discovery of the general requirements for polypeptide sequences to be cooperatively foldable stable proteins and realization that such requirements can be achieved by sequence selection. That put the field strongly into the realm of biology ("Nothing in Biology makes sense except in the light of Evolution" (Theodosius Dobzhansky)) The physics-based fundamental approach to protein folding dominated theoretical thinking in the last decade (reviewed in [1-4]) and its successes brought theory and experiment closer together

At the present stage we seek better understanding of how protein folding problem is *actually* solved in Nature. In this sense the protein folding field has made a full circle as attention is again focused on specific proteins and details of their folding mechanism. However these questions are asked at a new level of sophistication of both theory and experiment. Understanding of general principles of folding and vastly improved computer power makes it possible to develop tractable models that sometimes achieve atomic level of accuracy. Further, better general understanding of requirement for polypeptide sequences to fold, lead to establishment of direct links between protein folding and evolution of their sequences This development opened an opportunity to employ powerful methods of bioinformatics to test predictions of various folding models, in addition to more traditional tests of models against experiment After all, evolution presents a giant natural laboratory where sequences are designed to fold and function and availability of vast amounts of data certainly calls for its use to better understand folding of proteins at very high resolution. At the same time in vitro

Email Shakhnovich@chemistry.harvard.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

experimental approaches progressed to the point that very accurate time- and structure-resolved data are available. A close interaction with experimentalists helps to keep theorists honest by providing detailed tests of theories and simulation results.

In this review, which to a great extent reflects the thinking of the author on the subject, we will first summarize basic questions and present simple, coarse-grained models that provide a basis for a fundamental understanding of protein folding thermodynamics and kinetics. Then we will discuss more recent developments (over last five years) that focus on detailed studies of folding mechanisms of specific proteins, and finally we will briefly discuss some outstanding questions and future directions.

2. Random and designed heteropolymers – a fundamental model of protein folding

2.1 Random heteropolymers do not fold cooperatively

At the very basic level of coarse-grained microscopic models, statistical mechanics provided tools that facilitated our understanding of many fundamental and universal properties of proteins. A fundamental statistical-mechanical model of a protein, is a heteropolymeric molecule[5]. Its study provided many insights into thermodynamic and kinetic properties of proteins[5-8].

Studies of protein folding using coarse grained protein models followed two routes; A phenomenological approach was proposed by Bryngelson and Wolynes who *postulated* certain type of energy landscape (Random Energy-Model-like) for a protein-like molecule and explored consequences of such postulated energy landscape for protein thermodynamics [6] and kinetics [9]. The Random Energy Model was introduced by Derrida as a simplest model of spin glasses [10]. It is a phenomenological model that assumes that a system has M microstates (in the case of proteins each microstate is a conformations) and that energies of these microstates represent statistically independent random values drawn from Gaussian distribution. Bryngelson and Wolynes postulated just that for energies of different conformations of a protein-like heteropolymer. In addition to that they postulated that proteins have also a special conformation-native state - and that each aminoacid can be either in its native conformation or in any of *v* non-native ones. The authors adopted “The Consistency Principle” proposed by Go [11] (termed in [6] as “Principle of minimal frustrations”) by assuming that when aminoacids are in their native conformations their intrinsic energy, secondary structure energy and pairwise interaction energy is lower than for interacting aminoacids that adopt non-native conformation.

An alternative approach was proposed by Garel and Orland [12] and Shakhnovich and Gutin. [5] It is based on a statistical-mechanical analysis of a microscopic model that does not assume any landscape or conformational preferences *a'priori*. Rather, it derives energy landscape of a model protein from “first principles” – i.e. taking into account only a polypeptide chain connectivity and known set of interactions - and evaluates its consequences for thermodynamics and kinetics of folding.

The statistical-mechanical model defines a microscopic Hamiltonian, i.e. how energy of a conformation depends on the coordinates of all its atoms and on (fixed) protein sequence:

$$H(\{r_i\},\{\sigma_i\})=\sum_{i<j}B(\sigma_i,\sigma_j)U(r_i-r_j) \tag{2.1}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Where a conformation is determined through set of its atomic coordinates $\{r_i\}$. The protein chain's sequence is $\{\sigma_i\}$, the interaction energy between aminoacids of types σ_j and σ_i, depends on distance between them (via potential energy function $U(r_i-r_j)$) and their chemical identities – via interaction potential matrix B. The partition function of the model protein is a sum over all its conformations:

$$Z=\sum_{\text{conf}} g(r_i - r_{i+1}) \exp\left(-\frac{H(\{r_i\},\{\sigma_i\})}{kT}\right) \tag{2.2}$$

where $g(r_i-r_j)$ is a function describing connectivity of a chain [13]; it accounts for the chemical structure of the polypeptide representing (conditional) probabilities that residue i+1 is found around r_{i+1} when preceding residue, i is at r_i. Several forms for the function g were proposed in the literature [13,14]; selection of g corresponds to the model choice of the local (along the sequence) interactions; such choice determines mechanism of flexibility of a polypeptide chain. In principle Eqs.(2.1)-(2.2) are sufficient to fully evaluate sequence-dependent thermodynamic properties of a protein model. In practice their solution and analysis presents a formidable task both conceptually and technically.

Conceptually, the issue is what questions can be meaningfully asked within such a theoretical framework? It is quite clear that a low-resolution description is not suitable for prediction of thermodynamics and kinetic properties of *specific* proteins. Apparently, this class of coarse-grained models may be most suitable to address questions related to generic properties of proteins, common to all of them or to a wide range of protein sequences. Some questions that received much attention in the context of coarse-grained analytical models are:

1. What are the general requirements for sequences to be protein-like, i.e. to have a *stable* unique native structure as its lowest energy conformation?
2. Which sequences fold cooperatively (i.e. thermodynamically two-state) into their native conformation?
3. Are thermodynamic requirements of stability and cooperativity of the native conformation sufficient to make this conformation kinetically accessible, or is additional sequence selection to ensure kinetic accessibility necessary?

A key technical difficulty in studying the heteropolymer model of proteins is that proper averaging over sequences is required. This represents both a conceptual and technical challenge. Conceptually the difficulty is that one has to select such properties of a heteropolymer whose average values are representative of the majority of individual realizations, i.e. whose probability distributions are sharply peaked around average values. In this case evaluation of averages will be meaningful as it will describe a majority of individual molecules. *Physical quantities, whose averages are representative of a majority of realizations of a random system, are called self-averaging*. It was shown, first in theory of spin glasses that free energy (i.e. –kTlnZ) is a self-averaging quantity, while e.g. the partition function itself, Z, is not self-averaging. This can be understood if one realizes that very rare, atypical realizations of sequences (e.g homopolymers) can make exponentially large contribution to the partition function. As a result, despite the fact that such realizations are extremely rare (e.g. the probability to have a polyvaline molecule of N residues in the ensemble of randomly synthesized sequences is 20^{-N}) the overall contributions from such atypical sequences to the average partition function may be significant since their energy in some conformations (e.g. compact globule) may be very low, so low that their Boltzmann factor exp(-H/kT) in Eq.2.2 overwhelms the weight 20^{-N} corresponding to the slim probability to find such sequence). As a result average partition function may be heavily affected by sequences that are very atypical

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

members of the ensemble of protein sequences. On the other hand, contributions of very atypical sequences to *free energy* are at most ~N and such contributions from highly atypical sequences are easily overwhelmed by exponentially low probability of their occurrence.

Therefore, in order to obtain representative description of protein thermodynamics in analytical heteropolymer model, one should average, over sequences in the ensemble, *the free energy* of a protein chain

$$<F(T)>=-kT\sum_{\{\sigma\}}P(\{\sigma\})\log(Z(\{\sigma\},T) \tag{2.3}$$

where <> denotes average over all sequences, $P(\{\sigma\})$ is probability of occurrence of a sequence $\{\sigma\}$ in the ensemble and summation is taken over all sequences. The next and even more conceptually difficult question is over which ensemble of sequences to take average in Eq.(2.3). Averaging over unbiased ensemble of all possible sequences (i.e. assuming P=const in Eq.(2.3)) means that protein sequences are treated as being randomly selected from the pool of all possible sequences, i.e. no evolutionary selection (pressure) on protein sequences is assumed. Averaging over biased ensemble of sequences corresponds to evolutionary selected sequences. Thus possible evolutionary selection enters the theory via the probability distribution $P\{\sigma\}$ in sequence space (see below).

Averaging in Eq.(2.3) is a daunting task because the partition function to be averaged enters it under logarithm. However it is possible to evaluate <F> in Eq.2.3 using replica approach which was first proposed by Edwards and Anderson [15] and then significantly developed further by Parisi and coworkers [16] in the context of Spin Glass studies. The replica method is an ansatz based on the relation:

$$<\log Z>\grave{=}\lim_{n\to 0}\frac{<Z^n>-1}{n} \tag{2.4}$$

and observation that $<Z^n>$ is relatively easy to evaluate when n is integer – it is the average, *over all sequence realizations*, partition function of n identical systems (replicas, hence replica method). While analytic continuation of expression (2.4) to noninteger values of n is a mathematically very challenging task whose subtleties are not still fully understood, the technique was sufficiently developed in Spin-Glass theory to provide major insight into its equilibrium and non-equilibrium properties.

Heteropolymer theory as the basis for a fundamental understanding of protein folding was developed within the framework of the replica approach by Shakhnovich and coworkers [5, 17-19]. Detailed analysis based on of Eqs(2.1-4) revealed not only thermodynamic properties of random heteropolymers but provided major insights into nature of their energy landscape. It turns out that replica averaging over sequences results in an emergence of the order parameter that turns out to be extremely useful to understand the general properties of energy landscape of heteropolymers. In order to see this we consider the simplest case of contact Hamiltonian:

$$H(\{r_i\})=\frac{1}{2}\sum_{i,j}^{N}B\left(\sigma_i,\sigma_j\right)\delta\left(r_i-r_j\right) \tag{2.5}$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

where δ denotes that two aminoacids interact (with energy $B(\sigma_i,\sigma_j)$ depending on their types σ_i,σ_j when they are in spatial proximity to each other. (An important non-specific three-particle interaction term is omitted in (2.5) for brevity; full analysis is in [5]). Further, assume, following [5], that interaction energies $B_{ij}= B(\sigma_i,\sigma_j)$ can be approximated as independent random values drawn from Gaussian distribution, i.e.:

$$P\left\{\sigma\right\}=\prod_{i,\ j} p\left(B_{ij}\right) \tag{2.6}$$

and

$$p\left(B_{ij}\right)=\frac{1}{\left(\pi B^2\right)^{1/2}} e^{-\frac{\left(B_{ij}-B_0\right)^2}{2B^2}} \tag{2.7}$$

where B is a standard deviation of interaction energies between different types of aminoacids and B_0 is average interaction: if $B_0 < 0$ attraction prevails, in average, giving rise to tendency to chain collapse and if $B_0 > 0$ repulsion prevails, on average.

Averaging of Z^n over sequences leads to expression:

$$\langle Z^n\rangle=\int\prod_{\alpha=1}^{n}\prod_{i=1}^{N-1} g\left(r_i^\alpha - r_{i+1}^\alpha\right) e^{-\frac{\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i,j}^{N} B_{ij}\delta\left(r_i^\alpha - r_j^\alpha\right)}{kT}} \prod_{i,\ j=1}^{N} p(B_{ij}) dB_{ij} \prod_{i,\alpha}^{N,n} dr_i^\alpha \tag{2.8}$$

Here new “replica index” α appeared as a direct consequence of averaging n-th power of the partition function. One can view it mnemonically as averaging the partition function of n identical sequences that do not interact between themselves. Averaging over sequences in (2.8) (i.e. integration over dB_{ij}) is performed first; it amounts to evaluation of many independent Gaussian integrals. The result of averaging over sequences is emergence of an effective Hamiltonian such that:

$$\langle Z^n\rangle=\int\prod_{\alpha=1}^{n}\prod_{i=1}^{N-1} g\left(r_i^\alpha - r_{i+1}^\alpha\right) e^{-\frac{H_{\text{eff}}\left\{r_i^\alpha\right\}}{kT}} \prod_{i,\alpha}^{N,n} dr_i^\alpha \tag{2.9}$$

where

$$H_{\text{eff}}\left\{r_i^\alpha\right\}=\frac{1}{2}\ \tilde{B}\sum_{\alpha,i\neq j}\delta\left(r_i^\alpha - r_j^\alpha\right)-\frac{B^2}{4kT}\sum_{\alpha\neq\beta}\sum_{i\neq j}\delta\left(r_i^\alpha - r_j^\alpha\right)\delta\left(r_i^\beta - r_j^\beta\right) \tag{2.10}$$

Where $\tilde{B} = B_0 - B^2/2kT$ is renormalized (due to heterogeneity of interactions) “average” interaction strength. The second term is most important as it introduces new and extremely valuable order parameter that “mixes” different replicas (we remind the reader that Greek letters α,β etc denote replicas here).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$q_{\alpha\beta}=\sum_{i}\delta(r_i^{\alpha}-r_j^{\alpha})\delta(r_i^{\beta}-r_j^{\beta}) \tag{2.11}$$

whose simple physical meaning can be understood when one considers analogy between replicas (marked by index α,β etc) and configurations of the heteropolymer chain in its deep energy minima where it spends significant amount of time. Note again that δ-symbols in Eq. (2.11) count contacts i.e. it is 1 if monomers i and j are in contact (i.e. within certain short distance from each other) and 0 otherwise. Apparently the order parameter introduced in Eq. 2.11 counts the number of common contacts, i.e. structural overlap, between chains in two configurations corresponding to deep energy minima. The quantity that provides a comprehensive description of the energy landscape of the heteropolymer is then

$$P(Q)=\sum_{\{r_{\alpha}\},\{r_{\beta}\}}p(\{r_{\alpha}\})\,p(\{r_{\beta}\})\delta(Q-q_{\alpha\beta}) \tag{2.12}$$

where $p(\{r_{\alpha}\})$ is Boltzmann probability to be in state where the chain has coordinates $\{r_{\alpha}\}$. It is quite clear that only deep minima contribute to P(Q) because only for them Boltzmann probabilities p have noticeable values. The physical meaning of the equation 2.12. is simple. If one statistically samples conformations with their thermal probabilities (so that only conformations residing in deep energy minima contribute) then P in eq. (2.12) is probability that conformations from two minima have structural similarity Q. In other words P statistically characterizes the landscape in terms of how structurally different deep minima are.

The detailed calculations and analysis carried out along these lines in series of publications [5,17,18] (reviewed in [19]) provide a comprehensive description of the thermodynamic properties and energy landscape of random heteropolymers. It turns out that properties of random heteropolymers depend on dimensionality of space in which they are embedded with d=2 being a critical dimension separating two qualitatively different types of behavior. The analysis of low-dimensional case $d \leq 2$ was carried out in [20,21] where it was shown that energy landscape in this case is hierarchical, “smooth” in a sense that most low-energy conformations have significant structural similarity to the conformation with lowest energy, “native” one. It was argued in [21] that this property of energy landscape of low-dimensional heteropolymers is due to a very important role that polymer bonds play in this case: in compact states of low-dimensional polymers the majority of contacts appear to be between residues that are near to each other along the chain. While the low-dimensional heteropolymer case is of little relevance to proteins, the replica-space variational approach developed in [21] to treat such heteroplymers, was used by Mezard and Parisi to study random manifolds [22] and since then has been adopted in various fields including studies of polymer gels [23] and certain types of fermionic systems including high-Tc superconductors.[24]

The full analysis for a more relevant case of three-dimensional space [5,20] showed that the “energy landscape” of random heteropolymers is “rugged” in the sense that it consists of several deep energy minima of comparable (differing by just few kT per molecule) energies but conformations belonging to these minima are structurally unrelated. These deep energy minima which are structurally very different from the native state can serve as traps en route to the native state – hence their possible importance for folding kinetics. Thermodynamically a significant fraction of random heteropolymers can be stable in the “native state” (lowest energy conformation) [25] but that can happen only at low enough temperature and, most importantly,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the transition to the native state upon temperature decrease is gradual, akin to the transition to zero entropy state in the Random Energy Model[5,10]

The approximation of mutually statistically independent Gaussian-distributed energies of interactions *between aminoacids*Eq.(2.5) simplifies calculations significantly. It corresponds to the case when the number of aminoacid types is large [20]. The opposite case – of only two types of aminoacids, such as hydrophobic and polar, – was solved in 1993 by Sfatos et al [17]. In this case one can no longer assume independence of interaction energies between aminoacids and a new theoretical formalism (a version of Stratonovich-Hubbard transformation) was developed to tackle this issue. A new factor has to be considered in case of heteropolymers with two types of aminoacids - a possibility of microphase separation of aminoacids of different types (e.g. separation between hydrophobic core and hydrophilic surface). An interesting results of the analysis of the "two-letter" heteropolymers is that microphase separation and chain "freezing" (i.e. dominance of one or very few lowest-energy structures) may in certain cases compete with each other, e,g. chain freezing may prevent microphase separation under certain conditions (see [26] where complete phase diagram of a two-letter random heteropolymer is presented). However the energy landscape in the case of "two-letter" random heteropolymers appears to be the same as for the model of independent interactions – consisting of sets of deep energy minima corresponding to conformations that are structurally unrelated to each other. A general case of multi-letter heteropolymers was considered in [18]. A detailed, more technical, discussion of these issues and further references can be found in the '97 review by Sfatos and Shakhnovich [19].

2.2 Theory of evolutionary selected sequences: Protein-Like cooperative behavior

The main conclusion from the analysis of random heteropolymers is that they do not exhibit many protein-like properties such as, cooperativity of their folding transition [276]. Further it was shown that native structures of random heteropolymers are extremely susceptible to mutations: Probability that a random mutation in a random heteropolymer does not result in a dramatic change of native structure was found in [28] to be very slim. Apparently such instability to mutations is not conducive to proper evolutionary selection and is in direct disagreement with genomic observations.

The inadequacy of random heteropolymer model to describe proteins is perhaps not surprising as proteins are biological macromolecules whose sequences underwent evolutionary selection. In particular, it was first posited by Go [11] that proteins should have special properties, such as, "consistency between different types of interactions and structures", [11] or, later by Bryngeslon and Wolynes, that all interactions between aminoacids that are in their native conformations are energetically preferable. by a certain margin [6]. Bryngelson and Wolynes carried out a kinetic analysis of the same model. Their kinetic assumption was that attaempts at transitions occur between states whose energies are uncorrelated and the dynamics (acceptance or rejection of the attempt to move between states) is governed by Metropolis criterion. The conclusion from calculations presented in [9] was that there exists a particular temperature, called T_g ("g" stands for glass) that at all temperatures at or below T_g folding time of a protein equals Levinthal time [9]. According to the Bryngelson and Wolynes calculations fastest folding apparently occurs in their model at infinite temperature (see Fig. 3 of [9]) but the reason for this unphysical result may be due to dependence of parameters of the model on temperature. The Bryngelson and Wolynes study of kinetics within the REM approximation and their prediction of glass transition was further analyzed by Gutin et al [29]. Besides pointing out to technical issues with Bryngelson and Wolynes kinetic REM calculation[9], these authors carried out folding simulations for lattice model within a broad range of temperatures and for several native structures. They found no signature of glass transition in these simulations – just a pure Arrenhius dependence of folding rate on temperature and exponential distribution of folding

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

times. Gutin et al proposed a simple REM-based phenomenological model of kinetics that correctly reproduced temperature dependence of folding rates in simulations [29]

Analytical replica-based study of the microscopic model similar in spirit to that of Go was performed in 1989 by Shakhnovich and Gutin [30]. The interaction Hamiltonian was assumed in [30] to be Go-like:

$$H\left(\{r_i\}\right)=-\frac{1}{2}\sum_{i,\,j}^{N}B\delta\left(r_i-r_j\right)\delta\left(r_i^0-r_j^0\right)+\frac{1}{2}\sum_{i,\,j}^{N}B_0\delta\left(r_i-r_j\right) \tag{2.13}$$

where $\{r^0\}$ is the set of coordinates of the native conformation, B_0 is average interaction energy. The first term in (2.13) is a manifestation of the Go model: it posits that interactions between aminoacids, which are in contact in the native conformation, are energetically favorable by energy margin B. The Go model Eq.(2.13) presented in [30] features an important property: the native conformation, having n_c contacts is separated by extensive energy gap $-Bn_c$ from the set of misfolded compact conformation (molten-globule like). This is a defining feature of most Go-models, at least in 3-dimensional space. The full statistical-mechanical analysis of the model in (2.13)[30] (where replica method was used to average free energy over all possible native conformations $\{r^0\}$) showed that in this case the transition to the native state occurs as a true cooperative, first-order-like phase transition.

While earlier works [6,9,11,30] relied explicitly on the assumption that aminoacids in their native conformations or making native contacts have special energetic preference, a more general thermodynamic condition for heteropolymers to be protein like was discussed in [25]. The authors of [25] studied the conditions for thermodynamic stability of the unique native state and introduced explicitly the concept of energy gap, i.e. energy difference lowest energy (native) state and lowest energy misfold as the main factor that determines thermodynamic stability of the native state. Further, they determined the probability that heteropolymers with unique native state can be found in “one-shot” selection from the pool of random sequences, at certain temperature. They found that one-shot selection is able to find (with low but non-vanishing probability) sequences that have large gap and, correspondingly, stable native structure. However the condition of thermodynamic stability of the native state, found in [25] while indeed requires large enough gap (several kT) it does not require that the gap is extensive in chain length (i.e. proportional to chain length when different proteins are compared). According to [25] the probability to find a sequence with stable native state in a “soup” of random heteropolymers becomes extremely low if temperature exceeds T_c - the temperature of the freezing transition in the heteroplymer model of [5,20] (same as T_g of [6,9])

In 1992 Goldstein et al [31] presented a phenomenological model that explicitly assumed, without resorting to special “native-like” interactions, that native state is separated by energy gap from the set of non-native conformations,. Their reasoning, was that in order to be able to fold proteins must be stable at temperature above T_g, i.e. their unfolding temperature T_f must be higher than T_g. Further they introduced ratio of T_f / T_g as criterion of protein foldability and sough to optimize energy parameters for protein Hamiltonian to maximize this quantity., In fact the “glass transition temperature” of Goldstein et al [31] is equivalent to putative freezing transition temperature in a fully random system that is identical to the phenomenological protein model – has the same set of states - but without the single unique, specific natïve state. A detailed analysis and critique of the concept of glass transition in heteropolymer systems can be found in [29]. In a somewhat similar vein Camacho and Thirumalai [32] suggested a “foldability criterion” of T_f / T_θ - a ratio between folding temperature and random collapse temperature. Dinner et al [33] provided a comparative analysis of various foldability criteria.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

In fact the T_f / T_g criterion of Wolynes et al is equivalent to the requirement that the native state is separated by energy gap from misfolds [31]. The difference between this important criterion of Wolynes and colleagues and earlier gap analysis of Shakhnovich and Gutin [25] is that $T_f / T_g > 1$ criterion implies that energy gap is extensive, i.e. proportional to chain length (that can be discerned from Eq.2 of [31] upon straightforward additional analysis). In contrast the analysis in [25] suggested that thermodynamic stability of the native state alone does not require extensive gaps. However extensive gaps provide not only stability to the native state but also cooperative, first-order like folding transition.

An important insight from microscopic theory [30] and phenomenological models, [27,31], is that existence of extensive energy gap between low-energy native conformation and lowest energy non-native, misfolded, conformation – is *sufficient* to make folding transition cooperative, first-order like as it is indeed observed in many wild-type proteins [34]. . However at that time (late eighties early nineties) it was unclear to many researchers whether a large (extensive) energy gap is also a *necessary* condition for cooperative protein folding. Indeed theoretical analysis [35] suggested that cooperative transition may originate from other physical factors such as side-chain ordering, while energy gap, being still very important to stabilize the native state at room temperature [25] does not need to be extensive in chain length. While phenomenological models clearly highlighted a possible role of extensive energy gap, it was not entirely convincing at the time. The issue that concerned many researchers at the moment was that it was not clear how large energy gaps can be achieved in a realistic evolutionary scenario where sequences are allowed to vary in evolution but not physical interactions between aminoacids. “Maximal consistency” Go models essentially posited that interactions between aminoacids depend on whether they are neighbors in their native states or not. Such postulate is not entirely physical - e.g. interaction between, say, Valine and Tryptophane is the same in any conformation regardless of whether these two aminoacids are neighbors in the native state of the protein or not. “The minimal frustration” model of Bryngelson and Wolynes postulated that each aminoacid has a special “native” conformation[6] but it is not very clear how that may come about physically: the same aminoacid can have different native conformations in different proteins and it is also hard to imagine that aminoacids keep memory of their native conformations in any other conformation: when proteins are synthesized: Aminoacids do not “know” what their native conformation would be. Equally it is hard to imagine, on physical grounds, that set of states of a protein can feature a multitude of non-native, liquid-like states and just one, single native conformation as was assumed in [31]. For such an idealized density of states the first-order folding transition emerges by construction: As temperature decreases a protein has no other “choice” rather than to make a discrete jump to the postulated single native state. However in reality a protein's density of states is not discrete with a single native conformation at the bottom and a gap devoid of any conformations in between, but a continuous plethora of states varying from very native-like to totally dissimilar to native. (Reminder: the gap is defined as energy difference between the native conformation and lowest energy *structurally dissimilar* conformation). So, in reality, it becomes much less obvious if transition to the native state is the first-order one even if a sequence has an extensive gap. In fact this issue can be resolved only in microscopic, not phenomenological studies. As noted earlier, a microscopic study for Go-model like interactions indeed shows first-order-like transition to the native state [30]; however, such transition is the first order one only for 3-dimensional Go-heteropolymers. In lower-dimensional case $d \le 2$ even an extensive gap does not guarantee a cooperative behavior – because the set of partly folded states is organized in 2d-heteroplymer differently than in 3d-heteropolymer[21]. This fact also calls for caution in interpreting results of folding simulations of square lattice models.

Essentially phenomenological models such as [6] postulate some “ends” (e.g. cooperative transitions). However, they are completely agnostic of “means”, namely physical evolutionary mechanism by which extensive gaps, giving rise to such transitions, can be achieved *by*

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

sequence selection in evolution, even in principle. It is this conceptual difficulty of phenomenological “minimal frustration” and Go models that caused some skepticism about them and, by implication, about the concept of *extensive* energy gap at the time. (While the key role of energy gap in providing protein-like stability to the native state was clearly stated in [25], the analysis in [25] did not require extensivity of the gap) This fundamental issue was resolved in our work in 93 and 94 [27,36-38] where we showed that extensive gaps can be achieved by *sequence selection* alone within entirely physical microscopic model with physically realistic Hamiltonian. (see below for more details). The theoretical development [27,36,38] reconciled microscopic evolutionary models with phenomenological approach of Go and coworkers and Wolynes and coworkers, providing finally a coherent view on necessary and sufficient evolutionary requirements for polypeptide sequences to be protein-like.

(On a more technical note, in phenomenological models [3931] the gap is defined as energy difference between native state and *average* energy of the misfolded, “liquid-like” conformations - ΔE in Eq.2 of [31]. This definition differs from [25,27,40,41] where gap is defined as energy difference between the native state and *lowest energy misfold* that is structurally dissimilar to the native state. While one definition is related to the other by a simple additive sequence-independent parameter there is also a technical difference between the two: The parameter ΔE playing the role of gap in [31] is extensive in protein length even for random sequences (.which can be estimated if one sets $T_f = T_g$ in Eq.2 of [31]) while according to definition of Shakhnovich and coworkers such extensive gap exists only for special evolutionary selected sequences, while for random sequences it is ~few kT per molecule and does not grow with molecule size). However this difference is purely technical, perhaps even terminological. In fact, as we noted before, the “Tf/Tg” criterion of Wolynes and coworkers [31] is essentially equivalent to the requirement of extensivity of the gap.

Detailed simulations of simple lattice models showed the importance of gap as the main determinant of protein-like behavior, both thermodynamically and kinetically [33,40,41]. In the study of Sali et al [40] 200 random 27-mer sequences were generated and their folding was simulated using Monte-Carlo dynamics. The advantage of the 27-mer lattice model is that all its compact conformations can be enumerated [42,43] so that the ground (native) state can be known exactly if energy function is such that native states are guaranteed to be compact. In addition the availability of exhaustive conformational set made it possible to rigorously estimate energy gap. It was shown that sequences with large gaps are the ones that exhibited fast folding to the native conformation.[40]. This result was further confirmed and extended in a subsequent study [33] where different folding criteria were compared. The findings in [40] showed that large gap is *necessary* to provide fast folding. However this study was limited to one chain length – 27 residues – and it could not address the question of whether gap should be extensive or not.

Perhaps the most conclusive demonstration that energy gap is necessary and sufficient for cooperative and fast folding was obtained in computer experiments where stochastic sequence design procedure generated sequences with large gaps and it was shown that such sequences do indeed fold cooperatively and fast to their native conformations [3744] (see below, Ch**3**)..

Microscopic analytical replica theory of heteropolymers with *evolutionary selected* sequences was developed in [38,45,46]. The key idea is that now averaging of free energy in eq.(2.3) should be over the ensemble of evolutionary selected sequences. Technically that means that probability to find a sequence P in eq.(2.3) should now be properly biased towards correct, sequence ensemble, namely selected sequences that have large (and extensive) energy gap between their native conformation and collection of misfolds. A direct, (yet impractical) way to achieve this is to consider only sequences that fold with certain (very low) energy E into their native conformation, i.e.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$P_E(\{\sigma_i\})=\delta(E-\sum_{i<j}B(\sigma_i,\sigma_j)U(r_i^0-r_j^0)) \tag{2.14}$$

where δ is Dirac's delta-function that limits ensemble only to sequences that have energy E in their native conformation, and $\{r^0\}$ represents set of atomic coordinates of native structure for which sequences have been selected. (Technically eq.(2.14) biases sequences to have low energy in their native state, not large gaps. However as was shown in [27,36,38] and will be argued later, under certain conditions low native energy translates into large gap). Averaging free energy with a biased sequence ensemble Eq.(2.14) corresponds to consideration of only special sequences that are selected to fold into their lowest energy structure with significant energy gap,

However, practical calculations with sequence ensemble Eq. (2.14) are not feasible. One approach based on mean-field approximation that presents $P(\{\sigma\})$ as a product of single-site residue probabilities was proposed by Saven [47]. in the context of combinatorial protein design.

Another approach is to use a canonical distribution instead of Eq,(2.14). It was pointed out in [27,36] that sequence probability distribution given by Eq.(2.14) is equivalent to microcanonical sequence space ensemble in statistical mechanics. As usual, it is more convenient to deal with canonical ensemble, i.e. instead of a rigid requirement that all sequences have given (low) energy E in their native conformation (Eq.(2.14) impose a less restrictive and perhaps more biologically realistic requirement that ensemble of protein sequences is biased by evolutionary selection towards protein-like sequences, having low enough energy in the native state but this bias is not absolutely restrictive. Such a bias was introduced in [27,36,38,45] in the form:

$$P_{T_{\text{sel}}}(\{\sigma_i\})=\exp\left(-\frac{H\left(\{\sigma\},\{r^0\}\right)}{T_{\text{sel}}}\right)=\exp\left(-\frac{\sum_{i<j}B(\sigma_i,\sigma_j)U(r_i^0-r_j^0)}{T_{\text{sel}}}\right) \tag{2.15}$$

where T_{sel} is “selective temperature” that represents the degree of evolutionary selection on protein sequences (lower T_{sel} corresponds to stronger pressure). An extended analysis of thermodynamics of designed protein-like sequences was carried out by Wilder and Shakhnovich [45] It differed from an initial analysis of Ramanathan and Shakhnovich [38] and that of Pande et al [46] in that it extended the consideration beyond pure mean-field analysis taking into account fluctuations in order parameters (in one-loop approximation) as well as possibility of a two-step replica symmetry breaking (RSB) in the overlap order parameter.(Replica symmetry breaking (RSB) corresponds to equilibrium solutions where the overlap order parameter $q_{\alpha\beta}$ depends on replica indices” α and β. The physical meaning of RSB is that different “replicas” – conformations of the chain in deep energy minima – have different structural overlaps, which in turn reports on the complex structure of energy landscape. The specific nature of RSB is an indicator of the structure of the energy landscape in the model [48]). It was established in [45] that one-step RSB is still a stable solution for the problem and new phase diagram for the model was presented. It differs slightly from the original one proposed in 1994 [38] due to a more accurate approximation, however qualitatively it is similar to the earlier version [38] and also predicts cooperative, first-order phase transition between the native and disordered states for designed sequences and absence of cooperative transition for random sequences.

The phase diagram of protein-like heteropolymers in variables (T_{sel},T) is shown in Fig,1..

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

A major insight from evolutionary heteropolymer theory is that. random sequences can be stable at low enough temperature in their lowest energy ("native") conformations. However the transition to such "folded" states appears to be gradual, with numerous intermediate metastable states.[5]. This prediction from theory was tested by Goldberg and coworkers in an elegant experimental study [49]. These authors isolated a 101-residue fragment beta-2-subunit of Escherichia coli tryptophan synthase (ECTS). In the intact ECTS the fragment makes most of its interactions with the rest of the protein so that isolated fragment can be viewed as an essentially random sequence. The fragment forms compact conformation with some secondary structure but does not fold cooperatively as revealed by calorimetric van't Hoff criterion [50].

2.3 How many aminoacid types are needed to design a protein?

Computer experiments [37] and theory [38,45,46] showed that it is indeed possible to select sequences that exhibit protein-like behavior with large gaps. However not every heteropolymer is amenable to such evolutionary selection. Specifically there should be proper diversity of interactions to make it possible to find a sequence that has its native energy separated by a large gap from the decoys. Diversity of interactions is achieved when aminoacid alphabet is diverse. In particular it was pointed out in [37,45] that, under certain conditions, no sequences may exist for proteins having only two types of aminoacids (i.e. hydrophobic and polar, as in the HP model [51]) that could stabilize unique native conformations. The inadequacy of two-letter heteropolymers was also noted in [39] and directly confirmed in a lattice model study [52]. A mean field analysis based on application of the Random Energy Model [53] showed that two factors play a role in determining whether a polypeptide chain can have an energy gap. One is the diversity of interactions that is determined by the diversity of aminoacid alphabet, i.e. the number of aminoacid types. Another factor is chain flexibility, reflected in total number of its conformation. In particular, if a polypeptide chain has the total number of residues N and the number of conformations *per residue* is γ then the total number of conformations is

$$M=\gamma^N \tag{2.16}$$

The analysis presented in [53] showed that the necessary condition for protein-like sequences (that have large gap) to exist should be:

$$m_{\text{eff}}>\gamma \tag{2.17}$$

where

$$m_{\text{eff}}=\exp\left(-\sum_{i=1}^{20}p_i\ln p_i\right) \tag{2.18}$$

is "effective" number of aminoacid types (corrected from naïve number 20 to account for possible disparities in their compositions p_i). The effective estimated maximal gap for best designed sequences:

$$G_{\max}=N\ln\frac{m_{\text{eff}}}{\gamma}\left(2B^2\right)^{1/2} \tag{2.19}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Where B is a standard deviation of interaction energies between aminoacids. The importance of chain flexibility parameter γ can be easily understood because greater γ give rise to a greater size of conformational space of misfolds (or, “decoys”) (see eq.2.16). In turn, greater number of decoys makes it more probable that some of them have low enough energy to close the gap between decoys and the native state. This analysis suggests that making the polypeptide more rigid by introducing local interactions (most prominent of them are of course hydrogen bonds) leads to improved energy gaps and as a result, improved ability to fold. This conclusion is in agreement with results of recent all-atom simulations [54] which showed that neglect of hydrogen bonding potential results in deterioration of discriminating ability of all-atom two-body potential (see below Ch **5** for more details).

Kaya and Chan [55] tested many predictions of theory in a careful and comprehensive computer experiment. They studied cooperativity of folding transition in several popular lattice models: 2-letter 27-mer model of Shakhnovich and Gutin [27], 3-letter 27-mer model of Socci et al [56], 20-letter 36-mer model of Gutin et al [57], a 48-mer Go model [58], “solvation” 2-letter HP model [59] and short 20-letter model with side-chains of Thirumalai et al [60]. Kaya and Chan applied rigorous experimental van't-Hoff criterion to determine cooperativity of the folding transitions in these models [50]. In complete harmony with theoretical predictions they found that Go model (essentially infinite number of letters) and 20-letter models are most cooperative while short chain models as well as 2- and 3-letter models are much less cooperative, consistent withy theoretical predictions [37]. Further Kaya and Chan found that 2-dimensional lattice model proteins do not fold cooperatively. Again, this finding is consistent with heteropolymer theory [7] which predicts that 2- and 3-dimensional heteropolymers exhibit very different behavior (see above and [20]).

2.4. How important is native structure for protein cooperativity? Structural determinant of “downhill folding”

So far we focused on sequence selection aspect of protein cooperativity. However equally important is a structural aspect of the problem – how does folding cooperativity depend on the native structure of a protein? This question was first addressed by Go and Taketomi who studied simple two-dimensional lattice model with Go-type interactions [61]. These authors studied the relative role of short- and long-range (along the sequence) interactions and concluded that long-range interactions are essential for cooperativity while short-range interactions accelerate the folding and unfolding transitions. The implication from this study is that folding into structures with less long-range interactions will be less cooperative Govindarajan and Goldstein [62] conducted a detailed study of the effect of native conformation on sequence optimizability, i.e. existence of sequences with large enough gaps. Consistent with Go and Taketomi they found that prevalence of local interactions in a native structure makes it more difficult to find optimized sequences for them. In their analysis they used the T_f / T_g criterion and found that its value deteriorates for sequences that fold into structures with more local contacts. Based on the assumption that T_f / T_g serves as a predictor of how fast sequence can fold they concluded that folding will be slow into structures with many local contacts. Abkevich et al [63] addressed this question by designing sequences for three native structures of lattice 36-mer. One structure was chosen to have predominantly local contacts, another structure was selected to have almost exclusively non-local contacts and the third structure was picked randomly and had both non-local and local contacts in some average proportion. Consistent with earlier conclusion the cooperativity dramatically depended on the proportion of local conacts. In fact the structure with local contacts only did not fold cooperatively at all despite sequence design aimed at providing large gaps! Rather it folded in a continuous manner akin to the second-order rather than to the first-order transition. The structure with predominantly non-local contacts was very cooperative. The analysis of folding kinetics for these three structures revealed a more complex picture than suggested by both Taketomi and Go and Govindarajan and Goldstein. It turned

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

out that at respective temperatures when folding is fastest the sequences whose native structure had mostly local contacts folded faster than sequences that had their native states in other two structures consistent with Taketomi and Go prediction. However, at the condition when native state is stable, folding was fastest into the structure with most non-local contacts – more in line with Govindarajan and Goldstein view. This is perhaps not surprising. The cooperative transition occurs in narrow temperature range so that even slightly below T_f the protein may be already stable. When transition is not cooperative it requires much lower temperature to stabilize the protein resulting not surprisingly in slow folding at the condition when native state is stable.

The interest in the criteria of protein cooperativity was revived recently when Munoz and coworkers found a protein, BBL that exhibited thermodynamically non-cooperative behavior [64]. Based on this observation the authors posited that this protein should also exhibit non-cooperative kinetics, i.e. downhill folding. Downhill folding was also observed for other, mostly redesigned, proteins [65,66] Most recently, Zuo and coauthors [67] analyzed possible structural determinants of folding cooperativity of several proteins. They found that fraction of non-local contacts is an excellent predictor of cooperativiy or lack thereof: proteins with fraction of non-local contacts below certain threshold all exhibited non-cooperative, or downhill, folding. This analysis fully confirms earlier theoretical predictions [61,63].

3. Protein design – practical and evolutionary aspects

3.1 Stochastic algorithms to design sequences with large energy gaps

The idea to select folding (large gap) sequences from the canonical ensemble (Eq.(2.8)) immediately suggested a *practical* approach to find such sequences. Indeed any stochastic search in sequence space that converges to canonical distribution will do the job. Such method was first developed in [27,36] – Monte Carlo in sequence space. One issue that needs to be addressed in such search is that it can converge to homopolymeric sequences composed of residues that attract each other most strongly. Indeed such solution will certainly lead to low energy in the native conformation, but it is flawed. The reason is that in fact energy gap between the native state and set of misfolds needs to be maximized, not just energy of the native state. The simplest (albeit not necessarily most optimal or most realistic, from evolutionary standpoint) solution to that problem was proposed in [27]: to run stochastic Monte-Carlo search in sequence space to minimize energy of the native state *under constraint of constant aminoacid composition*. This idea appeared successful in preventing the convergence to homopolymer sequences providing sequences with optimized energy gaps. The reason why such approach is successful was explained in [27]. The low energy boundary of conformations in the misfolded set depends primarily on aminoacid composition. At the same time energy of the native conformation for which search in sequence space is carried out depends on sequence. Therefore minimization of energy of the native conformation while keeping aminoacid composition constant provided a simple way to maximize the energy gap.

This approach to sequence design while being conceptually simplest is perhaps not the optimal because it, by construction, is not able to find also an optimal aminoacid composition. Besides that, there is no condition of constant aminoacid composition for natural proteins: compositions vary between organisms and between proteins in genomes [68]. Several improvements were suggested. First, as a proxy of energy gap, Z-score in the native conformation [69]:

$$Z(\{\sigma\})=\frac{E_{\mathrm{NAT}}\ (\{\sigma\})-E_{\mathrm{av}}\ (\{\sigma\})}{D_E\ (\{\sigma\})} \tag{3.1}$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

can be optimized in sequence space. Here E_{NAT} ({σ}) is energy of sequence {σ} in the native ("target") conformation, E_{av} ({σ}) and D_E ({σ}) are average energy and its dispersion (over all M conformations) of sequence {σ} :

$$E_{\mathrm{av}}=\frac{\sum_{\mathrm{conf}} E(\{\sigma\},\mathrm{conf})}{M}; \qquad D_E=\frac{\left(\sum_{\mathrm{conf}} (E(\{\sigma\},\mathrm{conf})-E_{\mathrm{av}})^2\right)^{1/2}}{M^{1/2}} \tag{3.2}$$

Apparently homopolymeric solutions do not optimize Z-score - rather Z=0 for homopolymers because in this case $E_{NAT}= E_{av}$. Z-score optimization of sequences was first developed in [63] for lattice model proteins and was further extended to real proteins in [70,71]. In particular, Takada and coworkers designed novel sequences for a known protein having three-helix bundle structure [71] using the Z-score optimization as well as (for comparison) energy minimization approach with given aminoacid compositions. The authors used a simplified protein representation where aminoacids were represented as spheres. Several of the designed sequences were synthesized and one of them exhibited protein-like properties: significant helical content, cooperative unfolding transition (melting) and significant chemical shifts as judged by one-dimensional ^1H. NMR. However the structure of this designed protein was not determined so it is hard to say whether this design was fully successful.

Another approach to design optimal sequences was proposed in [72] where sequences {σ} that maximize Boltzmann probability to be in the native state at a given temperature T:

$$p_{\mathrm{NAT}}\ (T)=\frac{e^{-E_{\mathrm{NAT}}\ \{\sigma\}/kT}}{\sum_{\mathrm{conf}} e^{-E(\{\sigma\},\mathrm{conf})/kT}} \tag{3.3}$$

are sought.

Exact evaluation of sum over all conformation in the partition function in the denominator of (3.3) is not feasible. Instead, an approximation based on cumulant expansion of partition function was used in [72]. This approach opens the possibility to design proteins with selected thermal properties – from mesophilic to hypethermophilic ones. It also accounts for free energy difference between folded and unfolded states (the latter is accounted for via estimate of the partition function).

Further developments of stochastic Monte-Carlo sequence design procedures followed two tracks. First, it was applied to design of model lattice proteins in [37] and real proteins (with extension to all atom model of a protein and significant development of force-fields to realistically represent protein energetics) by Kuhlman and Baker [73-75] and by Mayo and coworkers [76]. DeGrado and coworkers [77] used combinatorial design approach of Saven. In particular, Kuhlman and Baker were able to design a sequence that folds into a new fold [75]. In contrast to the work of Takada [71] they used all atom-representation of proteins, i.e. accounted for side-chain packing. Folding to the target structure was confirmed by crystallographic analysis. This remarkable result provides fundamental experimental support to the main conclusion from statistical-mechanical protein folding theory that low energy in the native state (i.e. large energy gap) is necessary and sufficient for sequence to be protein-like and foldable. Earlier this key conclusion from theory was proven in simulations [37] where sequences were designed to have large energy gap for an arbitrarily chosen target and were shown to fold into

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

that target (see Fig.2). The Kuhlman and Baker work [75] is an experimental counterpart of an earlier computer experiment [37] shown in Fig.2

3.2 Using protein design to understand protein evolution: evolutionary dynamics of protein sequences and designability of protein structures

The second direction of development and application of stochastic sequence selection methods is to consider them as simple models of natural evolution. Along these lines two important sets of results were obtained. First one can seek better understanding of evolutionary processes that result in formation of fold families, i.e. collections of sequences of various degree of homology that fold into a particular structure. Sequence family expansion under structural constraints, was explored in significant theoretical detail by Dokholyan and Shakhnovich [70]. In this work the authors developed the Z-score design method for real protein structures and used it to design sequences to fold into several common folds. They followed the temporal progression of the sequence design and sequence families that emerged. The authors found that protein sequence evolution could be understood in terms of a “free energy landscape” in sequence space. Local exploration of sequence-structure pockets (which correspond to local minima on the evolutionary landscape, see Fig.3) occurs on some timescale and represents the diffusion of orthologs and paralogs with respect to one another within this pocket. The pocket itself is defined by a key set of residues that are constrained to certain amino acids in order for that set of sequences to support folding into a given structure, a fact that results in the conservation of specific amino acids or amino acid types at certain positions within the sequence family [70]. On a separate evolutionary time scale, some sequences cross “barriers” in this landscape and seed new local minima. These local minima may be unrelated from the standpoint of sequence comparison. The new sequence pocket may be subsequently explored on a shorter timescale with certain residues constrained. Sometimes these transitions result in structures that are similar to the original structure. In this case, comparison of the two sequence pockets demonstrates that the *identity* of the conserved residues differs between the two but the structural similarity is maintained because the relative *positions* of these conserved residues do not change. In other cases structural similarity is not maintained and a brand new fold is discovered. Dokholyan and Shakhnovich explored a model of protein evolution involving several protein structures and found that those residues with low substitution rates in their model tended to have low “Conservatism of Conservatism” (CoC) entropies [70,78,79]. The CoC quantity, first introduced in [78] and further studied in [79], considers families of sequences that belong to the same fold and identifies positions that are highly conserved within families (i.e. have low sequence variance) and tend to be highly and universally conserved in the set of families of the fold (i.e. positions that have low sequence entropy in many families within the fold) [70,79].

A second direction where an analogy between protein design and sequence/structure evolution can be explored, is to provide an estimate of the number of sequences that can fold into a given protein structure [36,53]. The goal of this analysis is to address an important problem in evolutionary structural biology as to why some protein folds are more abundant than others. A proper sampling in sequence space makes it possible to estimate the number of sequences that fold into a given structure, i.e. its designability [53,80-82]. Such calculations were carried out for several proteins in [53] and for many more (using a somewhat different sequence sampling strategy and analysis) in [83]. It was found for simple models [8481] and confirmed for real proteins [83] that different protein structures may have vastly different designabilities. Then the question is what is a structural determinant of protein designability? The initial insight came from the work of Finkelstein and coauthors who used Random Energy Model to estimate designability [80]. Within this approximation the overall compactness of a structure (total number of contacts between aminoacids) determines the designability of a protein. Subsequently Wolynes addressed this question and reached similar conclusion [82]. In his study Wolynes used the

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

approach of Shakhnovich and Gutin [36] to statistical mechanics in sequence space. He obtained cumulant expansion of free energy in sequence space up to the second order and also found that designability in this approximation is determined by proteins compactness. Subsequent analysis [53,85] showed that second-order truncation of the free energy expansion is equivalent to sequence-space Random Energy Model of Finkelstein However such approximation may be limited. E.g. it predicts that all maximally compact lattice conformations are equally designable – in direct contradiction with findings of Li and coworkers [81] and Goldstein and coworkers [86]. A more detailed theory developed recently by England and Shakhnovich [85] which allowed to obtain, under certain approximations, a closed form expression for free energy and entropy in sequence space, suggested that a particular property of a protein structure, namely traces of higher powers of its contact matrix (CM). (or equivalently, $\lambda_{\max}$, maximum eigenvalue of its contact matrix) may serve as a reliable predictor of protein designability. The CM of a protein of N aminoacids is an $N \times N$ matrix whose (m,n) element is 0 if aminoacids m and n are not in contact and 1 otherwise.

The physical explanation of the correlation between traces of powers of the CM and sequence entropy (i.e. designability) follows from the fact that these traces of powers of the CM reflect topological properties of the network of contacts within the structure [87]. For example, the trace of CM^2 simply gives the total number of contacts (or equivalently the total number of two step, self-returning walks) and the trace of CM^4 gives the number of length-4 closed loops in the network of contacts in the native structure of a protein and so on. One may also note that certain closed loops of contacts allow for optimal placement of amino acids that interact very favourably. For example, if four amino acids that strongly attract each other are folded into an architecture where they all interact favourably (e.g. when placed on four corners of a square, see Fig.4) this arrangement provides a greater contribution to the stability of the overall structure than configurations in which the same four amino acids are arranged linearly or in cases where the last contact is out of the contact range (Fig.4).

Such optimal placement of a sequence fragment of several strongly interacting amino acids allows for more sequences to be stable in the structure by relaxing energy constraints *for the rest of the sequence*. Thus the structures that provide certain features, such as availability of long closed loops of interactions and higher density of contacts per residue, are expected to be able to accommodate a wider variety of different sequences. This argument is similar in spirit to the derivation of Boltzmann distribution in Statistical Mechanics[88] and is similar to the justification for the “Boltzmann device” used in the derivation of knowledge-based potentials[80,89] for the study of protein folding and prediction of ligand binding energies.

The England-Shakhnovich structural determinant of designability, $\lambda_{\max}$, was tested using standard lattice model 27-mers whose maximally compact conformations could be exhaustively enumerated. The structures with highest and lowest maximum eigenvalues of their contact matrices can be found and their designabilities can be then directly compared by calculating S(E) which is (log) of the number of sequences that can fold into a given structure with energy E. This quantity can be calculated from Monte-Carlo sampling in sequence space using the analogy between statistics of sequences and statistical mechanics of canonical ensemble (Eq.(2.15)) [53]. The comparison shown in Fig.5 indeed indicates that structures that have greater maximal eigenvalue of their contact matrices (or, similarly, higher traces of powers of contact matrices) are indeed more designable: more sequences exist that can fold into them with low energy.

The analysis of sequence entropy curves presented in Fig.5 reveals another interesting feature – that it is easier to find thermostable sequences for more designable structures than for less designable ones. Indeed sequences that have exceptionally low energy in their native states can be found only for more designable structures – the blue curve on Fig.5 ends at a higher

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

energy than the red curve. This observation suggests a possible direct implication for structural genomics: that proteomes from more thermostable organisms will be statistically enriched with more designable structures. The comparative analysis of mesophilic and thermophilic proteomes from various sources confirmed this conjecture [90,91]. This finding is very important as it provides direct connection between protein folding, structural genomics (proteomics) and evolution of thermophilic adaptation.

Further, connection between protein evolution and designability is revealed in comparison between gene families of different sizes. The idea that designability may affect the size of gene families (so that more designable proteins can accommodate more sequences i.e. have gene families of greater size) was proposed by several researchers [80,81,84]. However in the absence of a structural determinant of protein designability such proposals were hard to evaluate. Now, structural determinants of protein designability are better understood so that a direct test of the hypothesis that designability effects the size of gene family could be carried out [87]. Statistically significant correlation between size of gene family and designability of protein structures that it encodes was indeed found [87]. However this correlation is limited because other factors such as evolutionary history effect the size of a gene family [92]. Indeed when factor of age of gene family is taken into account the correlation between designability and size of a gene family becomes more pronounced. Further it was found that more ancient proteins – i.e. the ones that are shared by all kingdoms of life – are significantly more designable., Furthermore, in a recent study of thermophilic adaptation, the proteomes of ancient hyperthermophiles, e.g. *P.furiosus* were found to be much more enriched in designable structures than that of hyperthermophiles that evolved as mesophiles but later recolonized hot environment[91]. This finding suggests that evolution progressed towards discovery of less designable proteins. This result can be explained by the observation that as evolution progressed in time, search in sequence space was facilitated simply because evolution had more time to explore it. The ability to explore sequence space more thoroughly relaxed restrictions on structures for which viable sequences could be found. This trend is also consistent with observations from simulations of evolution in lattice models [93].

4. From coarse grained to all-atom studies of protein folding kinetics

4.1 Discovery of specific nucleation in simulations and experiment

Studies of simple models indeed contributed considerably to our understanding of protein folding by emphasizing its universal aspects. They helped to focus our thinking on key common milestones along protein folding pathways such as transition states, on- and off- pathway intermediates [94-97], seen as ensembles of conformations. Importantly, many of the experimental studies were directly motivated by specific predictions and questions raised in theoretical studies. With regards to folding kinetics, an important theoretical discovery of nucleation mechanism via formation of specific folding nucleus [98] was made using coarse-grained –lattice- models. As defined in [98] a nucleus is a minimal folded fragment that results in inevitable subsequent unidirectional downhill descent to the native conformation. Such defined nucleus was termed “postcritical” in [98] to emphasize that no recrossing back to the unfolded basin occurs after its formation. A related definition of the folding nucleus as defining, common, structural feature of all conformations belonging to the Transition State Ensemble corresponds to “critical” nucleus suggesting probability to fold w/out recrossing back to unfolded basin as pfold=½, not just 1 as for postcritical nucleus of Abkevich et al.[98]. As noted in the original publication, [98] thus defined nuclei are related to each other. Folding nucleus was found to be specific in lattice model simulations [98]. Specificity of the nucleus means that a well-defined obligatory small fragment of structure needs to be formed in order to guarantee fast decent to the native state. This conclusion was reached in [98] based on the analysis of folding trajectories, i.e. search for the invariant minimal set of contacts whose appearance preceded subsequent fast folding. This way a *putative* nucleus was identified. Then control simulations

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

were run to make sure that simulations starting from conformations with pre-formed nucleus indeed rapidly descended to the native state without recrossing to the unfolded basin, i.e. that formation of the nucleus *guaranteed* subsequent rapid downhill folding. A modified and extended version of this approach was introduced later by Du *at al* and is now known as p_{fold} analysis [99] (see below).

Independently Guo and Thirumalai found nucleation mechanism in a different, off-lattice model [100,101]. These authors used 46-mer continuous model has aminoacids of 3 types that adopts three-pronged β -barrel structure. Guo and Thirumalai found that in several of their Langevin-dynamics simulation runs they “observed rapid formation of native hydrophobic contacts that is immediately followed by folding to the native state” [100]. The authors found that “nucleation sites” are found near the flexible loop regions. They also note that such mechanism is observed only in fractions of runs: roughly 40% of molecules reached their native state through a well-defined marginally stable intermediate.

Dokholyan et al also studied nucleation in an off-lattice model using Dicontinuous Molecular Dynamics simulations (see below) and dynamic criterion (akin to p_{fold}) to determine the TSE [102]. These authors observed specific nucleus for a generic protein model. Subsequently similar method was applied to determine the TSE in several SH3 domains where also nucleation scenario was observed [103] and location of nucleating residues appeared in good agreement with experimental ϕ-values (see below on of ϕ-values)

In experimental studies, Fersht and coworkers pioneered protein engineering approach to determine folding nuclei defined in a similar way - as residues most involved in folding transition states. They arrived, for two-state proteins such as Chymotrypsin Inhibitor 2 at a similar conclusion about specific nucleation [104]. Fersht and coworkers characterized three key residues involved in the specific nucleus of CI2 and the same residues were independently predicted as belonging to the nucleus in theoretical analysis in [105]. Fersht analyzed the results from protein engineering [104] and lattice simulations [98] and concluded that nucleation mechanism similar to the one found in lattice simulations [98] is a very plausible universal mechanism of folding for small two-state proteins. He coined the term “Nucleation-Condensation” to emphasize the fact that nucleus consists of residues that are uniformly distributed in sequence, hence bringing them together causes chain condensation. This is in contrast with earlier proposal by Wetlaufer who envisioned a nucleation mechanism based on condensation of a few residues that are nearest neighbors along the chain [106].

4.2 Chemical reaction or phase transition? “Energy landscapes” paradigm and its alternatives

Attempts to understand protein folding kinetics on theoretical grounds are deeply rooted in analogies with other, better studied systems. Of these two most powerful and conceptually very different ones are the analogy with chemical, or, perhaps, biochemical reaction [56,107-109] and the analogy with phase transition[98,110]. The major paradigm in thinking about chemical reactions is that of a low-dimensional energy landscape. The dynamics on energy landscape for a simple molecule(s) can be either ballistic or a diffusive motion along one or very few reaction coordinates. Reaction coordinate X in simple chemical kinetics is defined as one or very few coordinates (that is a function of all Cartesian coordinates that characterize the system) such that derivative of energy function E(X) (or, for many degrees of freedom, free energy function F(X)) gives the direction of reaction and the maximum corresponds to the transition state. The concept of reaction coordinate is highly non-trivial as it provides the relationship between equilibrium properties such as E(X) or F(X) and kinetics. The transition state is a kinetic separatrix that divides the direction of the reaction from going towards products to going towards reactants. Theoretical treatment of simple chemical reactions along well-defined reaction coordinates within the frameworks of the Transition State Theory or, for diffusive

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

dynamics, Kramers theory had been very successful. Therefore the appeal to pursue the chemical reaction analogy for protein folding is in the availability of a well developed theoretical formalism that can immediately be applied to the problem at hand. However the success of theoretical treatment of chemical reactions in simple molecules hinges heavily on a mere existence and proper selection of reaction coordinates. While this problem is relatively straightforward for simple molecules it becomes formidable for complex multi-particle systems such as proteins. The obvious difficulty here is that unlike simple molecules proteins are systems with many degrees of freedom. The implication of that is twofold. First, the “raw” energy landscape view is not helpful anymore because now such landscape is extremely multidimensional and is not conducive to meaningful insights. The possibility of a meaningful low-dimensional projections of the energy landscape are contingent on existence of identifiable reaction coordinate – an extremely non-trivial and yet unresolved problem (see below) Second, that unlike simple chemical reactions, entropic contributions are comparable to energetic ones in proteins so that energy alone does not determine the direction or path of the “folding reaction”.

An attempt to overcome this difficulty has been in pursuing the idea of dimensional reduction, i.e. projection via sampling on a few effective coordinates and analyzing the free energy landscape in such reduced space. In one of the first attempts along these lines, Shakhnovich and Finkelstein [35,111] (SF) introduced a simple “reaction coordinate” – volume of the whole molecule and developed an analytical model for free energy function F(V) under set of conditions such as assumption of affine deformation of the molecule. The SF theory took into account such factors as side-chain entropy and solvation in the discrete water molecule representation. The maximum in the F(V) profile curve was identified by SF [35] as the Transition State. It was noted that the folding barrier is entropic from unfolded to folded state and energetic as seen from the folded state and that the physical nature of the barrier is in the partial fixation of side chain uncompensated by proper decrease of energy and desolvation. Subsequent studies addressed the issue of desolvation of the protein core upon folding transition in more detail in simulations [112,113]. This Shakhnovich-Finkelstein theory [35] was viewed at that time as describing first-order like phase transition from the molten globule to the native state which was perceived by us at that time (with available experimental data at hand) [114] as the main cooperative transition upon protein folding. A subsequent study by Boczko and Brooks [115] used the same reaction coordinate – total volume of the molecule – but applied sampling and histogram technique with conformational clustering to determine the free energy profile F(V) and putative transition state for a small three-helix bundle.

The SF reaction coordinate – volume of the molecule - is limited in its ability to identify the actual folding transition – formation and thermodynamic dominance of a unique backbone conformation. To this end other reaction coordinates (order parameters) were proposed. Bryngelson and Wolynes used ρ - the fraction of aminoacids in their native conformation as an order parameter to measure the degree of folding [6]. Motivated by analytical theory of heteropolymers [5], Shakhnovich and Karplus (SK) introduced in a series of papers [41,116] two order parameters as a candidate reaction coordinates. One is the total number of *any* contacts between aminoacids –a parameter similar to total volume of the molecule. It reports on overall compaction of the molecule regardless whether it is folding to the native state or just a collapse to any of misfolded compact conformations. Another, much more specific and important reaction coordinate introduced by SK is Q, which is the fraction of *native* contacts in a conformation. This parameter is defined as:

$$Q=\frac{N_{\text{native}}}{N_{\text{total}}}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where N_{native} is the number of contacts in a conformation that are also present in the native state, and N_{total} is the total number of contacts in the native state. At present the SK reaction coordinate Q appears standard in most publications using the “chemical reaction” protein folding analogy [3,56,117,118]. The “free energy landscape” for Q, i.e. F(Q) was first obtained for lattice model via thermodynamic sampling by Sali et al [41]. These authors introduced a version of the histogram method that provided the density of states as a function of energy E and Q from equilibrium sampling. Sali *et al* derived density of states for the protein model based on a straightforward observation that native state in this model is unique i.e. that density of states at native energy is strictly 1. Sali *et al* also obtained thermally averaged energy as a function of Q and entropy as a function of Q. They identified conformations at $Q = Q^*$ where $F(Q^*)$ is at maximum as transition states and estimated their number from the same histogram technique. A generalization of this approach to more than one order parameter was proposed by Dinner and coauthors [119]

The paper by Sali et al [41] caused some debate in the literature (see critique and response to it in [120]). The authors of subsequent publication [56] concurred with the criticism of Sali *et al*[41] offered by Chan [52]. Nevertheless they adopted many of the approaches first introduced by Sali *et al:* order parameter Q, the histogram approach to Q sampling and, in their Fig.5, obtained the F(Q), E(Q), S(Q) plots for a similar (but not identical) lattice 27-mer model that are virtually indistinguishable from those presented in Fig.4 of Sali et al [41]. Both Sali et al and Socci et al found, not surprisingly, that F(Q), S(Q) and E(Q) plots are very temperature dependent. F(Q) is a two-minima function corresponding to native and unfolded states and cooperative barrier-crossing between them at some temperature. E(Q) is smooth monotonic function at high temperature and is less monotonic with additional pronounced minimum at low Q corresponding to a populated low-energy misfolded state at low temperature. Further, Socci et al considered Q as a reaction coordinate for Kramers equation formalism for the F(Q) profile to study kinetics for this model. The Kramers equation based approach was further developed in [117] and reviewed in [3,118]

An alternative kinetic analogy is that of a phase transition. Since folding is a cooperative process akin to a first order phase transition, our understanding and intuition about kinetics of phase transitions (with the caveat that an intrinsically small system is considered) could provide some guidance into the folding kinetic mechanism. This analogy was recognized and exploited by Abkevich et al [98] in defining the folding nucleus as a *minimal fragment of new phase* (folded state) that inevitably (i.e. without recrossing back) converts into the folded state. Thinking along these lines helped researchers to focus on important question of whether folding nucleus is specific – i.e. whether this minimal fragment of the new phase is the same or similar in all folding events or is random and varies from folding event to folding event (but its size may need to exceed some critical value). As pointed out earlier, kinetic analysis carried out in [98] and many subsequent *kinetic* studies [99,102] of the folding transition supported the specific nucleus view as did many experiments. The phase transition view was further discussed by Pande and coworkers [110]. Finkelstein and coworkers [121-123] used the phase transition view to analyze dependence of folding kinetics on length and temperature. Putting the analysis of folding reaction firmly on the ground of established facts and theories about first-order phase transitions these authors further demystified protein folding cast in terms of Levinthal paradox

While the chemical reaction analogy organically focuses on the transition states for the folding reaction, the key in phase transition analogy is also the transition state, but, with its emphasis on entropy, it focuses on the *transition state ensemble*, (TSE) i.e. ensemble of conformations that is defined dynamically: as having probability p_{fold} ½ to fold and ½ to unfold [99]. The advantage of the “phase transition” analogy is that it gets physics right, i.e. from the beginning it recognizes the crucial role of entropy, along with energy in determining the kinetics mechanism. The difficulty is that there is no universal theory of kinetics of first-order phase

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

transitions and many aspects of it are very system-dependent so that exploiting this analogy does not bring us automatically to a satisfactory theory of folding kinetics.

Which analogy – chemical reaction or phase transition - is more helpful? While the answer to this question may seem to be subjective, reflective of an individual's scientific background (chemical reaction analogy is closer to chemists and biochemists while phase transition analogy is more natural to physicists) there is a significant difference between the two in terms of predictions that they make.

Firstly the chemical reaction analysis using SK order parameter Q as a global reaction coordinate predicts that barriers for protein folding are proportional to chain length N so that folding time scales with chain length as $\exp(\alpha N)$ [117]. The nucleation mechanism developed within phase transition analogy predicts folding time to scale as $\exp(\alpha N^{2/3})$ *at the midpoint of thermodynamic folding transition*[122]. A detailed analysis of experimental data carried out by Finkelstein [123,124] at the transition midpoints and Go model simulations by Takada [125] definitely supports the $\exp(\alpha N^{2/3})$ scaling. (a virtually indistinguishable $\exp(\alpha N^{1/2})$ scaling was proposed by Thirumalai [126,127]) At the conditions when the native state is stable the nucleation mechanism would predict that folding barrier is entropic due to loop closure entropy lost upon formation of specific nucleus [128,129] which implies much slower scaling of the folding time with chain length, as a power law N^{λ}, which was indeed observed in simulations [128] and is also not inconsistent with experiment. Thus we see that the straightforward chemical reaction approach based on Q as a reaction coordinate, fails to predict correct and physically meaningful chain length scaling of protein folding time. Why? In order to understand that let us consider a simpler problem: condensation of vapor into liquid. One can consider a natural global order parameter – reaction coordinate – which is a bulk density ρ. The "free energy landscape" $F(\rho)$ will feature two minima (liquid and vapor) with maximum at some $\rho = \rho^*$ reflective of the first order character of the condensation transition. The Kramers equation or Transition State Theory approach will identify states with ρ^* as TSE and will predict the rate of condensation as $\exp(\alpha N)$ making eventually any liquid condensation event impossible for kinetic reasons, in stark contrast with our everyday experience. The reason for such failure of reaction coordinate approach is clear: While using ρ - spatially uniform, average density - as an order parameter is fully justified to study thermodynamics of the liquid-vapor transition in the mean-field approximation, it cannot serve even as a basic approximation to study kinetics [110]. We know that transition states for condensation are qualitatively different from having uniform intermediate ρ^*. Rather it is a set of fragments of new phase (that appear due to fluctuations) – water droplets – in the sea of "old", vapor phase. However certain aspects of Transition State Theory will be applicable to calculate the rate of forming of such water droplets.

Another difference between the two predictions following from two approaches is in the nature of the transition state ensemble. Kinetic approach predicts specific nucleus for many models – from lattice models to all-atom protein simulations[98,102,130]. In contrast, the reaction coordinate approach, which identifies the TSE as set of conformations corresponding to the maximum of F(Q) curve obtained from equilibrium sampling. does not find specific nucleus for lattice model proteins [131]. Why would two different approaches give different answers to the same question about specificity of nucleus? The issue here is whether the putative TSE identified in the reaction coordinate approach is a true TSE. i.e. kinetic separatrix between folded and unfolded states having $p_{fold}=1/2$. While some authors answered affirmatively to that question for idealized Go models of proteins [108,132], there is a considerable evidence that this is not so for more realistic, sequence-based and all-atom models with transferable potentials [99,133-136]. For further analysis of the relation between geometrical properties and location of kinetic separatrix see the work of Berezhlovskii and Szabo [137]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In this author's opinion, the Kramers equation approach to the kinetics on the F(Q) landscape is very problematic. The reason for our judgment is that original Kramers equation is derived from underlying dynamics given by the Langevin equation where noise is uncorrelated with coordinate and when fluctuation-dissipation theorem holds. To the best of our knowledge no such dynamics can be formulated for Q coordinate and therefore, fundamental relations such as the one between potential and force that form the basis of Langevin dynamics and Kramers equation do not hold in that case. Therefore while formally Kramers equation can be presented for the F(Q) “landscape”, its basis for the case at hand is uncertain.

In summary, while the debate of what is the best approach to theoretically describe protein folding kinetics is ongoing, it is this authors opinion that “physical” approach based on nucleation scenario within phase transition analogy is more physically sound than “chemical” approach motivated by “energy landscape” picture of simple chemical reactions. While the latter certainly claimed some success in quantitatively reproducing folding rates, failures to get it qualitatively right (e.g. incorrect chain length scaling) perhaps diminishes the success of quantitative agreements. However in all fairness a fully satisfactory folding kinetics theory is a matter of future, not the past and we can only guess its form and source of inspiration.

4.3 Folding funnels

A note on the widely used concept of folding funnels. The term “folding funnel” was introduced by Leopold et al [138] in the framework of a conceptually novel suggestion that some native structures may be kinetically accessible while other native structures may be not. These authors studied two sequences of lattice 27-mers – one that folded into a special structure and a random sequence. The first one was able to fold in 500,000 Monte-Carlo iterations while the second one was not. Leopold et al explained this difference by lack if kinetic accessibility for the second structure. The kinetic accessibility criterion was defined in [138] as the requirement that a “folding funnel” – a set of interconvertions between maximally compact 27-mer structures - that leads to the native state – exists for a given structure. Leopold et al state that “convergent kinetic pathways or “folding funnels” guide folding to a unique, stable native conformations”. In the same vein they concluded that “we introduce the concept of “folding funnels”, a kinetic mechanism for understanding the self-organizing principle of sequence-structure relationship” Similarly several other authors view folding funnel as a kinetic concept. David Wales in his textbook [139] writes “The set of monotonic sequences that lead to a particular minimum was termed a “basin” and in this sense a “basin” is analogous to a “folding funnel” described in terms of a collection of convergent kinetic pathways…” (p.246). Similarly Ozkan and coauthors [140] present funnels as kinetic concept. These authors studied a simple 2-dimensional lattice model and concluded that “folding in this model is fast, multichannel and funnel-like in the sense that conformations are fed by higher energy conformations and pour into lower energy ones…”

The key prediction of the “folding funnel” theory of Leopold et al [138] is that some sequences cannot fold due to kinetic inaccessibility of their native structures despite the fact that they may be thermodynamically stable in them. This interesting prediction potentially suggests another selection criterion for protein structure. While the work of Leopold et al did not provide an estimate of how severe this requirement is (i.e. which fraction of 27-mer structures is kinetically inaccessible) the one example that they provided – a randomly chosen sequence whose native state was deemed kinetically inaccessible – suggested that perhaps a significant fraction, if not a majority of structures may be kinetically inaccessible and only some special ones would be accessible. (Indeed in the opposite case when majority of structures are kinetically accessible the kinetic accessibility as a selection criterion would be irrelevant). However in lattice model simulations carried out over past 15 years we and others did not encounter a single kinetically inaccessible lattice structure for 27-mer as well as for longer chains. For example, the study

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

[141] addressed the question of how folding rate depends on chain length. To that end folding into 20 randomly selected lattice structures with chain lengths in the range of 10-100 units was studied using sequence design procedure described in **Ch.3** (this work can be viewed as a "high-throughput version of the computational experiment presented in Fig.2) and no lattice structure was found to be kinetically inaccessible. Similarly the study of 200 random sequences by Sali et al showed that energy gap is a single predictor of ability of a sequence to fold regardless of its native structure [40]. Others (see e.g.[142,143]) folded numerous lattice structures using the same design-folding approach as highlighted in Fig.2 and they did not report instances when kinetically inaccessible structures were encountered. That is not to say that folding rate does not depend on the native structures at all: several researchers found and discussed such dependence [62,63,143,144]. However variation of rates between different lattice native structures was found to be within approximately an order of magnitude [143,144] – i.e. well within normal folding rate variation for natural proteins [145].

Another, perhaps more widely used (or assumed) meaning of a folding funnel is that of special properties of the energy landscape presented as energy of a protein $E(X_1, X_2\ldots)$ projected into a small set of coordinates [146,147]. In their model, Bryngelson and Wolynes presented mean energy as a function of fraction ρ of aminoacids in their native conformation [6] Sali et al [41] projected energy surface of a model protein on SK order parameter Q using sampling and histogram technique as explained above (Sali et al also presented F(Q) and S(Q) functions). In both cases the resulting effective energy depended on temperature.

The concept of folding funnel, or "funneled landscape" in this "landscape" version is a statement that such projected E(X) function is monotonic, pictorially resembling a "funnel" perhaps with some fine structure reflecting its "ruggedness" [148]. In some cases "smooth funnel" and "rugged funnel" terms are used to highlight certain intuitive aspects of E(X) function. For example the Bryngeslon and Wolynes function $E(\rho)$ is always smooth-monotonically decreasing and the Sali et al function E(Q) was perfectly monotonically decreasing or "funneled" at high enough temperature even for random sequences. This is not surprising since both functions represent potentials of mean force and their monotonic behavior follows from general thermodynamic rules.

This interpretation of folding funnel is intuitively highlighted by cartoon representation of "folding funnels" that can be found in the literature [147]. Axes are usually not labeled in cartoon representations i.e. coordinates X_1, X_2 … are not specified. However selection of coordinates to present a "folding funnel" (in its second, "landscape" interpretation) is a key issue and the results depend crucially on how coordinates for E-projection are selected. This issue is highlighted in the work of Ozkan and coauthors [140] who studied folding mechanism of simple 2-dimensional lattice 16-mer within Go model approximation of energetics. Go-models are deemed to be archetypical "smooth funnels" [108]. Indeed if energy is plotted vs. SK reaction coordinate Q (the number of native contacts) the E(Q) is a perfectly monotonic function (by definition), indeed invoking associations with a "funnel-like" landscape. However, the authors of [140] used another set of coordinates obtained from principal value decomposition of the conformational space of the 16-mer. The first two principal axes were used to create E(X_1, X_2) surface and the result is that this surface *for the same Go model* is extremely rugged, or as authors of [140] put it "Using the singular value decomposition we show an accurate representation of the shapes of the model energy landscapes. They are highly complex funnels".

So, for the same simplest 16-mer Go model a funnel can be "smooth" (if SK Q coordinate is used) or "highly complex" (which even does not visually resemble a funnel if coordinates of Ozkan et al are used). Furthermore, in a recent study [149] Krivov and Karplus show that projection of energy function on pre-selected coordinates may be grossly misleading as it conceals the true complexity of the conformational space and physics associated with that. The

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

authors state that "…the standard funnel picture of protein folding should be revisited". In the same vein Caflisch argued that projection of (free) energy landscape into a specific coordinate (in his case SK Q) can be misleading [136]. He showed, for a small peptide, that such projection groups together structurally and kinetically different conformations by mixing, for example, in the same Q-bin conformations from native, denatured and transition state ensembles [136].

Another complication is that for a complex system with many degrees of freedom free energy rather than energy determines, in principle, the folding process. In this sense the E(X) graphs may not be reflective of the folding process et al!. Entropic part of the free energy in this reduced representation comes from sampling over all degrees of freedom unconstrained by selection of projection coordinates X. This makes such "landscape funnel" plots also dependent on the temperature.

However the key issue with "landscape funnels" is that relation of "funneled" (or "non-funneled") landscapes to folding kinetics is entirely unclear as explained in the previous chapter. This is again dramatically illuminated by Ozkan and coworkers [140]. Looking at the energy landscape for their 16-mer Go model (Fig.9 of [140]) one would immediately infer a trap-dominated complex folding scenario resulting in non-exponential kinetics (relation between traps and non-exponential kinetics was rigorously established in [98]). However the actual kinetics observed is perfectly exponential and the detailed kinetic mechanism revealed by master equation approach could not have been inferred looking at the "energy landscape" for the model. The study of Ozkan et al [140] puts the utility of "energy landscape" perspective on protein folding kinetics into question primarily because energy landscapes do depend dramatically on the choice of coordinates in which the "landscape" is plotted. The coordinate of choice should be a "true reaction coordinate". (TRC) In this case free energy gradients will be indicative of direction of the folding process, as explained above but such TRC is not known and even its mere existence is a matter of debate. A candidate for TRC – SK parameter Q – advocated by some researchers [108] was shown to be inapplicable even for a relatively simple peptide with realistic transferable potential [135,136]. Therefore, unless the TRC is found, the "landscape funnels" will remain a highly arbitrary and perhaps misleading concept. On the other hand, the utility of the concept of the "kinetic folding. funnels" advocated by Leopold et al [138] hinges on the ability to define kinetic connectivities in protein models of realistic size and assumptions about dynamics of the system.

We showed in this chapter that there is a significant variance of opinion in the literature as to what "folding funnel" is. Unfortunately, until the community converges on a clear definition of the "folding funnel", the use of this term is bound to generate significant amount of unnecessary confusion.

4.4 Structural determinants of protein folding rate: contact order and its alternatives

The accumulation of experimental data stimulated the search for empirical correlations between folding rate and structural properties of proteins and some were found indeed. One of the most interesting of them is Relative Contact Order

$$\mathrm{RCO}=\frac{\frac{1}{N_c}\sum_{i<j}(j-i)}{N} \tag{4.2}$$

(where N_c is the total number of contacts between aminoacids in a protein, N is the total number of aminoacids and the sum is taken over all (properly defined) contacts between aminoacids.) which was shown to be a good predictor of folding rates for several proteins [150]. More recent

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

experimental studies found numerous exceptions from that correlation both for mutants of already studied proteins [151] and several newly studied ones [152,153] (some many orders of magnitude off predicted rate[153]). It was shown in the original publication that Relative (i.e. normalized by N) Contact Order as given by Eq. (4.2) is a good predictor of folding rate. However, in a more recent revision of the concept published by the same authors it is now argued that Absolute Contact Order (defined in the same way as eq. (4.2) but without N in the denominator) is a good predictor of folding rates. At the same time other, more simple structural determinants such as fraction of local [1] and non-local, long-range contacts [154] were argued to be equal, or better predictors of folding rate. A comparison and analysis of various predictors for a set of 18 proteins was recently made by Kuznetsov and Rackovsky [155]. These authors argued that (1) Values of the correlation between folding rate and contact order are very data set dependent: values as high as 0.81 for 12 proteins [150] or as low as 0.64 for 18 proteins [156] have been reported. (2) A highly significant correlation between log(k) and secondary structure content has been found [157] (3) Both strength and distribution of the interactions have been shown to play an important role in determining folding rates. [156] However, contact order is a purely geometric property and does not account for these factors. Further, Kuznetsov and Rackovsky showed that sequence-based determinants such as propensity to form various types of secondary structure can serve as equally good determinants of folding rate [155]. Ivankov and Finkelstein proposed a similar sequence-based predictor of folding rates also based on secondary structure propensities [158] Apparently a further objective study that takes into account all available data is needed to clarify which structure-based or sequence-based parameters (if any) can serve as a unique and most reliable predictor of folding rates.

4.5. Evolutionary traces of nucleation mechanism.- Conservatism of Conservatism analysis

An important observation was made in [98] that location of the folding nucleus in structure is conserved between many model proteins that folded into the same structure despite having very different non-homologous designed sequences. Experimental studies of nucleation in non-homologous proteins that have similar structures arrived at similar conclusion [159-161]. These results provided the basis for “structure-centric” view according to which any folding potential (including Go) that leads to folding into a given structure would provide a robust picture of the pathway including the location of the nucleus.

The observation that folding nucleus is conserved between proteins belonging to the same fold has an interesting possible evolutionary implication. Indeed, if one assumes that evolutionary pressure was exerted to control folding rates (e.g. to prevent protein aggregation to happen before proteins fold) then folding nucleus residues, being “accelerator pedals” for folding are under universally stronger selective pressure in all proteins of the same fold (but not necessarily the same function). This hypothesis suggests an approach to detect folding nuclei from bioinformatics analysis[78,79]. The issue here is that residues in proteins may be conserved for various reasons – their importance for stability, function, interaction with other proteins and perhaps their role in folding kinetics. How to distinguish between these different factors? The insight comes from two observations: First, proteins having similar structures but very different sequences and functions still may have similarly located folding nuclei. That allows one to rule out functional conservation by properly comparing proteins with differently located active sites/regions. Second, the conservation for stability manifests itself in a very strong correlation between residue buriedness in structure and its conservation [70,79]. Therefore residues that are *more conserved than expected from buriedness factor alone* are under additional pressure, besides stability. Thus universally conserved (in all protein families having given fold) residues that are outliers (towards higher conservation) from the buriedness-conservation correlations are good candidates to represent folding nucleus for a fold in question. However one has to be careful in estimating conservation because here comparison is made between proteins having same fold but vastly different sequences so that naïve multiple sequence alignment between

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

them is not possible. Rather, one has to determine conservation profiles within families of homologous proteins (i.e. within each minimum on Fig.3) and then, using structural alignment, compare conservation profiles to determine which positions appear to be *universally* conserved. Of course identities of universally conserved residues may vary from family to family as shown schematically on Fig.3; it is the fact of their universal conservation in corresponding structurally aligned positions (see Fig.3) that determines their possible special role as belonging to folding nucleus. The detailed analysis of this property, called Conservatism of Conservatism (CoC) in [79] provided predictions for the folding nuclei in five common folds. In some cases such as (α/β) plaits or Rossman fold (CheY) the folding nucleus was already determined from protein engineering analysis (ϕ-values) [104,162,163] and the predictions are in good agreement with experiment. In other cases, most prominently for Ig-fold proteins, the CoC analysis predicted precise locations of the nucleus residues for all proteins having that fold [79]. We noted in [79] an interesting phenomenon of “circular permutation” of aminoacids in the Ig-fold nucleus. We found that folding nucleus always contained a 100% conserved Tryptophan residue but its location in the nucleus varied from family to family as if nucleus residues were making circular permutations upon transition from one family to another. Also in some cases strong hydrophobic contacts in the nucleus observed in one family was replaced by a disulfide bond in another family. In a series of papers Clarke and coauthors studied experimentally folding nuclei in Ig-fold family of proteins [164,165] and found that indeed folding nuclei appeared conserved between different proteins of this superfamily and its location was in agreement with earlier predictions [79].

It is still a subject of a considerable debate as to whether protein folding nuclei are under additional evolutionary pressure as it is posited here. While such suggestion was made by us in [78,79,141] and was used there to successfully predict folding nuclei in several proteins, Plaxco and coauthors argued against it [166]. These authors sought correlation between ϕ-values and sequence entropy in simple multiple sequence alignment and found it for some proteins but not for others. In response Mirny and Shakhnovich [167] argued that evolutionary pressure on folding nuclei is in addition to other selection pressures such as ones for stability and function. To this end a careful CoC analysis [7879] is necessary to detect such additional pressure. A simple multiple sequence alignment used in [166] would likely fail to detect additional pressure on folding nuclei. In response to that Plaxco and coauthors, [168] while emphasizing specific nucleus scenario of protein folding essentially reiterated their original argument based on the analysis of multiple sequence alignments, making next round of rebuttals redundant.

4.6 Topology-based folding models

The RCO correlation with folding rate and related observations motivated the development of a class of highly simplified models that allowed a detailed analysis under extremely limiting set of assumptions. One of such assumptions is that a conformation of a model protein should consist of two contiguous “native” parts separated by no more than one disordered fragment. [156,169,170]. Nevertheless analysis of the putative transition states (identified as maxima of low-dimensional free energy projections) in such models revealed some consistency with reality as found in comparison of “predicted” ϕ-values with experimental ones. Overall it is sometimes difficult to judge the measure of success of these analyses because in many cases the actual residue-by residue predictions of ϕ-values were not reported. Another important control that needs to be done is whether predicted correlation is much better than trivial null models such as correlation between ϕ-values and buriedness of an aminoacid in the structure, or the number of contacts that an aminoacid makes in the native conformation.

This line of research was extended by Plaxco and coworkers [171,172] who proposed the so-called “topomer search model” (TSM). A basic assumption of the TSM is that the rate-limiting step

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in folding is an essentially unbiased, diffusive search for a conformational state called the native topomer defined by an overall native-like topological pattern.

A comprehensive analysis of feasibility of the TSM was presented in a recent work by Wallin and Chan [173] These authors examined key conclusions of the TSM using extensive Langevin dynamics simulations of continuum C_α chain models. A careful determination of the probabilities that the native topomers are populated during a random search, as TSM posits, apparently fails to reproduce the folding rates predicted by the TSM, with discrepancy reaching for some proteins up to 70 orders of magnitude. Not surprisingly, simulations in [173] indicate that an unbiased TSM search for the native topomer amounts to a Levinthal-like process that would take an impossibly long average time to complete. Furthermore, Wallin and Chen argued that intra-protein contacts in all native topomers (which are predicted to be Transition States in the TSM) exhibit no apparent correlation with the experimental ϕ-values for these proteins.

This analysis of Wallin and Chan teaches us several important methodological lessons. First it shows that in protein folding as in any other field of science the models must be as simple as possible but not simpler. Second, it shows that a partial success of a model, in this case phenomenological correlation between a structural parameter (in the case of the TSM - number of long range-contacts) and experimental observable – folding rate – while encouraging may not serve as a proof of validity of a model. Rather a model must be physically consistent and be consistent with *all* available data or at least if partial inconsistencies do exist the model must offer an explanation for them. While these simple recipes may seem trivial, they are not always easy to follow when such a complex process as protein folding is modeled.

On a more general note a question arises as to the utility of oversimplified topology-based models. The role of theory in protein folding is to provide insights into thermodynamic, kinetic and evolutionary mechanisms that are not directly available from experiment. The agreement with experiment is necessary to validate the model's assumptions. Validation of the model makes believable the theoretical conclusions that go beyond direct experimental observation. However in this case the models assume mechanisms that are hardly realistic, such as two stretches of native structure separated by no more than one disordered loop. Karanicolas and Brooks pointed out that such models may not provide a reliable microscopic mechanism of protein folding[174]. A question then remains as to what one learns from oversimplified models.

4.7 A brief note on experiment

On the other hand, remarkable progress has been achieved over the last several years in experimental studies of protein folding. More advanced experimental techniques were developed that allowed researchers to significantly extend the time resolution of their kinetic experiments to low microseconds, using such approaches as laser T-jump and continuous-flow [175].[176]. Single-molecule techniques are used to probe folding thermodynamics and kinetics [177-179]. These and many other experimental studies provided a much more detailed experimental view on protein folding temporal and spatial progression that either overcame or have a potential to overcome such traditional limitations as loss of information due to ensemble averaging or lack of time resolution to detect intermediates or properly evaluate burst phases. To this end the discussion between Roder's group and Baker group concerning the intermediates in protein G folding is noteworthy: while Baker's experiments using traditional stopped-flow equipment and W43 fluorescence as a single probe revealed no intermediates, [180], the use of more time-sensitive continuous-flow apparatus made it possible to discern a major on-pathway folding intermediates [97]. Furthermore, a careful analysis of chevron plots for several proteins carried out recently by Kiefhaber and coworkers revealed slight yet noticeable curvature in the unfolding branch which can serve as evidence of transient intermediates or multiple transition states as well as possible effect of mutations on unfolded state [181,182]. Similarly, Clarke and coworkers analyzed non-linearity of chevron plots in several

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Ig-fold proteins and concluded that its most likely origin is in the existence of parallel folding pathways passing through distinct transition states and that denaturant may shift the dominant pathway[183]. Radford's group work on helical Bacterial Immunity proteins also revealed complex pathways, including intermediates stabilized by non-native interactions in some of them and the possibility to change the complexity of a folding pathway via mutations[184] Further insights into detailed picture of protein folding landscape can be obtained from AFM pulling experiments [185,186] pulling that probe free energy profile along complementary reaction coordinates. In a recent work, Marqusee and Bustamante used optical tweezers to induce complete mechanical unfolding and refolding of RnaseH [187]. A great advantage of optical tweezers over AFM is that they allow a much slower rate of pulling making experimental conditions closer to equilibrium. That allows to better relate single-molecule results to bulk experiments and simulations opening an exciting possibility to observe experimentally transitions in single molecules that so far could be seen only in simulations.

4.8 Towards microscopic description of the Transition State Ensemble

This brief and by no means complete account of recent experimental work in protein folding nevertheless illustrates impressive advances that provide detailed structural information about many aspects of folding mechanisms. *However, at this point experiment probably reaches the limit of its ability to provide structural insights without simulations*. This calls for very accurate computational models that match precision of experimental information and allow unambiguous structural interpretation of experimental data. Of special importance is structural characterization of Transition State Ensembles – a turning point (dynamic separatrix, see above and [99,137,188,189]) on the free energy landscape from which a protein is committed to fold.

Structural description of the TSE is impossible without simulations because it corresponds to an unstable state whose experimental detection is very difficult. While ingenious experimental approaches based on protein engineering methods provide extremely valuable information about possible interactions in the TSE [104,162] a structural model of the TSE can be obtained only from high-resolution simulations. However full folding simulations to determine the TSE *ab initio* are difficult for many proteins (see however [130,190,191]). To this end approaches that incorporate experimental data such as ϕ-values, into simulations have been proposed by several groups.

The Daggett group employed unfolding simulations analysis based on the premise that unfolding is the microscopic reverse of folding, This assumption was questioned by several authors [192,193] who showed that unfolding when simulated at different conditions than those of normal folding experiments may not represent direct inverse of folding. Such difference in simulation conditions may results in significant differences in observed pathways. Nevertheless Daggett and coworkers found that their proposed models of transition states are consistent with experimental ϕ-values and in some cases they were able to predict mutations that significantly affect the folding rate in some proteins [194].

4.9 Insights from simulations of all-atom Go-model proteins

While successes of some of all-atom simulations are encouraging [195] they are still limited to very short proteins or peptides, in some cases study unfolding rather than folding and sometimes rely on very small number (less than 10) trajectories. At an intermediate level of complexity Go models of various degree of detail proved useful. As we said earlier, in Go model only interactions between groups [133,196] or atoms [130,197198] that are neighbors in the native state are treated as attractive in any conformation. The benefit of such models is that they “solve” folding potential problem by guaranteeing that the correct native state is a global energy minimum. Their obvious shortcoming is that knowledge of native structure is needed in order to build such potentials and also they may underestimate non-native interactions in

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

some cases [199]. However in many cases they are the only potentials that allow full folding simulations from random coil to native state and as such provide extremely detailed insights into folding mechanisms for model proteins. Following this route we developed a novel and powerful tool – all atom Monte-Carlo dynamic simulations [197]. The method takes into account all heavy atoms of the protein and uses a move set consisting of a combination of local and non-local moves. Calibration of the move set appeared to be a major undertaking that included comparison with dynamics of short peptides undergoing helix-coil transition and comparison of rates observed in simulations (in terms of number of Monte-Carlo steps) and experiments where such data are available. That included the second beta-hairpin from protein G whose folding rate is known from experiments by Eaton and coworkers [200] as well as α-helices [201] and several small proteins. In all cases the observed folding rates were highly linearly correlated with experimental ones and the results on dynamics of helix coil transitions were consistent with MD simulations data and experiment [201,202203]. These results provided sufficient evidence that the developed technique is accurate enough to be useful for modeling folding mechanisms of small proteins, and we embarked on the studies of protein folding atomic level of detail. The first protein that we studied, Crambin, was mostly a proof-of principle study that showed that using atomic potentials that includes realistic steric interactions and contact Go atom-atom potentials we obtained numerous successful folding trajectories, for real proteins, at the atomic level of detail, using available computational resources Nevertheless even this first study provided strong evidence about complexity of folding pathways and relative role of energetic factors, backbone and side-chain geometries in defining folding pathways.

The next major undertaking in this direction was to simulate complete folding of a protein that has been well-characterized in experiment. Ig-binding domain of staphylococcal protein G[130] This protein appeared to be an ideal model as it is relatively short and is relatively fast-folding (3-5 ms) and there is a plethora of experimental data to compare with. [97,180,204]. This project presented us with numerous challenges including the need to carefully calibrate short-range potentials (mostly H-bond) relative to long-range Go-energetics. This was accomplished by setting the strength of *non-specific* backbone hydrogen bonds to comply with thermodynamic data on stability of *isolated* elements of protein G secondary structure.

The simulation [130] revealed a complex picture of protein G folding that entails parallel pathways converging to common Transition State Ensemble (Fig.6). The transition state ensemble contains specific nucleus of six hydrophobic residues, consistent with general picture of nucleation mechanism and consistent with available ϕ-values (see below),

This study taught us several lessons, most important of which are that ensemble averaging (as is done in most experiments) and selection of experimental probe/reaction coordinate (e.g. W43 fluorescence) may significantly affect apparent picture towards sometimes misleading conclusions. It emphasizes a crucial role that simulations must play in interpreting experiment – "Only theory decides what we manage to observe".(A.Einstein)

4.10 Using experimental constraints to obtain folding nucleus at atomic resolution

The results of the most structurally informative protein engineering method [162] are often "visually" interpreted as "high ϕ-value residues belong to the nucleus, while low ϕ-value ones do not". Such reasoning being qualitatively acceptable in some cases, sometimes misleads, E.g. I76 in Chymotripsin Inhibitor 2 (CI2) shows low ϕ-value in many mutations [104]. However a careful double-mutant study attributes it to folding nucleus [205]. In another example, of protein G the highest ϕ-values are observed in the turn of the second hairpin, while ϕ-values in other locations are noticeably lower [180]. While this observation points out the importance of the hairpin it is hard to imagine a TSE (i.e. set of conformations with p_{fold}=0.5) where only one hairpin is folded while the rest of the protein is not.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

The qualitative character of the "visual" interpretation of protein engineering method was noted by Fersht and Daggett [206] who insightfully pointed out that ϕ-values should be treated as experimental constraints akin to NOESY in NMR determination of protein structure. This idea was further developed by Vendruscolo [207,208] and coworkers who used ϕ-values to reconstruct *putative* transition state of acylphosphatase - one of the proteins studied by Dobson and coworkers using protein engineering methods[159]. Vendruscolo and coauthors reconstructed putative TSE for this protein using ϕ values as constraints in high-temperature unfolding simulations (using initially reduced C_α model [208] and later all-atom representation of proteins [207]). However they did not test whether the proposed conformations represent true TSE, i.e. set of conformations for which transmission coefficient to the folded state $p_{fold}=0.5$ [99].

All-atom simulations provide a unique opportunity to address this issue. First we carried out the analysis of TSE for CI2 [209] – perhaps the best characterized protein in terms of ϕ-value analysis [104]. We showed there that ϕ-values correctly specify, in general, the TSE: $<p_{fold}>$ over putative TSE appeared to be close to 0.5. The work presented in [209] was like a "proof of principle" both for p_{fold} calculations and ϕ-value analysis. Our subsequent study [188] presented a much more detailed picture of the TSE for protein G folding. In particular it clarified a number of key issues related to the ϕ-value analysis:

A. What is the minimal number of ϕ-value constraints to enable reliable reconstruction of TSE?

B. What is the relation between ϕ-values of residues reported in various mutations and its role in forming the TSE?

The all-atom simulation of protein G[188] provide some answers to these questions for that protein. In particular it was shown that upon gradual addition of ϕ-value constraints the $<p_{fold}>$ (<> means averaging over many starting conformations of the *putative* TSE, which are generated using constraints derived from experimental ϕ-values following the Vendruscolo approach[208]) first grows and then saturates reaching limiting values of 0.5. Most importantly, distribution of the p_{fold} values over constraint-generated *putative* TSE starting conformations is pronouncedly bimodal: many conformations are found with low and high p_{fold} and relatively few in between, with $p_{fold}=0.5$. This is perhaps not surprising because the TSE corresponds to free energy maximum, i.e. it is comprised of least stable conformations (see Fig.7). However this simple observation clearly indicates that no reliable structural characterization of the TSE without p_{fold} analysis is possible. In particular the models of transition states based only on constraints may be sometimes misleading. For example an unverified model of the TSE for SH3 domains (based on constraints only) posits that the TSE for these proteins has native-like topology and is structurally close to native state for all three SH3 domains studied. [183]. However a careful analysis of SH3 TSE that includes p_{fold} verification presents a completely different picture: of highly polarized TSE with well-defined small nucleus but with significant part of the chain disordered almost as much as in the unfolded state. [98,133,210]. It should be noted that folding nucleus in SH3 domains (as well as in other studied proteins [188,211]) is "diffuse" *in sequence:* it is comprised of residues that are uniformly distributed throughout the sequence. However the residues belonging to folding nucleus are well packed in *space* in the TSE conformations. This is very clear from the pfold-based analysis of contact maps in pre-TS ($p_{fold}< 0.5$) conformational ensemble, TSE ($p_{fold}=0.5$) and post-TS ($p_{fold}>0.5$) conformational ensemble. Contact maps are constructed to show contacts that are most probable in corresponding ensembles. Of special importance are differential contact maps between TSE and pre-TS ensembles. (Fig.8) Apparently such differential contact map shows only contacts that are most important for TSE: without them the TSE is not reached. These are the contacts that are necessary to form in folding TSE, i.e. *nucleation contacts*. The analysis of nucleation in SH3 domains reveals important necessary structural feature primarily central β-sheet consisting of strands 2-4. It is a necessary feature because it is always present in all TSE

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

conformations. However it is not sufficient to form this *β* strand to reach TSE. Indeed the same *β*-sheet is formed in pre-TS ensemble. In other words formation of the central β-sheet is very important but it does not guarantee that the TSE is reached. What does? The answer to this question comes from the analysis of differential contact map between the pre-TS ensemble and the TSE which points out to contacts that are key to TSE i.e. that appear only in TSE but not before. The analysis of differential contact maps revealed that a few key contacts are crucial for the TSE. (Fig.8) These contacts are between residues that are spread all over the sequence but form a tight cluster in structure. These residues constitute *folding nucleus* for SH3 folding: its formation is key to reaching the TSE. Also this set of contacts, corresponding to folding nucleus corresponds to common, invariant feature among all TSE conformations.

4.11 Sequence or Structure? Insights from High-Resolution Simulations

One of the most debated issues in protein folding is what determines folding pathways: final structure or protein sequence. While this question may sound somewhat scholastic (since sequence always determines final structure) it is not: there are many proteins that have similar structure but very different sequences and the relevant question is whether such proteins have similar or different folding mechanisms. This question has a long history. An early indication that structure may be a more robust determinant of the folding mechanism than sequence was made in [98]. This proposal was based on lattice model study. Subsequently several authors arrived at similar conclusions using various techniques [169,212]. (see above **4.4** for a more detailed discussion of evolutionary implications of this finding) However in some cases the apparent exceptions from the perceived robustness of folding pathway were found. For example, in small helical protein Im7 mutations changed observed pathway – from apparent two-state to three-state [213] folding mechanism. Similarly Baker and coauthors showed that structurally similar proteins G and L have different distribution of ϕ-values [214] suggesting that these two proteins may have different folding pathways However a detailed analysis based on simulations of protein G in structure-centric Go model [130] showed that certain features of folding pathway are flexible and certain features are robust. In particular, there may be many pathways leading to nucleus formation passing through various metastable intermediates. This aspect is flexible as mutations can easily shift distribution between different paths and stability of the intermediates. However all these pathways converge to a single nucleation step and the structure of the nucleus is robust in a sense that it is mostly determined by the final structure of the protein (see Fig.6). Proteins having different sequences but similar structures have very similar folding nuclei. This conclusion is supported by experimental studies. E.g. Radford and coworkers showed that despite the fact that two homologous helical proteins – Im7 and Im9 fold via two and three-state mechanism, the TSE structures of these proteins are very similar [215]. The apparent discrepancy between results for L and G proteins obtained by Baker and coworkers [180,214] can be attributed to difficulties of derivation of the TSE from “visual’ inspection of ϕ-values. Indeed, when detailed analysis using p_{fold} was carried out for protein G [188] (using experimental constraints and Go model simulations) its folding nucleus appears to consist of several tightly packed hydrophobic residues (consistent with other proteins such as S6[211], SH3[161,216], CI2[104] etc) rather than a β-turn as one would naively expect based on visual inspection of ϕ-values. The locus of the correctly determined nucleus appears invariant between proteins G and L. Similarly, location and composition of folding nucleus is invariant between three SH3 domains (spectrin, src and fyn) as revealed in a recent study [210]. Davidson and coauthors [217] suggest that that answer to the question “do proteins with similar structures fold via the same pathway?” is ambiguous. However our analysis based on combination of detailed high-resolution computations with experimental data gives less ambiguous answer: that *folding nucleus is a robust feature of a protein and its location is determined primarily by its final structure*. Other aspects of the folding pathway (e.g. how protein “ascends” to TSE) may be more sensitive to details of sequences and change even upon single mutations.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4.12 Discontinuous Molecular Dynamics (DMD) simulations: Domain swapping and amyloids

A complementary simulation method – Discontinuous Molecular Dynamics – was used in a number of studies to explore folding mechanisms in coarse grained models of folding[218102, 103,219-222]. This method is based on direct propagation of dynamics by solving energy and momentum conservation equations each time protein atoms interact between themselves or with "ghost" solvent particles. Several models were studied within Go model energetic prescription – from generic compact structure [102] to SH3 domain [103223] to amyloid-like aggregates [222,224]. The analysis of these simulations shows that the developed picture of specific nucleation is very robust between models and simulation techniques.

Further, a very promising model to study protein aggregation and amyloidosis [222] was developed within the DMD simulations approach. Energetics of this model is based on specific side-chain-like interactions combined with nonspecific backbone hydrogen bonding. This is a multiple chain Go model whereby the aminoacids interact following Go prescription not only for their own chain but between identical chains as well. The multichain Go model of Ding et al [222] provided an intriguing experimentally testable generic model of amyloid fibril formation. More recently the same model was used by Wolynes and coauthors for their study of dimerisation of SH3 domains with identical conclusions concerning domain swap mechanism [225] of aggregation and very similar structural model of dimers of SH3 molecules – precursors of amyloid fibrils.

In a more recent work [226] a sequence-based coarse-grained energetics model (as opposed to structure-based Go model) was developed to fold Trp-Cage miniprotein using DMD simulation technique. The authors of [226] note that suceess in folding of Trp-Cage miniprotein by this method and by atomistic MD simulations [191,227] may be attributable to specific features of folding and energetics of this miniprotein and may not necessarily be transferable to other cases.

4.13 Long-time side-chain and backbone dynamics – a glassy story

The all-atom Monte-Carlo simulations tool made it possible to address several problems that previous coarse-grained models were not able to approach due to their (over)simplified character. One of them is the issue of statistics and dynamics of side-chain packing – an aspect of protein folding that was recognized by many as a cornerstone [228,229]. The all-atom MC simulations were used to address this problem. First, a direct sampling of side-chain packing states was performed to resolve a long-standing issue[228]: how many side-chain packing arrangements are sterically compatible with a given backbone conformation? The analysis was performed for several models of sterics – from hard-shell to van-der-Waals soft-shell steric interactions with unexpected conclusion – that many (exponential in the number of side-chain degrees of freedom) conformations are compatible with a given backbone conformation[230]. Naturally this degeneracy is broken in real proteins by interactions so that native conformation of side-chains is energetically favored over alternatives (decoys). The side-chain packing decoys generated by this algorithm are used to develop atom-atom potentials for protein folding using potential optimization techniques[31,231-233].

The large conformational space of side-chains even in tightly packed state suggests that there may be a peculiar dynamics of their packing during folding. Again all-atom folding simulations proved an invaluable tool to address this difficult question – the analysis of many individual trajectories for protein G folding makes it possible to develop a very detailed picture of how side-chains get organized in folding process and the results are quite interesting. It appears that there is a broad distribution in time-scales for side-chain packing times even with apparent two-state kinetics but side-chains that constitute the nucleus are the fastest to acquire their

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

native conformation![234]. This result was obtained in [234] in simulations of a new lattice model with side-chains as well as in analysis of trajectories of all-atom simulations of protein G.

Further analysis of protein G folding trajectories revealed a complex folding scenario whereby major features of protein topology and packing of nucleus side-chains get established first concurrently with nucleation while side-chain packing of the rest of the structure occurs over longer time scale and is accompanied with backbone fluctuations (see Fig.9).

These longer time-scale fluctuations appear to be of a peculiar character resembling glass transition dynamics with its signature power law relaxation of many characteristics such as total energy. A detailed analysis of such relaxation processes requires a new theoretical approach based on mode-coupling theory [235-237]. A general theoretical formalism based on mode-coupling theory applicable to homo- and heteropolymer dynamics has been developed in [237]. It was shown there that in the low temperature regime a glass transition that would feature a long-time non-exponential relaxation of energy may indeed occur. However this is only a small initial step – a comprehensive theory that would treat directly side chain relaxation in proteins is a matter of future development.

4.14 From ensemble to single molecules – pulling and stretching

The analysis of protein G folding [130] suggested (perhaps not surprisingly) that ensemble averaging in experiments may conceal important features of folding pathways. To this end single-molecule studies appear to be a very important complementary approach to elucidate folding kinetics and landscapes. A first successful single-molecule study of protein folding was published by Hochstrasser and coauthors [179] A number of interesting studies followed [177,185,187,238-240]; in most cases protein stability/unfolding was probed in mechanical AFM experiments when the molecule was mechanically stretched (with notable exception of optical tweezers approach of Bustamante and coworkers [187]).

Theoretical foundation for understanding mechanical response of proteins in single-molecule experiments starts from analytical theory of mechanical properties of random heteropolymers [99,241]. This theory predicted a regime of gradual stretching of a heteropolymer when force comes close to a critical value f_c with intermediate structures resembling bead on a string. Furthermore a phase diagram of stretching heteropolymer was presented in variables temperature – stretching force that outlined the regimes where such intermediates can be observed. This results in behavior that is quite different from that of mechanical proteins most notable titin where domains unfold in a two-state manner at or around critical force [238]. However titin is a protein selected by evolution to perform mechanical functions. When a non-mechanical protein underwent stretching it exhibited much more gradual unfolding [240] – in complete agreement with theoretical predictions because from the point of view of special mechanical properties barnase studied in [240] is not an evolutionary selected protein. An interesting and important extension of this study is to develop a theory and simulations of mechanical proteins, i.e. the ones selected by evolution to perform mechanical functions – such as titin. This effort should combine simulations in coarse-grained as well as all-atom models and bioinformatics analysis aimed at determining which residues define mechanical robustness of such proteins. Interesting simulations along these lines were reported recently [242].

5. Towards realistic transferable sequence-based potentials for protein folding and design

The all-atom Monte-Carlo algorithm as well as several other efficient all-atom and coarse grained folding dynamics algorithms are valuable tools to study folding dynamics and thermodynamics. However, any folding study has two major components: a) search strategy/

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

dynamic algorithm and b) energy function that should select native structure as global minimum. The energy function used in most of all-atom studies described above is based on Go prescription. This may be a good choice to study folding mechanism as it indeed guarantees that the native state is global energy minimum. However it requires knowledge of the native structure (or at least NOESY constraints from NMR experiments) and may underestimate energetic contribution and persistence of some non-native contacts. The latter were shown to play a possible role in nucleus formation, as predicted in simulations and bioinfomatics analysis [243] and confirmed in experiment [244].

The next step therefore is to develop atomic sequence-based potentials for all-atom simulations that would not require the knowledge of the native state and that may be transferable between proteins. This task is extremely challenging as many who work in protein structure prediction and simulations may appreciate. A few avenues can be explored here. Fundamentals of simple knowledge –based approaches using quasichemical approximation of the type pioneered by Tanaka and Scheraga [245] and further developed by Miyazawa and Jernigan [246], were studied and generalized to atomic level of description [247248] by Skolnick and coworkers. In particular these authors addressed a difficult question of what should be considered a *reference state* for such potentials. Reference state issue concerns the statistics of pairwise frequencies in the case when no interactions are present. Obviously any meaningful statistical signal about interactions should manifest itself in differences between observed statistics of interatomic contacts in proteins and those of the reference state. Another class of approaches are Z-score and related optimization methods [31,232]. A more recent new approach to design atomic potentials for protein folding was developed in our lab. It is based on selection of atomic potentials to make realistic protein energetics resemble Go-based energetics as much as possible. To this end, in spirit of knowledge-based potentials the interactions often observed in protein structures are deemed more attractive, while non-existent interactions are more repulsive. The form of the new potential (called μ-potential) is designed to coincide with Go-potential when derived on one protein and be closest to Go in terms of energetic bias to native state when derived on independent training dataset of protein structures:

$$E_{AB}=\frac{-\mu N_{AB}+(1-\mu)\tilde{N}_{AB}}{\mu N_{AB}+(1-\mu)\tilde{N}_{AB}} \tag{5.1}$$

where E_{AB} is contact interaction energy between atom types A and B, N_{AB} is the number of AB pairs found in contact and $\tilde{N}_{AB}$ is the number of AB pairs in the database that are not in contact. μ is a parameter that determines the average interaction (repulsion or attraction); it can be chosen to provide uniform and high (10-20%) acceptance rate in Monte Carlo simulation by preventing overly rapid collapse or excessively slow compactisation. The advantage of the new potential eq. (5.1). is that interactions energies between all atom types are confined to the range of values (-1,1) avoiding occasional overestimation of repulsive interactions in quasichemical methods in cases when interactions are not observed in the database. A systematic comparison of all methods to derive atomic potentials – quasichemical approximation, μ-potentials and optimization techniques was analyzed in a recent paper [232] based on results of fold recognition in gapless threading and against standard sets of decoys. It appears that all derived potentials show significant degree of consistency in a sense that in all cases the dominant interactions contributing to stabilization of the native fold are the same – interaction between side-chain atoms of aliphatic groups. However in terms of performance (Z-score of the native conformation) μ-potentials perform better than quasichemical potentials and about as good as optimized non-transferable potentials. This is important given that that

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

μ-potentials were derived on an independent dataset of proteins and were not optimized to perform a specific task.

The first application of μ-potential was for folding of a small three-helix bundle protein It showed repetitive and systematic folding within 2A RMS from crystal structure.[249]. However this was not a fully transferable μ-potential - it was derived using statistics of contacts in native structure of protein A itself. However the potentials derived from different databases seem to be strongly correlated [249] which is an encouraging sign that the potential may be transferable. A more stringent test of atomic potentials was made recently [54]. The energy function used for this study represented a linear combination of explicit hydrogen bonding potential (well suited to stabilize helical but not beta conformation) and μ-potential derived on an independent database of 103 proteins that did not contain tested proteins of their homologs. 84 atom types were considered (same as described in [249]). Simulations performed on seven small nonhomologous alpha-helical proteins showed encouraging results providing in 6 out of 7 cases folding to less than 4A rmsd structures from the native state. The analysis of simulation results included clustering of structures and observation that largest disjoint cluster – Giant Component contained most native-like conformations. (Fig.10)

Various graph-theoretical measures were tried to select "best" prediction and it appeared that most connected conformations – the ones that have most similar conformations – appeared to be statistically closer to the native state. Energy alone was effective but not most effective predictor of the native-like conformations. One possibility, as pointed out by Baker and Shortle [250] that clustering procedure alleviates some inaccuracies that are present with inexact potentials taking advantage of possibly greater number of states surrounding the native structure of the protein rather than infrequent low-energy decoys. Heteropolymer theory is consistent with that view pointing out that "random decoys" are akin to deep minima in random heteropolymers and represent isolated small sets of conformations on a rugged landscape as explained in Chapter 2. while native-like structures are less randomly organized [19,38]. Caflisch observed a similar phenomenon using a different clustering approach - protein folding network [135].

Of special interest are control simulations carried out for this study [54]. Simulation of randomized sequence folding resulted in a collection of conformations from which native structure of simulated proteins could not be identified by energy or any graph-theoretical criteria. However, interestingly some infrequent conformations were found that exhibited relatively low (4.2 A) rmsd with native structure of some proteins. This result may reflect on some conclusions from distributed computing approach where many folding simulations are run independently on a grid of computers. Some conformations were found in distributed computing among many simulations that were close in rmsd to native structure of a small target protein – villin headpiece [195,251]. However these low rmsd conformations did not appear to be lowest energy ones. A possibility exists therefore that low-rmsd conformations observed in distributed computing simulations are result of random collapse rather than sequence-based energy-guided folding. A similar random control for distributed computing simulations is necessary to address this important concern.

Another control concerns the issue of relative importance of pairwise interactions vs. explicit hydrogen bonds in formation of proper protein-like conformations. To this end a number of simulations were performed using energy function in which explicit hydrogen bond term was turned off. Resulting conformations formed almost perfect hydrophobic cores and were as compact as native ones but did not contain any helixes (less than 1% helical content) (Fig.11)

This result while it appears almost obvious is nevertheless important in light of recent suggestions that geometrical/topological and generic factors alone (such as excluded volume,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

topological constraints, compactness) are sufficient to provide protein-like architecture of compact polypeptide globules (modeled as polymers with “finite thickness”) [252,253]. [254]. In further development, the same authors incorporated explicit hydrogen bond into their model [255] to explain existing protein architectures. This view appears more consistent with results of simulations and earlier proposal by Ptitsyn and Finkelstein [256]. Most recently Skolnick and coauthors showed that collection of compact structures with hydrogen bonding is able to reproduce complete PDB [257].

6. Concluding remarks. Is protein folding problem solved?

Well, the answer to this question depends of what “is” is (William Jefferson Clinton). While many (but not all) conceptual aspects of protein folding (that used to be centered around “Levinthal paradox”) appear well understood and established, there is a lot of room for development and further studies as indicated throughout this review. Perhaps in coming years we will see further progress in using predictive atomistic level model to achieve complete description of folding pathway for several proteins. In particular an important aspect of protein folding problem is currently poorly understood. At what stage of the folding pathway do side-chains get packed and fixed into their native rotamer states? All-atom Monte-Carlo simulations suggest that as protein “descends” after the nucleation stage most side-chains adopt their final native conformations via local fluctuations [234]. Side-chains belonging to nucleus get “frozen” earlier, when nucleus is formed. While probable this picture needs further testing both by experiment and other simulation techniques.

Decisive departure from structure-centric (Go) models to sequence based all-atom models that are capable to simulate full folding process from random coil to native ensemble of conformations is an urgent need and an emerging reality. While the consequences of such models for structural genomics are obvious it is equally clear that their study will have significant impact on further understanding of protein folding mechanisms. In a certain sense such atomic-level simulations will represent a “final solution” of the problem of the protein folding mechanism However, protein folding has been an active field for more than 30 years and probably all conceivable mechanisms had been proposed in the literature either as pure speculations or as insights from coarse-grained models. In this sense, “the final solution” of the problem of protein folding mechanism will most likely look like a multiple-choice problem rather than an “essay”-like solution presenting an entirely novel mechanism that nobody thought of in the past. Most likely the “final solution” will combine elements of many mechanisms that researchers observed in simplified models in a more pure forms, so that in a sense the best “multiple choice” answer will sound like “all of the above”. Nevertheless we are bound to witness decisive progress in studies of protein folding in the coming years.

Acknowledgments

I am very grateful to my past and present coworkers Alex Gutin, Chrys Sfatos, Victor Abkevich, Leo Mirny, Sharad Ramanthan, Juergen Wilder, Nikolay Dokholyan, Jun Shimada, Edo Kussell, Isaac Hubner, Eric Deeds, Boris Shakhnovich, Igor Berezovsky, Konstantin Zeldovich for their numerous most valuable contributions and insights during many years of work on protein folding. Oleg Ptitsyn was a great source of knowledge, support and inspiration for me since my entrance the field in early eighties till his untimely death in 1999. Our collaborations and numerous discussions with Shura Grosberg, Alexei Finkelstein, Martin Karplus, Ricardo Broglia, Guido Tiana, Feng Ding, Sergey Buldyrev, Gene Stanley, Bill Eaton, Dave Thirumalai, Jeff Skolnick, Joan Shea have been always valuable and enjoyable and I am grateful to them for that.

References

1. Mirny L, Shakhnovich E. Annu Rev Biophys Biomol Struct 2001;30:361. [PubMed: 11340064]
2. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Annu Rev Phys Chem 1997;48:545. [PubMed: 9348663]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

3. Plotkin SS, Onuchic JN. Q Rev Biophys 2002;35:205. [PubMed: 12599750]
4. Shea JE, Brooks CL 3rd. Annu Rev Phys Chem 2001;52:499. [PubMed: 11326073]
5. Shakhnovich EI, Gutin AM. Biophys Chem 1989;34:187. [PubMed: 2611345]
6. Bryngelson JD, Wolynes PG. Proc Natl Acad Sci U S A 1987;84:7524. [PubMed: 3478708]
7. Shakhnovich EI. Curr Opin Struct Biol 1997;7:29. [PubMed: 9032061]
8. Pande VS, Grosberg AY, Tanaka T. Biophys J 1997;73:3192. [PubMed: 9414231]
9. Bryngelson JD, Wolynes PG. Journal of Physical Chemistry 1989;93:6902.
10. Derrida B. Physical Review B 1980;24:2613.
11. Go N. Adv Biophys 1984;18:149. [PubMed: 6544036]
12. Garel T, Orland H. Europhysics letters 1988;6:307.
13. Lifshits IM, Grosberg A, Khokhlov AR. Rev Mod Phys 1978;50:683.
14. Flory PJ. Brookhaven Symp Biol 1960;13:89. [PubMed: 13700386]
15. Edwards E, Anderson P. Journal of Physics F- Metal Physics 1975;5:965.
16. Parisi G. Physics Reports 1980;67:25.
17. Sfatos CD, Gutin AM, Shakhnovich EI. Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 1993;48:465.
18. Sfatos CD, Gutin AM, Shakhnovich EI. Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 1994;50:2898.
19. Sfatos CD, Shakhnovich E. Physics Reports 1997;288:77.
20. Shakhnovich E, Gutin A. Europhysics Letters 1989;8:327.
21. Shakhnovich E, Gutin A. Journal of Physics A - Mathematical and General 1989;22:1647.
22. Mezard M, Parisi G. Journal de Physique I 1991;1:809.
23. Goldbart PM, Castillo HE, Zippelius A. Advances in Physics 1996;45:393.
24. Giamarchi T, Le Doussal P, Orignac E. Physical Review B 2001;64Article Number 245119
25. Shakhnovich EI, Gutin AM. Nature 1990;346:773. [PubMed: 2388698]
26. Sfatos CD, Gutin AM, Shakhnovich EI. Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 1995;51:4727.
27. Shakhnovich EI, Gutin AM. Proc Natl Acad Sci U S A 1993;90:7195. [PubMed: 8346235]
28. Shakhnovich EI, Gutin AM. Journal of Theoretical Biology 1990;149:537. [PubMed: 2062107]
29. Gutin AM, Sali A, Abkevich V, Karplus M, Shakhnovich EI. Journal of Chemical Physics 1998;108:6466.
30. Shakhnovich E, Gutin A. Studia Biophysica 1989;132:137.
31. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Proc Natl Acad Sci U S A 1992;89:4918. [PubMed: 1594594]
32. Camacho CJ, Thirumalai D. Proc Natl Acad Sci U S A 1993;90:6369. [PubMed: 8327519]
33. Dinner AR, Abkevich V, Shakhnovich E, Karplus M. Proteins 1999;35:34. [PubMed: 10090284]
34. Privalov PL. Adv Protein Chem 1979;33:167. [PubMed: 44431]
35. Shakhnovich EI, Finkelstein AV. Biopolymers 1989;28:1667. [PubMed: 2597723]
36. Shakhnovich EI, Gutin AM. Protein Eng 1993;6:793. [PubMed: 8309926]
37. Shakhnovich EI. Physical Review Letters 1994;72:3907. [PubMed: 10056327]
38. Ramanathan S, Shakhnovich E. Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 1994;50:1303.
39. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Proc Natl Acad Sci U S A 1995;92:3626. [PubMed: 7724609]
40. Sali A, Shakhnovich E, Karplus M. J Mol Biol 1994;235:1614. [PubMed: 8107095]
41. Sali A, Shakhnovich E, Karplus M. Nature 1994;369:248. [PubMed: 7710478]
42. Shakhnovich EI, Gutin A. J Chem Phys 1990;93:5967.
43. Chan HS, Dill KA. Annu Rev Biophys Biophys Chem 1991;20:447. [PubMed: 1867723]
44. Maddox J. Nature 1994;370:13. [PubMed: 8015593]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

45. Wilder J, Shakhnovich EI. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2000;62:7100. [PubMed: 11102067]

46. Pande VS, Grosberg AY, Tanaka T. Macromolecules 1995;28:2218.

47. Zou J, Saven JG. J Mol Biol 2000;296:281. [PubMed: 10656832]

48. Mezard M, Parisi G, Sourlas N, Tolouse G, Virasosoro M. Phys Rev Lett 1984;52:1156.

49. Chaffotte A, Guillou Y, Delepierre M, Hinz HJ, Goldberg ME. Biochemistry 1991;30:8067. [PubMed: 1868082]

50. Privalov PL, Khechinashvili NN. J Mol Biol 1974;86:665. [PubMed: 4368360]

51. Lau KF, Dill KA. Proc Natl Acad Sci U S A 1990;87:638. [PubMed: 2300551]

52. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. Proc Natl Acad Sci U S A 1995;92:325. [PubMed: 7816842]

53. Shakhnovich EI. Fold Des 1998;3:R45.

54. Hubner IA, Deeds EJ, Shakhnovich EI. Proc Natl Acad Sci U S A 2005;102:18914. [PubMed: 16365306]

55. Kaya H, Chan HS. Proteins 2000;40:637. [PubMed: 10899787]

56. Socci ND, Onuchic JN, Wolynes PG. J Chem Phys 1996;104:5860.

57. Gutin AM, Abkevich VI, Shakhnovich EI. Fold Des 1998;3:183. [PubMed: 9562547]

58. Pande VS, Rokhsar DS. Proc Natl Acad Sci U S A 1999;96:1273. [PubMed: 9990014]

59. Sorenson JM, Head-Gordon T. Fold Des 1998;3:523. [PubMed: 9889163]

60. Klimov DK, Thirumalai D. Fold Des 1998;3:127. [PubMed: 9565757]

61. Go N, Taketomi H. Proc Natl Acad Sci U S A 1978;75:559. [PubMed: 273218]

62. Govindarajan S, Goldstein RA. Proteins 1995;22:413. [PubMed: 7479714]

63. Abkevich VI, Gutin AM, Shakhnovich EI. J Mol Biol 1995;252:460. [PubMed: 7563065]

64. Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Munoz V. Science 2002;298:2191. [PubMed: 12481137]

65. Gruebele M. C R Biol 2005;328:701. [PubMed: 16125648]

66. Ma H, Gruebele M. Proc Natl Acad Sci U S A 2005;102:2283. [PubMed: 15699334]

67. Zuo G, Wang J, Wang W. Proteins 2006;63:165. [PubMed: 16416404]

68. Das R, Gerstein M. Funct Integr Genomics 2000;1:76. [PubMed: 11793224]

69. Bowie JU, Luthy R, Eisenberg D. Science 1991;253:164. [PubMed: 1853201]

70. Dokholyan NV, Shakhnovich EI. J Mol Biol 2001;312:289. [PubMed: 11545603]

71. Jin W, Kambara O, Sasakawa H, Tamura A, Takada S. Structure 2003;11:581. [PubMed: 12737823]

72. Morrissey MP, Shakhnovich EI. Fold Des 1996;1:391. [PubMed: 9080185]

73. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. J Mol Biol 2003;332:449. [PubMed: 12948494]

74. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. J Mol Biol 2002;315:471. [PubMed: 11786026]

75. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Science 2003;302:1364. [PubMed: 14631033]

76. Voigt CA, Gordon DB, Mayo SL. J Mol Biol 2000;299:789. [PubMed: 10835284]

77. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF. J Am Chem Soc 2005;127:1346. [PubMed: 15686346]

78. Mirny LA, Abkevich VI, Shakhnovich EI. Proc Natl Acad Sci U S A 1998;95:4976. [PubMed: 9560213]

79. Mirny LA, Shakhnovich EI. J Mol Biol 1999;291:177. [PubMed: 10438614]

80. Finkelstein AV, Badretdinov A, Gutin AM. Proteins 1995;23:142. [PubMed: 8592696]

81. Li H, Helling R, Tang C, Wingreen N. Science 1996;273:666. [PubMed: 8662562]

82. Wolynes PG. Proc Natl Acad Sci U S A 1996;93:14249. [PubMed: 8962034]

83. Meyerguz L, Grasso C, Kleinberg J, Elber R. Structure (Camb) 2004;12:547. [PubMed: 15062078]

84. Govindarajan S, Goldstein RA. Proc Natl Acad Sci U S A 1996;93:3341. [PubMed: 8622938]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

85. England JL, Shakhnovich EI. Phys Rev Lett 2003;90:218101. [PubMed: 12786593]

86. Buchler NE, Goldstein RA. Proteins 1999;34:113. [PubMed: 10336377]

87. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. Genome Res 2005;15:385. [PubMed: 15741509]

88. Landau, LD.; Lifshi*t*s, EM.; Pitaevski*i, LP. Statistical physics. Vol. 3rd rev and enl. Pergamon Press; Oxford; New York: 1978.

89. Grzybowski BA, Ishchenko AV, Shimada J, Shakhnovich EI. Acc Chem Res 2002;35:261. [PubMed: 12020163]

90. England JL, Shakhnovich BE, Shakhnovich EI. Proc Natl Acad Sci U S A 2003;100:8727. [PubMed: 12843403]

91. Berezovsky IN, Shakhnovich EI. Proc Natl Acad Sci U S A 2005;102:12742. [PubMed: 16120678]

92. Taverna DM, Goldstein RA. Biopolymers 2000;53:1. [PubMed: 10644946]

93. Tiana G, Shakhnovich BE, Dokholyan NV, Shakhnovich EI. Proc Natl Acad Sci U S A 2004;101:2846. [PubMed: 14970345]

94. Gutin AM, Abkevich VI, Shakhnovich EI. Biochemistry 1995;34:3066. [PubMed: 7893719]

95. Sosnick TR, Mayne L, Hiller R, Englander SW. Nat Struct Biol 1994;1:149. [PubMed: 7656032]

96. Qi PX, Sosnick TR, Englander SW. Nat Struct Biol 1998;5:882. [PubMed: 9783747]

97. Park SH, Shastry MC, Roder H. Nat Struct Biol 1999;6:943. [PubMed: 10504729]

98. Abkevich VI, Gutin AM, Shakhnovich EI. Biochemistry 1994;33:10026. [PubMed: 8060971]

99. Du R, Pande V, Grosberg A, Tanaka T, Shakhnovich EI. Journal of Chemical Physics 1998;108:334.

100. Guo Z, Thirumalai D. Biopolymers 1995;35:137.

101. Guo Z, Thirumalai D. Biopolymers 1995;36:83.

102. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. J Mol Biol 2000;296:1183. [PubMed: 10698625]

103. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Biophys J 2002;83:3525. [PubMed: 12496119]

104. Itzhaki LS, Otzen DE, Fersht AR. J Mol Biol 1995;254:260. [PubMed: 7490748]

105. Shakhnovich E, Abkevich V, Ptitsyn O. Nature 1996;379:96. [PubMed: 8538750]

106. Wetlaufer DB. Proc Natl Acad Sci U S A 1973;70:697. [PubMed: 4351801]

107. Jacob M, Schindler T, Balbach J, Schmid FX. Proc Natl Acad Sci U S A 1997;94:5622. [PubMed: 9159122]

108. Cho SS, Levy Y, Wolynes PG. Proc Natl Acad Sci U S A 2006;103:586. [PubMed: 16407126]

109. Lorch M, Mason JM, Sessions RB, Clarke AR. Biochemistry 2000;39:3480. [PubMed: 10727243]

110. Pande VS, Grosberg A, Tanaka T, Rokhsar DS. Curr Opin Struct Biol 1998;8:68. [PubMed: 9519298]

111. Shakhnovich EI, Finkelstein AV. Doklady Akademii Nauk SSSR 1982;267:1247. [PubMed: 7151680]

112. Sheinerman FB, Brooks CL 3rd. J Mol Biol 1998;278:439. [PubMed: 9571063]

113. Cheung MS, Garcia AE, Onuchic JN. Proc Natl Acad Sci U S A 2002;99:685. [PubMed: 11805324]

114. Ptitsyn O, Dolgikh DA, Gilmanchin RI, Shakhnovich EI, Finkelstein AV. Molecular biology 1983;17:451.

115. Boczko EM, Brooks CL 3rd. Science 1995;269:393. [PubMed: 7618103]

116. Shakhnovich E, Farztdinov G, Gutin AM, Karplus M. Physical Review Letters 1991;67:1665. [PubMed: 10044213]

117. Plotkin SS, Wang J, Wolynes PG. Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 1996;53:6271.

118. Plotkin SS, Onuchic JN. Q Rev Biophys 2002;35:111. [PubMed: 12197302]

119. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M. Trends Biochem Sci 2000;25:331. [PubMed: 10871884]

120. Chan HS. Nature 1995;373:664. [PubMed: 7854444]

121. Finkelstein AV, Badretdinov A. Fold Des 1997;2:115. [PubMed: 9135984]

122. Galzitskaya OV, Ivankov DN, Finkelstein AV. FEBS Lett 2001;489:113. [PubMed: 11165233]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

123. Garbuzynskiy SO, Finkelstein AV, Galzitskaya OV. J Mol Biol 2004;336:509. [PubMed: 14757062]

124. Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Proteins 2003;51:162. [PubMed: 12660985]

125. Koga N, Takada S. J Mol Biol 2001;313:171. [PubMed: 11601854]

126. Thirumalai D. Journal de Physique I 1995;5:1457.

127. Kouza M, Li MS, O'Brien EP Jr, Hu CK, Thirumalai D. J Phys Chem A Mol Spectrosc Kinet Environ Gen Theory 2006;110:671. [PubMed: 16405339]

128. Sfatos CD, Gutin AM, Abkevich VI, Shakhnovich EI. Biochemistry 1996;35:334. [PubMed: 8555193]

129. Fersht AR. Proc Natl Acad Sci U S A 2000;97:1525. [PubMed: 10677494]

130. Shimada J, Shakhnovich EI. Proc Natl Acad Sci U S A 2002;99:11175. [PubMed: 12165568]

131. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG. Fold Des 1996;1:441. [PubMed: 9080190]

132. Shea JE, Onuchic JN, Brooks CL 3rd. Proc Natl Acad Sci U S A 1999;96:12512. [PubMed: 10535953]

133. Ding F, Guo W, Dokholyan NV, Shakhnovich E, Shea JE. J Mol Biol. 2005in press

134. Settanni G, Rao F, Caflisch A. Proc Natl Acad Sci U S A 2005;102:628. [PubMed: 15644439]

135. Rao F, Caflisch A. J Mol Biol 2004;342:299. [PubMed: 15313625]

136. Caflisch A. Curr Opin Struct Biol. 2006

137. Berezhkovskii A, Szabo A. J Chem Phys 2004;121:9186. [PubMed: 15527389]

138. Leopold PE, Montal M, Onuchic JN. Proc Natl Acad Sci U S A 1992;89:8721. [PubMed: 1528885]

139. Wales, DJ. Energy Landscapes. Cambridge University Press; Cambridge: 2003.

140. Ozkan SB, Dill KA, Bahar I. Protein Sci 2002;11:1958. [PubMed: 12142450]

141. Gutin AM, Abkevich VV, Shakhnovich EI. Physical Review Letters 1996;77:5433. [PubMed: 10062802]

142. Faisca PF, Telo Da Gama MM, Ball RC. Phys Rev E Stat Nonlin Soft Matter Phys 2004;69:051917. [PubMed: 15244857]

143. Faisca PF, Telo Da Gama MM. Biophys Chem 2005;115:169. [PubMed: 15752600]

144. Jewett AI, Pande VS, Plaxco KW. J Mol Biol 2003;326:247. [PubMed: 12547206]

145. Jackson SE. Fold Des 1998;3:R81. [PubMed: 9710577]

146. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Proteins 1995;21:167. [PubMed: 7784423]

147. Dill KA, Chan HS. Nat Struct Biol 1997;4:10. [PubMed: 8989315]

148. Onuchic JN, Wolynes PG. Curr Opin Struct Biol 2004;14:70. [PubMed: 15102452]

149. Krivov SV, Karplus M. Proc Natl Acad Sci U S A 2004;101:14766. [PubMed: 15466711]

150. Plaxco KW, Simons KT, Baker D. J Mol Biol 1998;277:985. [PubMed: 9545386]

151. Lindberg M, Tangrot J, Oliveberg M. Nat Struct Biol 2002;9:818. [PubMed: 12368899]

152. Liu C, Gaspar JA, Wong HJ, Meiering EM. Protein Sci 2002;11:669. [PubMed: 11847289]

153. Jones K, Wittung-Stafshede P. J Am Chem Soc 2003;125:9606. [PubMed: 12904024]

154. Gromiha MM, Selvaraj S. J Mol Biol 2001;310:27. [PubMed: 11419934]

155. Kuznetsov IB, Rackovsky S. Proteins 2004;54:333. [PubMed: 14696195]

156. Munoz V, Eaton WA. Proc Natl Acad Sci U S A 1999;96:11311. [PubMed: 10500173]

157. Gong H, Isom DG, Srinivasan R, Rose GD. J Mol Biol 2003;327:1149. [PubMed: 12662937]

158. Ivankov DN, Finkelstein AV. Proc Natl Acad Sci U S A 2004;101:8942. [PubMed: 15184682]

159. Chiti F, Taddei N, White PM, Bucciantini M, Magherini F, Stefani M, Dobson CM. Nat Struct Biol 1999;6:1005. [PubMed: 10542090]

160. Clarke J, Cota E, Fowler SB, Hamill SJ. Structure Fold Des 1999;7:1145. [PubMed: 10508783]

161. Martinez JC, Serrano L. Nat Struct Biol 1999;6:1010. [PubMed: 10542091]

162. Matouschek A, Kellis JT Jr, Serrano L, Fersht AR. Nature 1989;340:122. [PubMed: 2739734]

163. Lopez-Hernandez E, Serrano L. Fold Des 1996;1:43. [PubMed: 9079363]

164. Hamill SJ, Steward A, Clarke J. J Mol Biol 2000;297:165. [PubMed: 10704314]

165. Hamill SJ, Cota E, Chothia C, Clarke J. J Mol Biol 2000;295:641. [PubMed: 10623553]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

166. Plaxco KW, Larson S, Ruczinski I, Riddle DS, Thayer EC, Buchwitz B, Davidson AR, Baker D. J Mol Biol 2000;298:303. [PubMed: 10764599]

167. Mirny L, Shakhnovich E. J Mol Biol 2001;308:123. [PubMed: 11327757]

168. Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW. J Mol Biol 2002;316:225. [PubMed: 11851333]

169. Alm E, Baker D. Proc Natl Acad Sci U S A 1999;96:11305. [PubMed: 10500172]

170. Galzitskaya OV, Finkelstein AV. Proc Natl Acad Sci U S A 1999;96:11299. [PubMed: 10500171]

171. Makarov DE, Keller CA, Plaxco KW, Metiu H. Proc Natl Acad Sci U S A 2002;99:3535. [PubMed: 11904417]

172. Makarov DE, Plaxco KW. Protein Sci 2003;12:17. [PubMed: 12493824]

173. Wallin S, Chan HS. Protein Sci 2005;14:1643. [PubMed: 15930009]

174. Karanicolas J, Brooks CL 3rd. Proteins 2003;53:740. [PubMed: 14579364]

175. Eaton WA, Munoz V, Hagen SJ, Jas GS, Lapidus LJ, Henry ER, Hofrichter J. Annu Rev Biophys Biomol Struct 2000;29:327. [PubMed: 10940252]

176. Capaldi AP, Shastry MC, Kleanthous C, Roder H, Radford SE. Nat Struct Biol 2001;8:68. [PubMed: 11135674]

177. Schuler B, Lipman EA, Eaton WA. Nature 2002;419:743. [PubMed: 12384704]

178. Lipman EA, Schuler B, Bakajin O, Eaton WA. Science 2003;301:1233. [PubMed: 12947198]

179. Talaga DS, Lau WL, Roder H, Tang J, Jia Y, DeGrado WF, Hochstrasser RM. Proc Natl Acad Sci U S A 2000;97:13021. [PubMed: 11087856]

180. McCallister EL, Alm E, Baker D. Nat Struct Biol 2000;7:669. [PubMed: 10932252]

181. Sanchez IE, Kiefhaber T. J Mol Biol 2003;327:867. [PubMed: 12654269]

182. Sanchez IE, Kiefhaber T. J Mol Biol 2003;325:367. [PubMed: 12488101]

183. Wright CF, Lindorff-Larsen K, Randles LG, Clarke J. Nat Struct Biol 2003;10:658. [PubMed: 12833152]

184. Capaldi AP, Kleanthous C, Radford SE. Nat Struct Biol 2002;9:209. [PubMed: 11875516]

185. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, Fernandez JM. Nat Struct Biol 2003;10:738. [PubMed: 12923571]

186. Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, Smith DA, Perham RN, Radford SE. Nat Struct Biol 2003;10:731. [PubMed: 12923573]

187. Cecconi C, Shank EA, Bustamante C, Marqusee S. Science 2005;309:2057. [PubMed: 16179479]

188. Hubner IA, Shimada J, Shakhnovich EI. J Mol Biol 2004;336:745. [PubMed: 15095985]

189. Klimov DK, Thirumalai D. Proteins 2001;43:465. [PubMed: 11340662]

190. Ferrara P, Caflisch A. Proc Natl Acad Sci U S A 2000;97:10780. [PubMed: 10984515]

191. Simmerling C, Strockbine B, Roitberg AE. J Am Chem Soc 2002;124:11258. [PubMed: 12236726]

192. Finkelstein AV. Protein Eng 1997;10:843. [PubMed: 9415434]

193. Dinner AR, Karplus M. J Mol Biol 1999;292:403. [PubMed: 10493884]

194. Fersht AR, Daggett V. Cell 2002;108:573. [PubMed: 11909527]

195. Snow CD, Nguyen H, Pande VS, Gruebele M. Nature 2002;420:102. [PubMed: 12422224]

196. Karanicolas J, Brooks CL 3rd. J Mol Biol 2003;334:309. [PubMed: 14607121]

197. Shimada J, Kussell EL, Shakhnovich EI. J Mol Biol 2001;308:79. [PubMed: 11302709]

198. Clementi C, Garcia AE, Onuchic JN. J Mol Biol 2003;326:933. [PubMed: 12581651]

199. Paci E, Vendruscolo M, Karplus M. Proteins 2002;47:379. [PubMed: 11948791]

200. Munoz V, Thompson PA, Hofrichter J, Eaton WA. Nature 1997;390:196. [PubMed: 9367160]

201. Thompson PA, Eaton WA, Hofrichter J. Biochemistry 1997;36:9200. [PubMed: 9230053]

202. Hummer G, Garcia AE, Garde S. Proteins 2001;42:77. [PubMed: 11093262]

203. Hummer G, Garcia AE, Garde S. Phys Rev Lett 2000;85:2637. [PubMed: 10978126]

204. Kuszewski J, Clore GM, Gronenborn AM. Protein Sci 1994;3:1945. [PubMed: 7703841]

205. Ladurner AG, Itzhaki LS, Fersht AR. Fold Des 1997;2:363. [PubMed: 9427010]

206. Daggett V, Li A, Itzhaki LS, Otzen DE, Fersht AR. J Mol Biol 1996;257:430. [PubMed: 8609634]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

207. Paci E, Vendruscolo M, Dobson CM, Karplus M. J Mol Biol 2002;324:151. [PubMed: 12421565]

208. Vendruscolo M, Paci E, Dobson CM, Karplus M. Nature 2001;409:641. [PubMed: 11214326]

209. Li L, Shakhnovich EI. Proc Natl Acad Sci U S A 2001;98:13014. [PubMed: 11606790]

210. Hubner IA, Edmonds KA, Shakhnovich EI. J Mol Biol 2005;349:424. [PubMed: 15890206]

211. Hubner IA, Oliveberg M, Shakhnovich EI. Proc Natl Acad Sci U S A 2004;101:8354. [PubMed: 15150413]

212. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Nat Struct Biol 1999;6:1016. [PubMed: 10542092]

213. Friel CT, Beddard GS, Radford SE. J Mol Biol 2004;342:261. [PubMed: 15313622]

214. Jang S, Kim E, Shin S, Pak Y. J Am Chem Soc 2003;125:14841. [PubMed: 14640661]

215. Friel CT, Capaldi AP, Radford SE. J Mol Biol 2003;326:293. [PubMed: 12547210]

216. Martinez JC, Pisabarro MT, Serrano L. Nat Struct Biol 1998;5:721. [PubMed: 9699637]

217. Zarrine-Afsar A, Larson SM, Davidson AR. Curr Opin Struct Biol 2005;15:42. [PubMed: 15718132]

218. Zhou Y, Hall CK, Karplus M. Physical Review Letters 1996;77:2822. [PubMed: 10062054]

219. Jang H, Hall CK, Zhou Y. Biophys J 2002;83:819. [PubMed: 12124267]

220. Zhou Y, Karplus M. Nature 1999;401:400. [PubMed: 10517642]

221. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Fold Des 1998;3:577. [PubMed: 9889167]

222. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. J Mol Biol 2002;324:851. [PubMed: 12460582]

223. Borreguero JM, Ding F, Buldyrev SV, Stanley HE, Dokholyan NV. Biophys J 2004;87:521. [PubMed: 15240485]

224. Peng S, Ding F, Urbanc B, Buldyrev SV, Cruz L, Stanley HE, Dokholyan NV. Phys Rev E Stat Nonlin Soft Matter Phys 2004;69:041908. [PubMed: 15169044]

225. Yang S, Cho SS, Levy Y, Cheung MS, Levine H, Wolynes PG, Onuchic JN. Proc Natl Acad Sci U S A 2004;101:13786. [PubMed: 15361578]

226. Ding F, Buldyrev SV, Dokholyan NV. Biophys J 2005;88:147. [PubMed: 15533926]

227. Qiu L, Pabit SA, Roitberg AE, Hagen SJ. J Am Chem Soc 2002;124:12952. [PubMed: 12405814]

228. Richards FM, Lim WA. Q Rev Biophys 1993;26:423. [PubMed: 8058892]

229. Bromberg S, Dill KA. Protein Sci 1994;3:997. [PubMed: 7920265]

230. Kussell E, Shimada J, Shakhnovich EI. J Mol Biol 2001;311:183. [PubMed: 11469867]

231. Maiorov VN, Crippen GM. J Mol Biol 1992;227:876. [PubMed: 1404392]

232. Chen W, Shakhnovich EI. Protein Sci. 2005in press

233. Vendruscolo M, Mirny LA, Shakhnovich EI, Domany E. Proteins 2000;41:192. [PubMed: 10966572]

234. Kussell E, Shimada J, Shakhnovich EI. Proteins 2003;52:303. [PubMed: 12833553]

235. Takada S, Portman JJ, Wolynes PG. Proc Natl Acad Sci U S A 1997;94:2318. [PubMed: 9122192]

236. Thirumalai D, Ashwin VV, Bhattacharjee JK. Physical Review Letters 1996;77:5385. [PubMed: 10062790]

237. Pitard E, Shakhnovich EI. Phys Rev E Stat Nonlin Soft Matter Phys 2001;63:041501. [PubMed: 11308842]

238. Fowler SB, Best RB, Toca Herrera JL, Rutherford TJ, Steward A, Paci E, Karplus M, Clarke J. J Mol Biol 2002;322:841. [PubMed: 12270718]

239. Li H, Oberhauser AF, Fowler SB, Clarke J, Fernandez JM. Proc Natl Acad Sci U S A 2000;97:6527. [PubMed: 10823913]

240. Best RB, Li B, Steward A, Daggett V, Clarke J. Biophys J 2001;81:2344. [PubMed: 11566804]

241. Geissler PL, Shakhnovich EI. Phys Rev E Stat Nonlin Soft Matter Phys 2002;65:056110. [PubMed: 12059650]

242. Hyeon C, Thirumalai D. Proc Natl Acad Sci U S A 2003;100:10249. [PubMed: 12934020]

243. Li L, Mirny LA, Shakhnovich EI. Nat Struct Biol 2000;7:336. [PubMed: 10742180]

244. Viguera AR, Vega C, Serrano L. Proc Natl Acad Sci U S A 2002;99:5349. [PubMed: 11959988]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

245. Tanaka S, Scheraga HA. Proc Natl Acad Sci U S A 1975;72:3802. [PubMed: 1060065]
246. Miyazawa S, Jernigan RL. J Mol Biol 1996;256:623. [PubMed: 8604144]
247. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Protein Sci 1997;6:676. [PubMed: 9070450]
248. Zhang L, Skolnick J. Protein Sci 1998;7:112. [PubMed: 9514266]
249. Kussell E, Shimada J, Shakhnovich EI. Proc Natl Acad Sci U S A 2002;99:5343. [PubMed: 11943859]
250. Shortle D, Simons KT, Baker D. Proc Natl Acad Sci U S A 1998;95:11158. [PubMed: 9736706]
251. Zagrovic B, Snow CD, Shirts MR, Pande VS. J Mol Biol 2002;323:927. [PubMed: 12417204]
252. Banavar JR, Maritan A. Rev Mod Phys 2003;75:23.
253. Banavar JR, Maritan A, Micheletti C, Trovato A. Proteins 2002;47:315. [PubMed: 11948785]
254. Maritan A, Micheletti C, Trovato A, Banavar JR. Nature 2000;406:287. [PubMed: 10917526]
255. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A. Proc Natl Acad Sci U S A 2004;101:7960. [PubMed: 15148372]
256. Finkelstein AV, Ptitsyn OB. Prog Biophys Mol Biol 1987;50:171. [PubMed: 3332386]
257. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. Proc Natl Acad Sci U S A. 2006

Biography

Eugene Shakhnovich received his M.S. in 1981 in Theoretical Physics from Moscow University. In 1984, he received his Ph.D. in Theoretical Biophysics and Molecular Biology in from the Russian Academy of Sciences. He was a Research Fellow and Senior Research Fellow in the Institute of Protein Research of the then Soviet Academy of Sciences until his arrival to Harvard in 1990, where he held Assistant (1991-1995) and Associate (1995-1997) Professorships. He is now Full Professor of Chemistry, Chemical Biology and Biophysics (since 1997) at Harvard. His research interests include theoretical studies of Protein Folding, Evolution and Design, Rational Drug Design, theory of Complex systems, Bioinformatics and Theoretical Material Science. He is the author of more than 220 publications and a recipient of several awards and fellowships.

He is a member of several editorial boards. In 2001, he co-founded Vitae Pharmaceuticals, a vibrant pharmaceutical company which incorporates computational approaches developed by Shakhnovich lab into their drug discovery platform.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

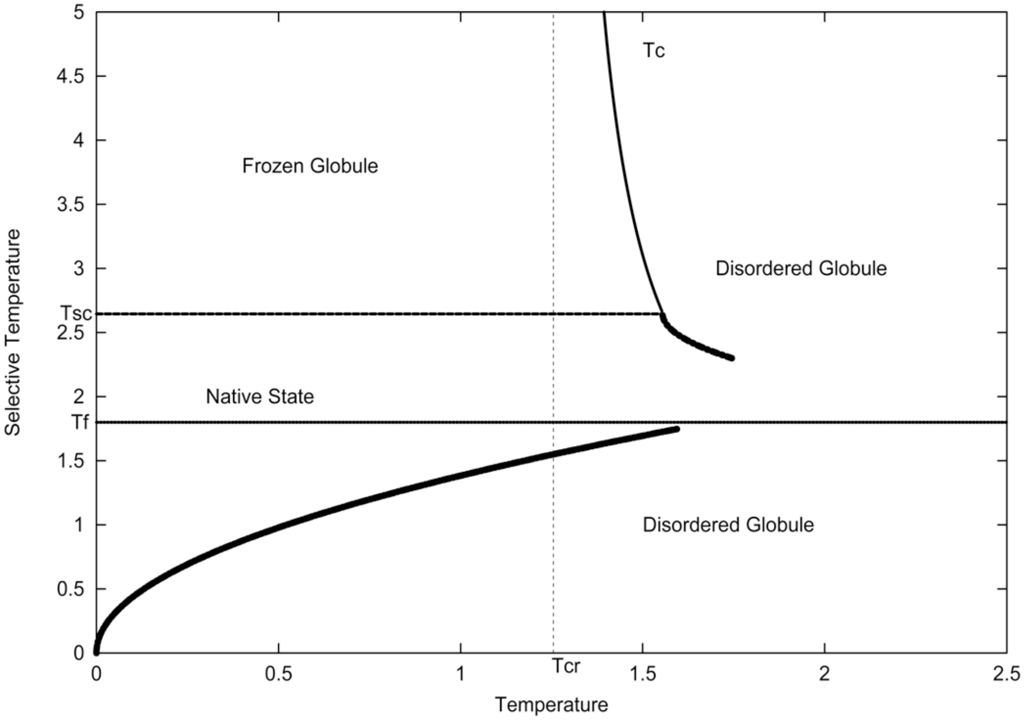


Figure.1.

Phase diagram for evolutionary selected protein-like heteropolymers. This phase diagram was derived in [45] for heteropolymers consisting of two types of residues – hydrophobic and polar.. High selective temperature corresponds to random sequences while lower selective temperature corresponds to protein-like evolutionary selected sequences. The transition from native state to disordered compact state is cooperative first order-like and gradual for evolutionary selected sequences (dashed line) and second order for random sequences (solid line around Tc). Reprinted with permission from [45]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

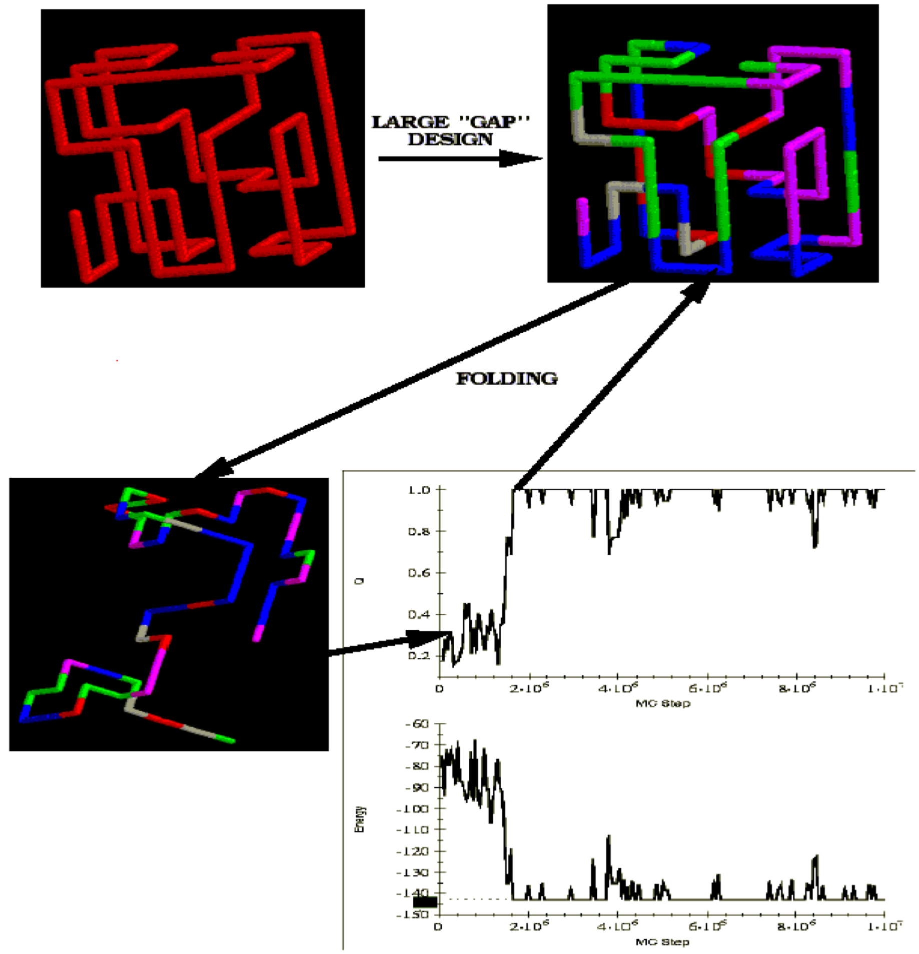
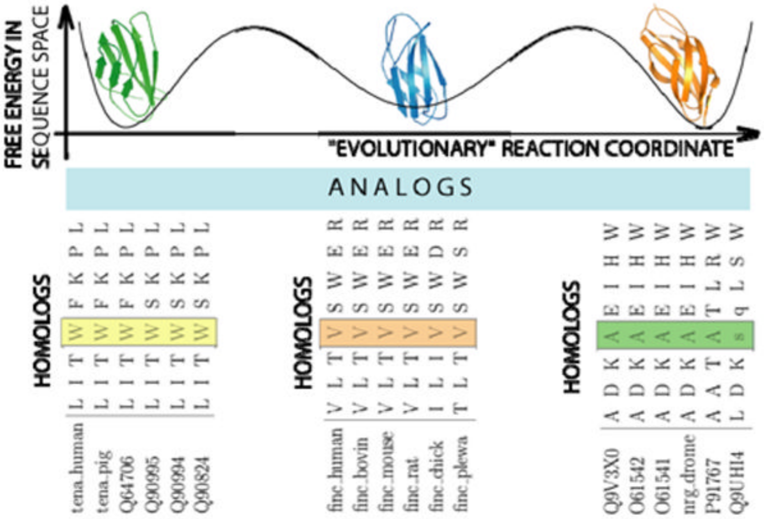


Figure.2.

Computational experiment showing that sequences designed with large energy gap fold cooperatively and rapidly into their native conformations[37]. First, a structure is chosen to serve as target, native conformation. Then sequences are designed (using Monte-Carlo search in sequence space with fixed composition) to have large energy difference (gap) between native conformation and set of structurally distinct misfolds. One of such sequences is memorized. Monte-Carlo folding simulations for this sequence start from an arbitrary random coil conformation and quickly and cooperatively converge to the target conformation for which the sequence was designed. The designed sequence has target conformation as its apparent global energy minimum as no conformations with energy lower than that of the target (native) conformation are found.


FREE ENERGY IN SEQUENCE SPACE
"EVOLUTIONARY" REACTION COORDINATE
ANALOGS
HOMOLOGS
HOMOLOGS
HOMOLOGS

tena_human	L	I	T	W	F	K	P	L
tena_pig	L	I	T	W	F	K	P	L
Q64706	L	I	T	W	F	K	P	L
Q90995	L	I	T	W	S	K	P	L
Q90994	L	I	T	W	S	K	P	L
Q90824	L	I	T	W	S	K	P	L

fnc_human	V	L	T	V	S	W	E	R
fnc_bovin	V	L	T	V	S	W	E	R
fnc_mouse	V	L	T	V	S	W	E	R
fnc_rat	V	L	T	V	S	W	E	R
fnc_chick	I	L	I	V	S	W	D	R
fnc_plewa	T	L	T	V	S	W	S	R

Q9V3X0	A	D	K	A	E	I	H	W
O61542	A	D	K	A	E	I	H	W
O61541	A	D	K	A	E	I	H	W
nrg_drome	A	D	K	A	E	I	H	W
P91767	A	A	T	A	T	L	R	W
Q9UHI4	L	D	K	s	q	L	S	W

Figure 3.

A schematic representation of the evolutionary processes that result in conservation patterns of amino acids. For a given family of folds, e.g. immunoglobulin (Ig) fold in this diagram, there are several alternative minima (3) in the hypothetical free energy landscape in the sequence space as a function of the “evolutionary” reaction coordinate (e.g. time). Each of these minima are formed by mutations in protein sequences at some typical time scales, τ_0, that do not alter the protein's thermodynamically and/or kinetically important sites, forming families of homologous proteins. Transitions from one minimum to another occur at time scales $\tau = \tau_0 \exp(\Delta G / T)$, where ΔG is the free energy barrier in sequence space separating one family of homologous proteins from another. At time scale τ mutations occur that would alter several amino acids at the important sites of the proteins in such a way that the protein properties are not compromised. At time scale τ the family of analogs is formed. In three minima we present three families of homologs (1TEN, 1FNF, and 1CFB) each comprised of six homologous proteins. We show 8 positions in the aligned proteins: from 18 to 28. It can be observed that at position 4 (marked by blocks) in each of the families presented in the diagram amino acids are conserved within each family of homologs, but vary between these families. This position corresponds to position 21 in Ig fold alignment (to 1TEN) and is conserved. We are very grateful to Nikolay Dokholyan for preparation of this figure.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

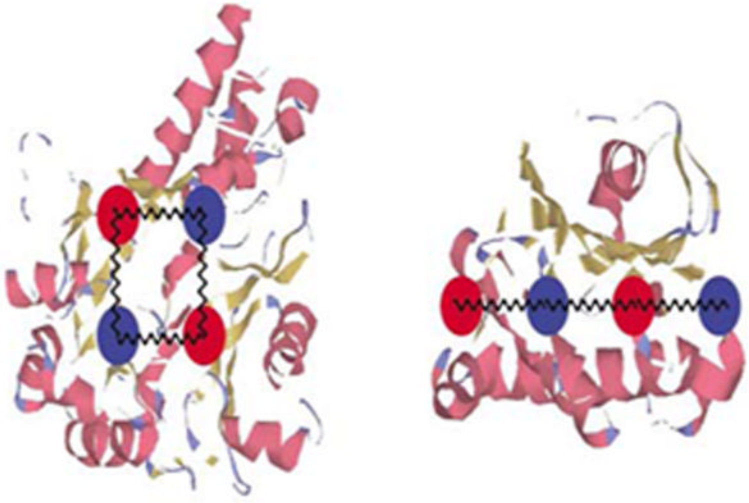


Figure 4.

An illustration of the physical reasons of why and how structure of a protein determines its designability. The balls schematically represent amino acids. Suppose that the interaction between the “red” amino acid and the “blue” amino acid is favorable and gives E = - 1. The configuration on the left yields lower energy –4, compared with right structures where contribution from interactions between these amino acids is only –3. Thus the 4-loop in the left structure contributes more to the stability of the structure overall allowing more freedom to select the remaining part of the sequence to obtain overall stabilization of the structure, Similar considerations apply to 3-loops, 5-loops etc. Reprinted with permission from [87].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

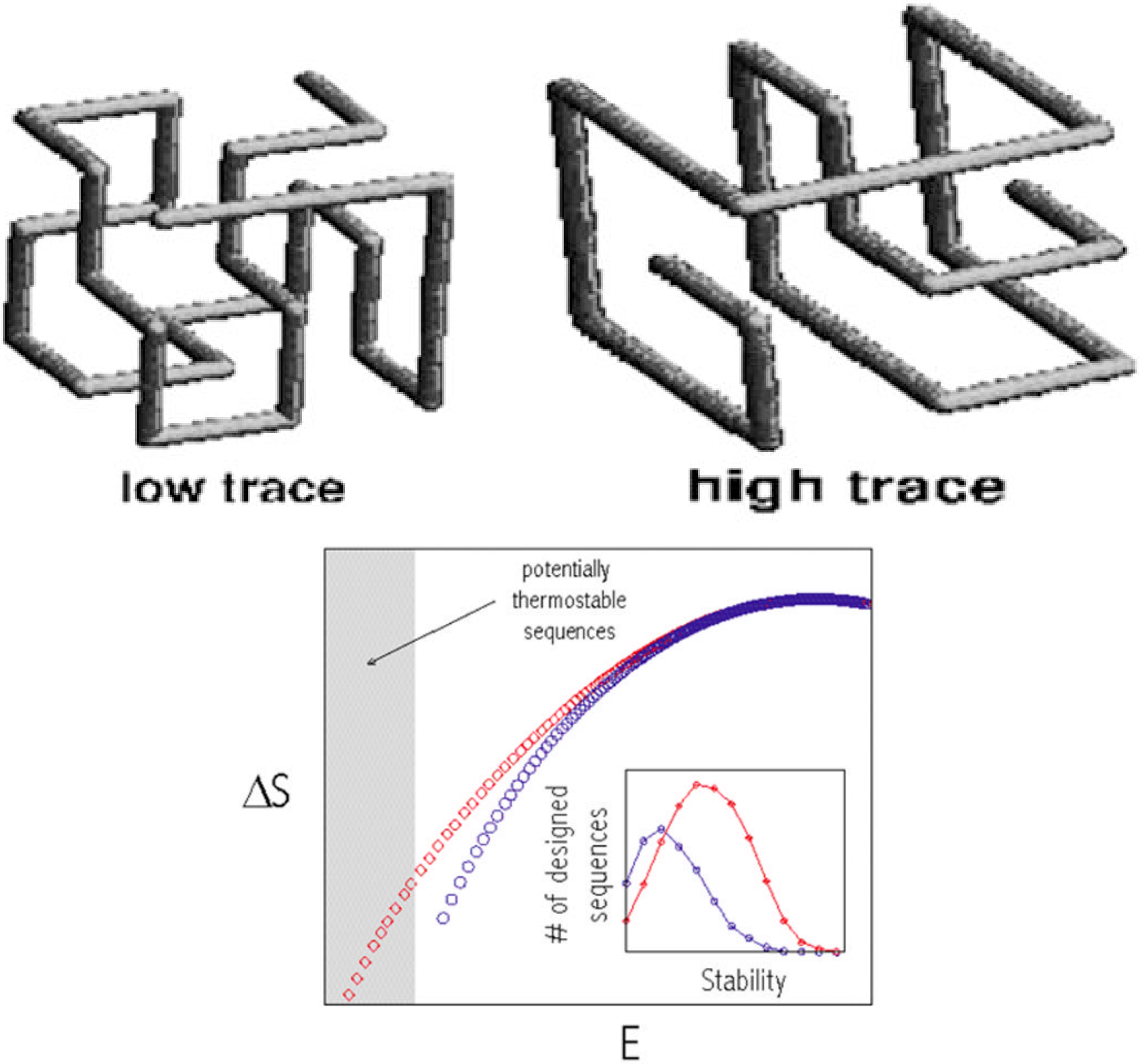


Figure.5.
(a) Two lattice structures – having highest and lowest predicted (by traces of their contact matrices) designabilities - and (b) counting of sequences that can fold into these structures with given energy. Δ*S* is entropy (log) of the number of sequences that fold into a given structure with a given energy counted from fully unconstrained statistics (at E=0). Blue points describe entropy of sequences designed for the low trace structure and red points are for high trace structure. The insert shows how many sequences can be stable (i.e. have high Boltzmann probability) in less and more designable structures respectively.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

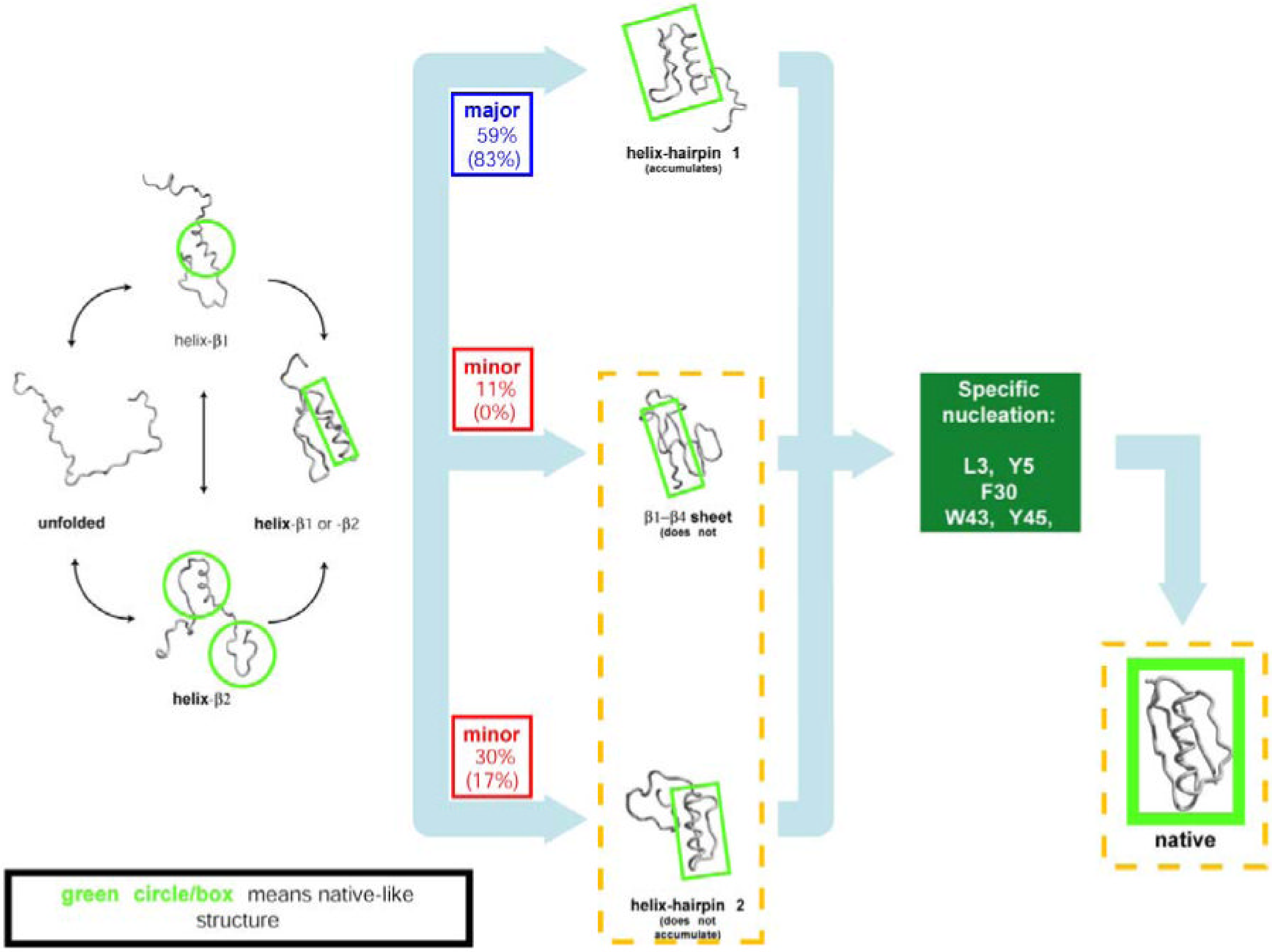


Figure 6.

Mechanism of folding of small protein G as derived from all-atom Monte-Carlo ensemble folding simulations with Go potential [130]. Parallel pathways through various helix-hairpin intermediates converge to common nucleation step that leads to final folding step. Reprinted with permission from [130]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

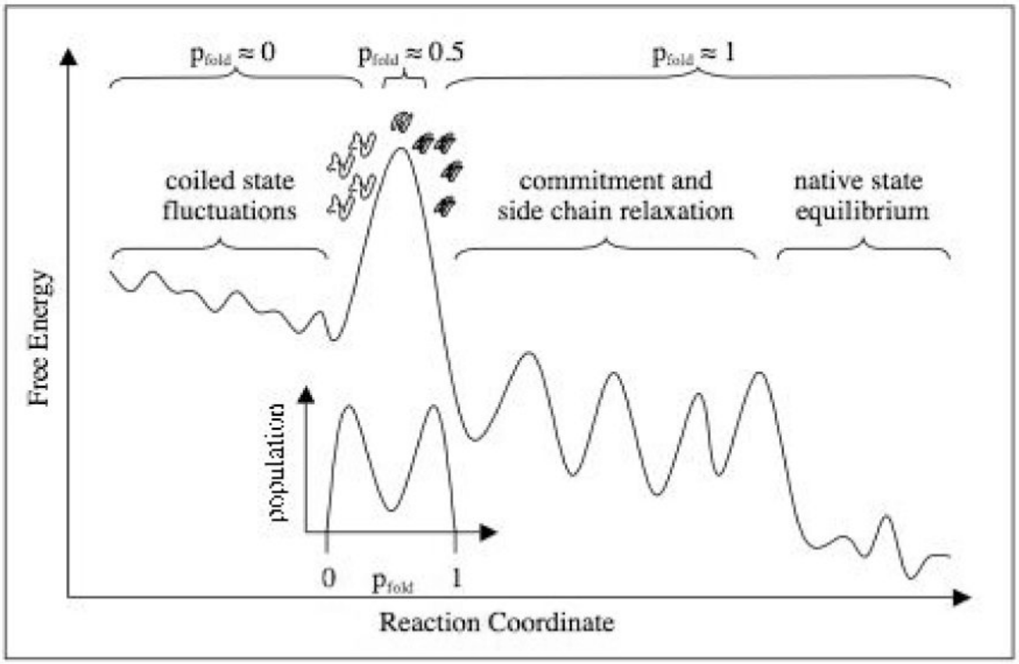


Figure 7.

A schematic representation of the putative free energy landscape and idea of p_{fold}. The transition state ensemble corresponds to the set of conformations at the "top" of free energy barrier (saddle point on free energy landscape). Passing the top of the barrier from unfolded to folded direction changes the dynamic behavior of the folding protein: it becomes committed to (in average) downhill folding. Folding dynamics starting from conformations on the "folded" side of the barrier always (apart from unlikely recrossing event) ends in the native basin, hence for these conformations probability to fold is 1. On the other hand folding dynamics that starts from conformations on the "unfolded" side of the barrier ends inevitably in the unfolded state; for such conformations p_{fold}=0. Conformations that belong to the barrier, i.e. transition state ensemble, have equal probability to fold and to unfold; for them p_{fold}=1/2. A rigorous definition of the Transition State Ensemble (TSE) is collection of conformations having p_{fold}=1/2. A detailed discussion of how to define and determine p_{fold} in realistic all-atom simulations can be found in [188]. An insert shows that ensemble distribution of p_{fold} is bimodal with TSE conformations corresponding to minimum probability. The hypothetical plot here is shown along hypothetical "reaction coordinate" for which the top of the barrier coincides with the TSE. The identity or even existence of such reaction coordinate is not known. Reprinted with permission from [188].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

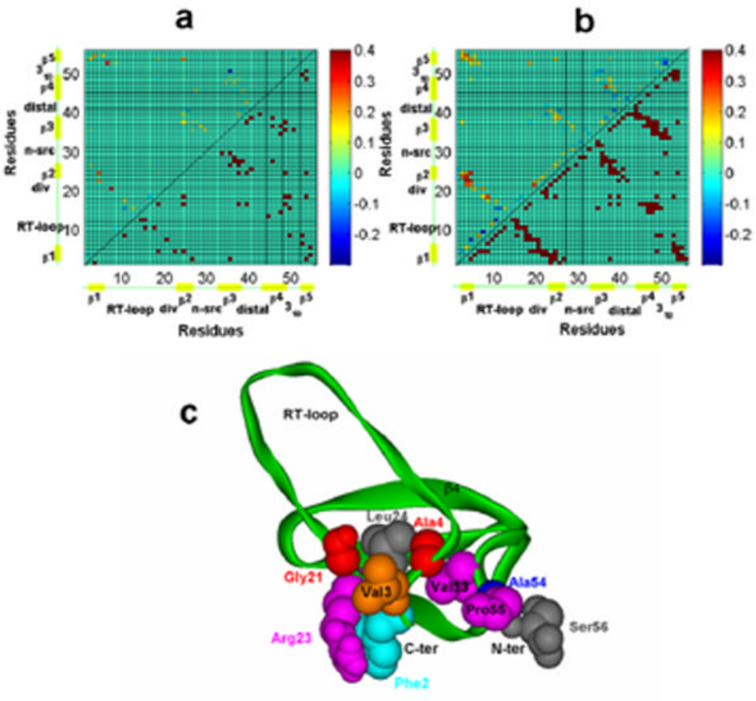


Figure 8.

Differential contact maps between pre-TS ensemble and TSE for src SH3 domain folding [133] (upper panels). (a) – for contacts between geometric centers of side chains, (b) – for contacts between C_β atoms. Lower panels on both contact maps correspond to native structure of the SH3 domain. (c) – Cartoon diagram of a sample TS structure determined by p_{fold} analysis. Residues with contact probabliy change from pre-TSE to TSE (as shown on upper panel of (b) greater than 0.1 are shown in space-filling scheme. They constitute a polarized folding nucleus for this domain. The figure and analysis are from [133]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

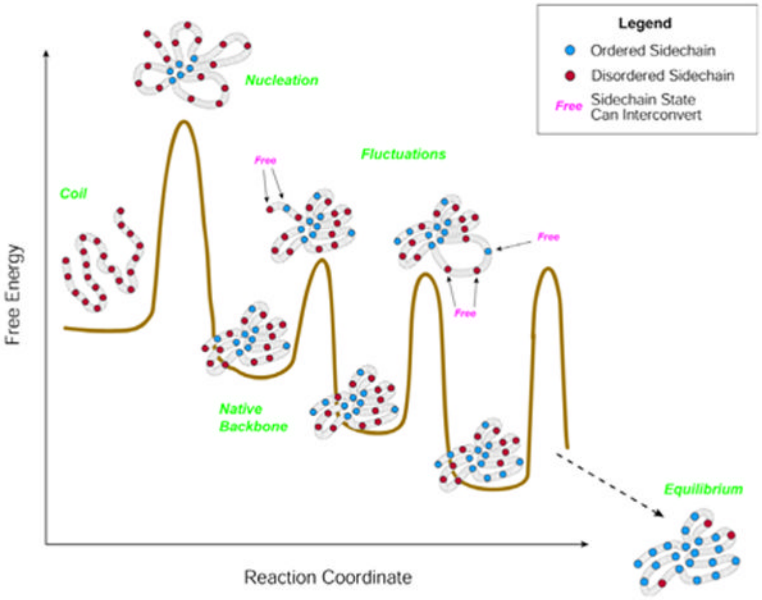


Figure .9.

A Schematic representation of full dynamic process of folding that includes side-chain organization. The main nucleation barrier is overcome first and leads to establishment of the overall fold. Subsequent dynamics includes local fluctuations of the backbone accompanied by progressive freezing of side-chains. Barrier heights are shown for illustrative purposes only and may be exaggerated and not representative of real situation. Reprinted with permission from [234]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

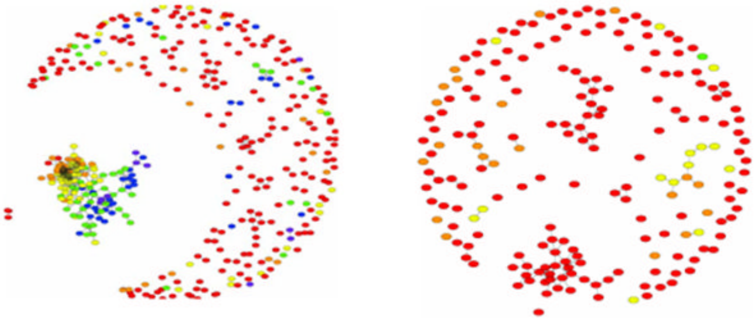


Figure 10.

Clustering of 200 conformations obtained in 200 independent simulation runs of all-atom MC folding algorithm with sequence-based transferable atomic μ-potential for protein A (1BDD) [54]. Each node corresponds to the lowest energy conformation obtained in each run and an edge is drawn between any two conformations if RMSD between them is less than 3.5A. Color code indicates RMSD from the native structure: purple: < 4A, blue: <5A, green: < 6A, yellow < 7A,orange< 8A, red: > 8A The central cluster- giant component – contains all native-like structures, while "peripheral" nodes are mostly misfolds. Figure on the right shows control: clustering of 200 conformations obtained in the same way but for random sequence with the same composition as for 1BDD. Comparison clearly shows that we observe sequence-guided non-trivial folding and that clustering focuses landscape for real sequence towards correct native structure. Reprinted with permission from [54]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

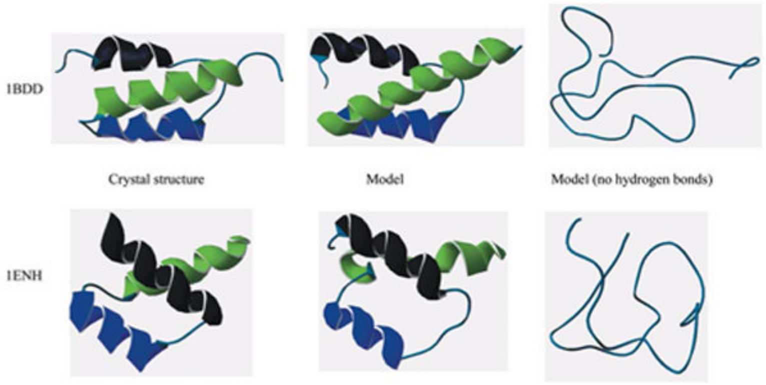


Figure 11.

Protein models from the PDB and representatives from simulation. Model simulations with full energy function (μ-potential pairwise interaction+ hydrogen bonding) fold to near native conformations while simulations without hydrogen bonding collapse without helices. Excluded volume and an attractive potential ensure a protein-like hydrophobic core and sidechain packing. However, representation of hydrogen bonding interactions is essential for formation of secondary structure