Published in final edited form as:

Technology (Singap World Sci). 2019 ; 7(1-2): 12–39. doi:10.1142/s233954781950002x.

A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme

Lawrence J. Williams¹, Brian J. Schendt¹, Zachary R. Fritz², Yonatan Attali¹, Robert H. Lavroff¹, Martin L. Yarmush²

¹Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 123 Bevier Rd., Piscataway, NJ 08854, USA.

²Department of Biomedical Engineering, Rutgers, The State University of New Jersey, 599 Taylor Road, Piscataway, NJ 08854, USA.

Abstract

Here we present a model to estimate the interaction free energy contribution of each amino acid residue of a given protein. Protein interaction energy is described in terms of per-residue interaction factors, μ . Multibody interactions are implicitly captured in μ through the combination of amino acid terms (γ) guided by local conformation indices (σ). The model enables construction of an interaction factor heat map for a protein in a given fold, allows prima facie assessment of the degree of residue-residue interaction, and facilitates a qualitative and quantitative evaluation of protein association properties. The model was used to compute thermal stability of T4 bacteriophage lysozyme mutants across seven sites. Qualitative assessment of mutational effects provides a straightforward rationale regarding whether a particular site primarily perturbs native or non-native states, or both. The presented model was found to be in good agreement with experimental mutational data ($R^2 = 0.73$) and suggests an approach by which to convert structure space into energy space.

Keywords

Coarse-Grain Protein Model; Protein Interaction Free Energy; Protein Engineering; Peptide Engineering; Protein Thermal Stability

INTRODUCTION

Biological processes are fundamentally realized at the atomic and molecular level and are mostly carried out by proteins. Molecular life science is teeming with protein studies and derived biotechnologies. Despite its tremendous success, biotechnology and its foundational molecular sciences lack a simple means to relate interaction energy to protein primary sequence, structure, and function^{1,2}. Advancement of such an understanding would significantly improve the pace and capabilities of the biotechnology industry³. Without it,

the community has grown to rely on intelligent guesswork made possible by genomics, proteomics, informatics, and other experimental tools for protein manipulation and structure determination⁴. The seminal insights of the founders of protein science^{5–10} largely have given way to methods of protein characterization that depend on all-atom computations^{11–16}. These powerful tools have not led to much-desired qualitative guidelines, intuitive understanding, or *prima facie* approximations of per-residue contributions to interaction free energy^{17–20}. It is an open question whether the difficulty in identifying governing principles of protein behavior is intrinsic to proteins or whether the difficulty is due to obfuscation by the surfeit of details generated by all-atom computations.

We began to explore the possibility of a simple model to understand protein energetics derived solely from protein sequence and approximate structure^{21,22}. A scale-invariant, nonatomistic approach could provide a straightforward understanding of proteins and enable *prima facie* analysis of sequence-dependent behavior^{23–27}. Moreover, it could substantively impact the biotechnology industry by simplifying the description of intra- and intermolecular protein interactions, facilitating protein design and engineering, and describing a direct connection between genomic sequence, protein structure, and biological function. At present, proteins are engineered to have enhanced binding affinity, thermostability, catalytic activity, and so on, through synthetic and recombinant technologies as well as by computational modeling^{28,29}. However, this process can be labor intensive and largely empirical without guidance provided by a predictive understanding of interaction free energy.

To overcome these limitations, we introduce a simple model to predict the per-residue protein interaction energy and demonstrate its applicability to understand thermal stability of the well-studied T4 bacteriophage lysozyme protein (**T4L**, Fig. 1a). We illustrate the ideas by focusing on single-point mutations of the wild-type protein, the use of a fractional exponent scale (Fig. 1b), and classification of mutations that are expected to primarily impact the energetics of the native state ensemble (NSE, Fig. 1c). Our findings suggest that this model has the potential to: (a) facilitate an understanding of the effects of mutations on protein thermal stability, (b) approximate well the interaction free energy changes that bear on the native state ensemble, and (c) convert protein structure space into energy space.

Summary of the model

Our model has three key features: (1) a high precision (fractional exponent) parameter for each amino acid residue (γ), (2) an index of four possible states that each residue can adopt (σ), and (3) normalization of a unique subset of residues for each amino acid in the sequence to determine each interaction factor (μ). These are briefly described in the following (see also Supplementary Information S1).

γ -Parameter —There are over 400 scales describing attempts to parameterize various properties of amino acid residues in proteins, such as structural propensities and hydrophobicity³⁰. However, there is only one scale-invariant set (Fig. 1b), and it is based on accessible surface area measurements^{24,25,31}. Using high-resolution crystallographic data, peptide segments as small as only three amino acid residues in length were computationally

separated from the parent protein as static objects, and the accessible area of the central residue was measured. The procedure was repeated for increasingly larger segments, each time measuring the accessible surface area of the central residue. Examination of thousands of segments revealed that each amino acid displayed scale invariance and obeyed a power law of the form $\alpha_r = N^{-\gamma}$, where α_r is the relative accessible surface area of the amino acid side chain and N is the length of the peptide segment analyzed. Although asymptotic reduction of amino acid side chain accessible surface area was noted early on^{32,33} and anticipated by steric crowding, it was not until the report of the γ scale³¹ that this behavior was shown to be distinctly different for each of the 20 amino acids. The exponents are fractional and range from approximately 1/16 (for lysine) to 1/4 (for cysteine) (Fig. 1b). The possible implications of this previously hidden dilation symmetry led us to formulate a model, which we term the Hidden Symmetry Model (HsymM), based on the information likely to be contained in these exponents and based on notions of how the effects captured by these terms might be propagated along the protein sequence (for the details of these conjectures, see Appendix).

σ -Parameter — Although protein conformation is conventionally defined by the backbone and side chain torsion angles^{34–36}, we anticipated advantages of an alternative conformational classification, which we term σ -index. The σ -index is a structural classification based on the closest-linked nearest neighbors (CLNN) of each amino acid residue. Briefly, CLNN are those residues with side chains near each other and separated by less than 4 residues in the sequence (see Supplementary Information S1 for a detailed description of CLNN determination and sigma assignment). The CLNN designation can be $\sigma = 1, 2, 3$, or 4 (Fig. 2). This index describes canonical secondary motifs in recognizable terms but also describes loops, turns, and coil structures. For example, a residue (i) that is part of a canonical beta strand would be assigned $\sigma = 2$, because $i + 2$ and/or $i - 2$ would be CLNN. Had that region of the protein been folded as an extended 3_{10} helix, the residue (i) would be assigned $\sigma = 3$, because $i + 3$ and/or $i - 3$ would be the CLNN. The CLNN approach thereby delimits local conformation according to relative amino acid side chain arrangement. This coarse treatment of conformation considers many variations of the local backbone and side chain dihedral angles as equivalent and accommodates small dynamical fluctuations^{37,38}.

μ -Parameter — Effective per-residue contributions to interaction energy (μ) are determined as a function of γ and σ terms (Eq. 1). Normalization of each intrinsic contribution (γ) of the *determining set* (θ_d^i) gives μ_i . The set, θ_d^i , consists of residue i , the CLNN ($i + \sigma$), along with a subset of residues nearby in the sequence that would constitute an extended regular σ motif — a surface composed of contiguous amino acid residue side chains centered on residue i . A physical interpretation of this is as a hydrated surface centered on the residue of interest, for example, the sort of extended motif that would be revealed upon partial unfolding^{19,39,40}. Each residue of this surface is considered critical and is therefore included in determining the interaction factor. Specifically, $i \pm n\sigma \in \theta_d^i$ where $n = 1, 2, 3, \dots$, and $n\sigma \leq 10$. The details of interaction factor (μ) determination and the related free energy calculations are given in Supplementary Information S1. Briefly, the free energy

of interacting residues i and j is related to the product of the interaction factors, μ , (Eq. 2), where G is free energy, ϵ relates medium effects, $\tau_{i,j}$ is an interaction efficiency term, and λ is a scaling factor with units of energy. Single-state analysis requires an estimate of entropic contributions. For the purposes of this study, we set $\epsilon = \tau_{i,j} = 1$ and we estimate λ in this study through fitting the **T4L** thermal data (see below). The implicit multibody interactions subsumed in the interaction factors (μ) aim to reflect solvation and other nonlocal effects that are hidden or obscured by multi-scale correlations. Although the scaling problem confounds protein free energy calculations, especially for per-residue free energy contributions^{17,41,42}, as shown below, this approach appears to describe interaction free energy successfully.

$$\mu_i^\sigma = (2 \xi_\sigma + 1)^{-1} \sum_{n=-\xi_\sigma}^{\xi_\sigma} \gamma_{i+\sigma n}, \quad \xi_\sigma = \left\lfloor \frac{\xi}{\sigma} \right\rfloor \quad (1)$$

$$\Delta G \sim -\epsilon \lambda \sum_{i,j(i \neq j)}^n \tau_{i,j} \mu_i \mu_j \quad (2)$$

RESULTS AND DISCUSSION

Native vs. non-native ensemble perturbations

The energy gap between the native and non-native ensembles determines the thermal stability of a protein fold^{43,44}. A decrease in thermal stability of a mutant protein reflects destabilization of the native state, stabilization of non-native states, or both; the converse holds as well. Experimental measures of thermal stability (G_{exp}) do not reveal which state or states are perturbed by mutation. Theoretical appraisal of thermal stability does not require knowledge of which states are perturbed, provided the energetics of a suitably representative set of a very large number of conformational microstates are computed accurately across appropriate timescales^{45–47}. This is difficult to achieve with certainty. The determination of mutational effects based on analysis of the native state alone is particularly challenging, because knowledge of whether the mutation primarily influences the native state is required.

Our model enables speculation of whether a mutation primarily influences native state stability and by how much (G_{calc}) based on evaluation of the native state alone. First, we categorize the impact of mutation on thermal stability as belonging to one of two classes (Fig. 1c). Mutation Class I (MC-I) are those mutants that are expected to primarily change the interaction factors of buried residues. Such changes are expected to correlate with protein stability, because these changes bear directly on stabilizing the native state. This presumes that the non-native state involves different residue–residue contacts and/or different backbone conformations^{48,49}. MC-I mutations are expected to minimally perturb the non-native state energetics. Mutation Class II (MC-II) are those mutants that change the interaction factors of surface residues or both surface and buried residues and therefore are

expected to perturb either the non-native states or both native and non-native states. This analysis suggests that the determination of whether a mutation is Class I can be assessed qualitatively by examining the ground state structure alone.

The set of residues whose μ factor would change upon mutation of site i determines the effect of the mutation on thermal stability. We term this group of residues the *member set* of site i (θ_m^i). It is important to note that θ_m^i and θ_d^i are not necessarily identical. The member set is composed of those residues whose μ values are impacted by mutation at site i , whereas the determining set is composed of those residues that are normalized to yield the μ value of site i . Analysis of θ_m^i can be used to assess whether mutation of residue i will be of Class I, as illustrated below.

μ -Profile of T4 lysozyme

T4L is one of the most extensively studied proteins^{50–53}. The availability of high-resolution structures for the wild-type protein and many of its mutant forms, combined with the corresponding thermal data, makes this protein an ideal candidate for comparative analysis⁵⁴. We focused on these structures because the single-point mutations of this protein do not significantly alter the native structure. Each possible μ factor for the wild-type sequence was calculated (Supplementary Table 1). We used the high-resolution structure of this protein (PDB ID: 3FA0⁵⁵) to assign a σ -index to each residue and the residue–residue contacts (see Supplementary Information S1 for details). Additionally, all mutants were assumed to retain these assignments. Figure 3 is a plot of the computed per-residue μ -factors for the native fold, along with the conventional α -helical and β -sheet motif assignments (top) and a color gradient (far right) to categorize and visualize relative μ values of the protein. These range from red (high/hot) to salmon, white, cyan, and then blue (low/cold). The protein structure heat-map, shown in Fig. 4a,b, provides a simple way to simultaneously visualize the approximate relative magnitude of interaction factors and the protein structure. For example, the protein exterior is dominated by cold residues. The interior, however, is dominated by hot residues, whose interactions significantly contribute to stability ($G \sim \mu\mu$). Supplementary Fig. 1 shows the heat maps of several other classic globular proteins, which also follow this trend. This model suggests that proteins may be organized such that the interaction factors are optimized. The canonical hydrophobic amino acid residues⁵⁶, that is, residues with nonpolar side chains, tend to be hot, whereas those with polar side chains, especially those with ionic functionality, tend to be cold. Approximately 25% of the residues in each of these classes, however, do not hold to these trends. Thus, the nature of the side-chain without regard for backbone conformation (σ -index) and sequence context is not a reliable predictor of μ factors. For example, in T4 lysozyme, some of the archetypal nonpolar residues are cold, such as L121, and some of the charged residues are hot, such as the catalytically relevant E11 (Fig. 4e,f).

Hot residue cluster analysis

The 31 residues of **T4L** with the highest μ factors, that is, the “hottest” residues, are indicated in the top shaded region of Fig. 3. The two largest hot clusters constitute the *C*- and *N*-terminal domain cores (Fig. 4c,d, respectively). The *N*-terminal domain, which is

known to be the less stable of the two, has a smaller core and has significantly fewer hot residues in the core than the *C*-terminal domain. An uninterrupted vein of hot residues connects the two domains. All 31 hottest residues (Supplementary Table 2), except for I9 and D159, make extensive contacts with other hot residues, and 21 of the 31 hottest residues colocalize in three clusters. Although the clustering leads to the protein interior being dominated by hot residues and the exterior being dominated by cold residues, there are some cool residues that are buried and some hot residues that are on the protein surface. For example, the most noticeable hot spot on the protein surface is the swath of contiguous hot residues that constitute the substrate binding region (Fig. 4b)^{57,58}.

Heat-map analysis

Many multimutant variants of **T4L** have been reported. In this study, we focused on 13 positions of wild-type enzyme that have been the subject of single-point mutations (see Supplementary Information S2). Virtually, all these mutants have been characterized in terms of both thermal stability and structure⁵⁴. We examined these mutants in detail. Supplementary Table 3 gives the θ_d^i and θ_m^i sets for each residue of this protein. On the basis of the location of each residue of the θ_m sets, seven sites (positions 3, 11, 115, 117, 119, 132, and 133) were assessed as MC-I mutations and therefore expected to contribute primarily to conformational ground state stabilization. For example, residue S117 rests near the center of an α -helix ($\sigma = 4$). As shown in Fig. 5a, site 117 is a member of the determining set of sites 107(1), 109(4), 113(4), 117(4), 121(4), 124(1), and 125(1) (i.e., θ_m^{117} , the number in parentheses indicates the σ -index of the residue listed). Consequently, mutation at 117 will impact the interaction factors of each member of this set in this fold. Many of the residues of θ_m^{117} , specifically 107, 117, and 121, are part of the hot *C*-terminal core cluster residues or contact these core residues. Therefore, mutation at 117 should bear primarily on ground state stability. The S117V mutant is the most stabilizing single point mutant known for **T4L** (ignoring covalent crosslinks). Interestingly, the thermal stability of this mutant has been difficult to rationalize^{53,54}. Within the framework of this model, the scale-invariant terms (γ) increase from serine to alanine to valine (Fig. 1b). Indeed, the interaction factors for the entire θ_m^{117} set will increase for S117A and S117V. The thermal stability is expected to increase because the majority of residues are buried and interact with (hot) core residues. The S117V mutant would be expected to increase in stability to an even greater extent over S117A. Similarly, residue M102 is buried, and θ_m^{102} includes some of the hottest *C*-terminal core residues in one of the largest clusters of hot residues in this protein (92(1), 93(1), 94(4), 95(1), 98(4), 102(4), 106(4), 107(1), 110(4) $\in \theta_m^{102}$, Fig. 5b). Hence, mutation at this site is also expected to bear primarily on ground state stability. However, relative to methionine, most mutations at position 102 would be expected to compromise stability (Fig. 1b). Thus, we used this model first to qualitatively assess the impact of the 13-point mutation sites, as MC-I or MC-II, and then to quantitatively assess the expected change in folding free energy for the 28 mutants distributed across these seven MC-I sites, as described in the following.

Quantitative assessment of mutant thermal stability

The change in folding free energy was calculated and compared to experimental data for the MC-I point mutants of T4L (Fig. 6, see also Supplementary Table 5). Briefly, free energy contributions are approximated by summing the products of interaction factors of residue (i) and its nearest neighbors (j) for the entire member set (Θ_m) of each mutation site (i.e., $\sum \mu_{i,j}$) for both the wild-type and mutant proteins. Changes in the backbone and side chain entropy at the site of mutation are also included (see Appendix, Eq. 3, and Supplementary Information S1 for details). It is noteworthy that the μ factors alone provide a reasonable, though less accurate, prediction of G_{calc} (see Supplementary Fig. 2). Although changes in local structure or among contacts between residues, and so on, could also be assessed, for simplicity, we approximate the structure of the mutants as perturbations of the wild-type protein with the same residue–residue contacts. The similarity of the crystal structures supports this approximation. Figure 7 illustrates the key aspects of the analysis with S117V. The S117V mutation significantly modulates the entire θ_m^{117} (Fig. 5c). For example, L121 is one of the residues impacted by this mutation. The 6 nearest neighbors of the L121 side chain are V87, L91, L118, W126, A129, and F153. All but A129 are calculated to have high μ -factors, that is, to be hot residues (Fig. 7a). In the wild-type, protein L121 is cold, but the S117V mutant converts L121 to a hot residue and thereby increases the estimated stability of the protein. Taken together, the S117V mutant is estimated to significantly increase protein stability — in agreement with experiment.

A contact schematic of nearest neighbors for L121 is illustrated in Fig. 7a (see inset). The central residue corresponds to i and the peripheral residues correspond to its nearest neighbors j (c.f., Eq. 2). Supplementary Table 6 shows the entire i, j array for the 13 mutation sites. We use the backbone and side chain confinement penalties described by Baxa *et al.*⁵⁹ to estimate these entropy contributions to free energy (Eq. 3, see Supplementary Information S1 for details). These data, the computed μ -factors, and the scaling factor λ ($\lambda = 12.5$ kcal/mol, Eq. 3) yielded G_{calc} , which is in good agreement with experimental results ($R^2 = 0.73$, AUE = 0.81 kcal/mol, Fig. 6).

Important correlations that govern the free energy are hidden, that is, there are contributions that are nonnegligible, nonlocal, and not easily identifiable by current models and rationales^{60,61}. Local interaction-based rationales, although possibly sound in terms of enthalpy, have been broadly recognized as untenable decompositions of free energy. Yet, perhaps because there is no alternative, this fundamental of thermodynamics is occasionally disregarded, and attempts to justify free energy-governed processes in terms of local atomic interactions are advanced. For example, in drug discovery, small molecule-protein binding free energy is sometimes rationalized in terms of individual interactions between the small molecule and the protein⁶². Recent studies based on sound theories use all-atom simulations to estimate free energy^{15,16,47,63}. These computed free energies are not decomposed into individual interactions, even in cases where free energy is binned on a per-residue basis^{20,64,65}, are determined by thermodynamic cycles that swap solvent arrays^{14,66–68}, or dissected in terms of many-body correlated changes^{60,61}. These approaches aim to sample over multiple timescales in an effort to solve the scaling problem of protein energetics.

Similarly, the model presented here does not attempt to decompose free energy into isolated individual interactions.

The structure of the model is reminiscent of earlier ideas on phase transition behavior of polymer-, protein-, and Ising-based statistical frameworks^{23,69–76}. The interaction factor, μ , defined in (Eq. 1), is a function of the average per-residue effects (γ) and local protein conformation (σ). The propagation length ξ , ($= 10$) corresponds to the standard hydropathic sliding window average, reflects sharp cut-offs of phase transition-like critical effects, and is consistent with extended Zimm–Bragg models. Importantly, the γ parameter includes critical factors such as solvation and excluded volume properties, among others. Although highly simplified, these ideas are in line with the energetic and dynamical properties of proteins being tuned to the energetic and dynamical properties of water, the long-range effects of solvation and liquid–vapor coexistence models, and protein–solvent fluctuations^{77–79}.

In addition to considering conformation in terms compatible with dynamic solvated surfaces (σ -index), the model uses multiple fractal terms defined for each amino acid. These are high-precision terms that uniquely capture properties related to accessible surface area of each amino acid residue in proteins. Proteins have been described in terms of fractal dimensionality, and much recent effort has been devoted to using protein fractal dimension to describe protein physics^{24–26,80–82}. Additionally, proteins have been described in terms of hybrid numerical–dynamical models to assess per-residue contributions to free energy^{20,64,65}. This study provides the first example of per-residue fractal dimension being used to describe protein energetics.

The generality of these findings requires further study. For example, although the other parameters of the model are not specific to **T4L**, additional studies are required to determine whether magnitude of the scaling parameter (λ) is specific to this protein. Nevertheless, the per-residue approximation of interaction free energy presented here appears promising, compares favorably to molecular simulations using state-of-the-art potentials and machine learning estimations, is easily visualized, and has the potential to be of considerable utility. Comparison to a recent study, where a state-of-the-art all-atom physical potentials simulation method (CHARMM36H) was used, shows the advantages of the presented model. A very good fit for the barnase mutant set was obtained through extended dynamics simulation (correlations of 0.67+)⁸³, and although computationally and setup intensive, it is indeed comprehensive. However, it does not enable straight forward per-residue interpretation, in contrast to the presented coarse-grained model.

CONCLUSION

Protein interaction free energy is the lynchpin to the understanding of many biological processes and for biotechnology applications, including protein and peptide engineering^{48,49,84}, design and development of biologics and small molecule drugs⁸⁵, among many others. Despite the temptation of describing protein stability and energetics in terms of local interactions such as ionic contacts and hydrogen bonding, these interactions alone are often too numerous, varied, weak, and delicately balanced to allow meaningful

characterization of per-residue contributions to interaction free energy. The model presented here outlines an approach to describe protein interaction energies based on sequence and per-residue contacts. This approach provides a straightforward basis to describe the individual and correlated effects of per-residue contributions to interaction free energy. As an example, T4 lysozyme was analyzed using the presented approach. Residues modeled to contribute significantly to interaction free energy of T4 lysozyme are depicted as hot in heat map renderings of the protein, with the interior of the protein dominated by hot residues and the exterior dominated by cool residues. The most notable exception to cool surface residues is the large hot patch on the surface that corresponds to the substrate binding surface of this enzyme^{57,58}. We used the model to classify mutations as primarily impacting ground state conformations (MC-I) or ground and/or excited state conformations (MC-II). In addition to these qualitative insights, differences in thermal properties, G , of T4 lysozyme single point MC-I mutants were calculated on the basis of the observed ground state structure, without recourse to dynamic simulations. Conveniently, the model requires insignificant computing power. The calculation of the values and generation of the heat maps require a few seconds of computing time on a dual processor (laptop). The agreement of this simple model with the experiment results is encouraging.

Although further studies on the scope and limitations studies of the model are warranted, the simplicity, directness, qualitative insights, and accuracy of the approach suggest potentially broad utility as well as complementarity to all-atom and bioinformatics-based protein modeling^{83,86,87}. The most exciting implications of this work are that the high complexity of protein structure can be reduced to a simple network of readily visualized and interpreted interaction factors and that it may be possible to convert structure space — so abundant in the wake of the structural genomics revolution — into energy space.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors wish to thank Prof. J. C. Phillips (Rutgers University) for introducing them to Ref. 31, Prof. Phillips and Dr. Doug Allan (Corning) for stimulating conversations, and Prof. M. A. Moret (Universidade Estadual de Feira de Santana, Brazil) for sharing the amino acid gamma exponents and standard deviations (*personal communication to LJW*). This work was partially funded through support from Corning Inc. and by NIH grants P41EB002503, R01EB020036, and T32GM008339.

APPENDIX

The two conjectures of the model flow from the two lines of inquiry. First, *What information is contained in, and what information is excluded from, power law exponents derived from accessible surface areas?* Conjecture #1: The exponents contain the fixed contributions to compaction for each amino acid residue, including the combined effect of solvation, steric/excluded volume, dispersion, and van der Waals interactions as well as stereoelectronic, polar, and Coulombic effects. The exponents do not contain contributions that vary from protein structure to protein structure, including temperature, medium effects, or sequence. Moreover, secondary, tertiary, and quaternary structural details, destabilizing charge-charge

interactions, and other fine structural details have also been removed. Idiosyncratic fluctuations have been averaged out of these exponents. Protein backbone and side chain entropy penalty factors would be expected to be largely absent as well, because these do not correlate strongly with side chain accessible surface area. Still, much information intrinsic to the amino acid residues in a protein context should be included in these exponents. A model that aimed to use these per-residue exponents as reduced complexity descriptors of protein interaction free energy would have to include the effects that are omitted and would have to build in the effects of sequence and conformation. Second, *How are these intrinsic per-residue effects propagated in proteins?* Conjecture #2: The effective contribution of an individual residue can be described by normalization of its intrinsic contributions with certain residues nearby in the protein sequence as a function of the amino acid residue's assigned conformation (σ -index, see Supplementary Information S1). A physical interpretation of this is as a hydrated surface centered on the residue of interest and corresponding to an extended motif revealed upon partial unfolding^{19,39,40}. Normalization in this way simultaneously includes multibody interaction, local structure, and local sequence effects and is described in more detail below. The model equation parameters are listed in Supplementary Information S1.

Proteins are considered at a level of coarseness that removes all molecular fine structure such that they are approximated as flexible chains with identical blobs uniformly spaced along the chain. The chain represents the protein backbone, blobs represent the side chains of residues, identical spacing is a consequence of proteins being composed of α -amino acid building blocks, and proximity of blobs is indicative of the impact of compaction. A set of simple rules determines the σ -index of each blob. Proximity of blobs renders them as nearest neighbors. As shown in Fig. 2, CLNNs are within four residues along the chain (i.e., σ -index of 1, 2, 3, or 4) and have side chains that are in contact. This guarantees that the σ -index refers to residues that are both clustered together and nearby in terms of sequence. Viewed linearly, a blob is assigned a σ -index based on the translational symmetry relationship — the number of steps within the interval determines the index. This is applicable to all residues in a protein, including extended strands, helical motifs, various turns, and loops. The CLNN definition captures regular secondary structure designations, that is, extended $\sigma = 2$ intervals correspond to beta strands, repeated $\sigma = 3$ intervals to the 3_{10} -helix motif, and repeated $\sigma = 4$ intervals to the canonical alpha-helix. The σ -index captures the essence of conformation and subsumes the many varied combinations of backbone and side chain angles of contacting residues.

Equation 1 expresses the ideas presented above in a compact form. Accordingly, for each residue i with index σ , the γ -values of θ_d^i are summed, normalized, and then assigned as interaction factor μ_i^σ . The μ term thus captures the effective sequence, conformation, and fluctuation dependence of each amino acid residue in a given fold of a given protein. As shown below, the interaction free energy of adjacent residues, i and j , is relatable to the product of adjacent interaction factors ($\mu_i^\sigma \mu_j^\sigma$). See also Supplementary Information S1.

Equation 2 indicates that the free energy is relatable to the sum of the products of interaction factors of residues i with nearest neighbors j . The sign indicates the favorability of low

energy; ϵ is a medium parameter, λ is a scaling factor (fit as 12.5 kcal/mol, see Fig. 6), and $\tau_{i,j}$ is an efficiency parameter of the i,j -interaction (e.g., polar/nonpolar compatibility). See also Supplementary Information S1.

Equation 3, given below, describes how the change in free energy upon mutation (ΔG) was computed. λ is a scaling parameter; both ϵ and $\tau_{i,j}$ are set to 1; T is the melting temperature in Kelvin of the wild-type protein; ΔT_m is the change in melting temperature of the mutant. Because only the native state is analyzed, S_i^{bb} and S_i^{sc} are protein backbone (bb) and side chain (sc) estimates of entropy change from non-native to native state. Equation 3 captures the net favorable interactions between residues i and j (as $\mu_i \mu_j$) and penalties of folding for wild type and mutant (as S_i^{bb} and S_i^{sc}). We use bb and sc entropy values determined from the evaluation of native and non-native state ensembles to approximate these confinement penalties⁵⁹. For simplicity, we set the number of neighbors (j) to a maximum of 6. Surface residues with less than 6 neighbors are assigned phantom solvent neighbors. Although formally based on pairwise nearest neighbor interactions, (Eq. 3) includes implicit multibody effects captured in μ_i^{σ} (Eq. 1). Even for mutants that adopt the same fold as the wild-type protein, there may be significant differences in many μ terms, and thus, multibody effects dominate the interaction free energy. See also Supplementary Information S1.

$$\Delta\Delta G = \left[-\lambda\epsilon \sum_{i,j(i \neq j)}^n \tau_{i,j} \mu_i \mu_j - (T + \Delta T_m) \sum_i^n (\Delta S_i^{bb} + \Delta S_i^{sc}) \right]_{\text{Mutant}} - \left[-\lambda\epsilon \sum_{i,j(i \neq j)}^n \tau_{i,j} \mu_i \mu_j - T \sum_i^n (\Delta S_i^{bb} + \Delta S_i^{sc}) \right]_{\text{Wild Type}} \quad (3)$$

REFERENCES

1. Bowman GR, Voelz VA & Pande VS Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* 21, 4–11 (2011). doi:10.1016/j.sbi.2010.10.006. [PubMed: 21081274]
2. Brenner S Life's code script. *Nature* 482, 461 (2012). doi:10.1038/482461a. [PubMed: 22358811]
3. Carlson R Estimating the biotech sector's contribution to the US economy. *Nature Biotechnol.* 34, 247 (2016). [PubMed: 26963545]
4. Burley SK et al. In *Protein Crystallography: Methods and Protocols*. (eds.) Wlodawer A, Dauter Z & Jaskolski M. (Springer, New York, pp. 627–641, 2017).
5. Baldwin RL Energetics of protein folding. *J. Mol. Biol.* 371, 283–301 (2007). doi:10.1016/j.jmb.2007.05.078. [PubMed: 17582437]
6. Tanford C How protein chemists learned about the hydrophobic factor. *Protein Sci.* 6, 1358–1366 (1997). doi:10.1002/pro.5560060627. [PubMed: 9194199]
7. Fersht A *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Macmillan, 1999).
8. Richards FM Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* 6, 151–176 (1977). doi:10.1146/annurev.bb.06.060177.001055. [PubMed: 326146]
9. Anfinsen CB Principles that govern the folding of protein chains. *Science* 181, 223–230 (1973). [PubMed: 4124164]
10. Tanford C & Reynolds J *Nature's Robots: A History of Proteins* (OUP Oxford, 2003).
11. Perez A, Morrone JA, Brini E, MacCallum JL & Dill KA Blind protein structure prediction using accelerated free-energy simulations. *Sci. Adv.* 2 (2016). doi:10.1126/sciadv.1601274.

12. Shaw DE et al. Atomic-level characterization of the structural dynamics of proteins. *Science* 330, 341–346 (2010). doi:10.1126/science.1187409. [PubMed: 20947758]
13. MacCallum JL, Perez A & Dill KA Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci.* 112, 6985–6990 (2015). doi:10.1073/pnas.1506788112. [PubMed: 26038552]
14. Wang L et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137, 2695–2703 (2015). doi:10.1021/ja512751q. [PubMed: 25625324]
15. Lens elink EB et al. Predicting binding affinities for GPCR ligands using free-energy perturbation. *ACS Omega* 1, 293–304 (2016). doi:10.1021/acsomega.6b00086. [PubMed: 30023478]
16. Ford MC & Babaoglu K Examining the feasibility of using free energy perturbation (FEP+) in predicting protein stability. *J. Chem. Inf. Model.* 57, 1276–1285 (2017). doi:10.1021/acs.jcim.7b00002. [PubMed: 28520421]
17. Piana S, Klepeis JL & Shaw DE Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 24, 98–105 (2014). doi:10.1016/j.sbi.2013.12.006. [PubMed: 24463371]
18. Freddolino PL, Harrison CB, Liu Y & Schulten K Challenges in protein-folding simulations. *Nat. Phys.* 6, 751 (2010). doi:10.1038/nphys1713. [PubMed: 21297873]
19. Skinner JJ et al. Benchmarking all-atom simulations using hydrogen exchange. *Proc. Natl. Acad. Sci.* 111, 15975–15980 (2014). doi:10.1073/pnas.1404213111. [PubMed: 25349413]
20. Gu J & Hilser VJ. Predicting the energetics of conformational fluctuations in proteins from sequence: A strategy for profiling the proteome. *Structure* (London, England: 1993) 16, 1627–1637 (2008). doi:10.1016/j.str.2008.08.016.
21. Henzler-Wildman K & Kern D Dynamic personalities of proteins. *Nature* 450, 964 (2007). doi:10.1038/nature06522. [PubMed: 18075575]
22. Zwier MC & Chong LT Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol.* 10, 745–752 (2010). doi:10.1016/j.coph.2010.09.008. [PubMed: 20934381]
23. Lesne A & Lagues M Scale Invariance: From Phase Transitions to Turbulence (Springer Science & Business Media, 2011).
24. Phillips JC Scaling and self-organized criticality in proteins I. *Proc. Natl. Acad. Sci.* 106, 3107–3112 (2009). doi:10.1073/pnas.0811262106. [PubMed: 19218446]
25. Phillips JC. Scaling and self-organized criticality in proteins II. *Proc. Natl. Acad. Sci.* 106, 3113–3118 (2009). doi:10.1073/pnas.0811308105. [PubMed: 19124778]
26. Reuveni S, Granek R & Klafter J Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure. *Proc. Natl. Acad. Sci.* 107, 13696–13700 (2010). doi:10.1073/pnas.1002018107. [PubMed: 20639464]
27. Banerji A & Ghosh I Fractal symmetry of protein interior: What have we learned? *Cell. Mol. Life Sci.* 68, 2711–2737 (2011). doi:10.1007/s00018-011-0722-6. [PubMed: 21614471]
28. Koga N et al. Principles for designing ideal protein structures. *Nature* 491, 222 (2012). doi:10.1038/nature11600. <https://www.nature.com/articles/nature11600#supplementary-information> [PubMed: 23135467]
29. Lehmann M, Pasamontes L, Lassen SF & Wyss M The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta* 1543, 408–415 (2000). doi:10.1016/S0167-4838(00)00238-7. [PubMed: 11150616]
30. Kawashima S et al. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205 (2008). doi:10.1093/nar/gkm998. [PubMed: 17998252]
31. Moret MA & Zebende GF Amino acid hydrophobicity and accessible surface area. *Phys. Rev. E* 75, 011920 (2007).
32. Creamer TP, Srinivasan R & Rose GD Modeling unfolded states of peptides and proteins. *Biochemistry* 34, 16245–16250 (1995). doi:10.1021/bi00050a003. [PubMed: 8845348]
33. Creamer TP, Srinivasan R & Rose GD Modeling unfolded states of proteins and peptides. II. backbone solvent accessibility. *Biochemistry* 36, 2832–2835 (1997). doi:10.1021/bi962819o. [PubMed: 9062111]

34. Ramachandran GN & Sasisekharan V In *Advances in Protein Chemistry*. (eds.) Anfinsen CB, Anson ML, Edsall JT & Richards FM. (Academic Press, Vol. 23, pp. 283–437, 1968). [PubMed: 4882249]
35. Ramakrishnan C & Ramachandran GN Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. *Biophys. J.* 5, 909–933 (1965). [PubMed: 5884016]
36. Hollingsworth SA & Karplus PA A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol. Concepts* 1, 271–283 (2010). doi:10.1515/bmc.2010.022. [PubMed: 21436958]
37. Zhang T, Faraggi E & Zhou Y Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins* 78, 3353–3362 (2010). doi:10.1002/prot.22842. [PubMed: 20818661]
38. DuBay KH & Geissler PL Calculation of proteins' total side-chain torsional entropy and its influence on protein-ligand interactions. *J. Mol. Biol.* 391, 484–497 (2009). doi:10.1016/j.jmb.2009.05.068. [PubMed: 19481551]
39. Englander SW, Mayne L, Kan Z-Y & Hu W Protein folding — How and why: By hydrogen exchange, fragment separation, and mass spectrometry. *Annu. Rev. Biophys.* 45, 135–152 (2016). doi:10.1146/annurev-biophys-062215-011121. [PubMed: 27145881]
40. Best RB, Hummer G & Eaton WA Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci.* 110, 17874–17879 (2013). doi:10.1073/pnas.1311599110. [PubMed: 24128758]
41. Meirovitch H, Cheluvuraja S & White RP Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding. *Curr Protein Peptide Sci.* 10, 229–243 (2009). [PubMed: 19519453]
42. De Vivo M, Masetti M, Bottegoni G & Cavalli A Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* 59, 4035–4061 (2016). doi:10.1021/acs.jmedchem.5b01684. [PubMed: 26807648]
43. Sanchez-Ruiz JM Protein kinetic stability. *Biophys. Chem.* 148, 1–15 (2010). doi:10.1016/j.bpc.2010.02.004. [PubMed: 20199841]
44. Karshikoff A, Nilsson L & Ladenstein R Rigidity versus flexibility: The dilemma of understanding protein thermal stability. *FEBS J.* 282, 3899–3917 (2015). doi:10.1111/febs.13343. [PubMed: 26074325]
45. Zwanzig RW High-temperature equation of state by a perturbation method. I. nonpolar gases. *J. Chem. Phys.* 22, 1420–1426 (1954). doi:10.1063/1.1740409.
46. Pohorille A, Jarzynski C & Chipot C Good practices in free-energy calculations. *J. Phys. Chem. B* 114, 10235–10253 (2010). doi:10.1021/jp102971x. [PubMed: 20701361]
47. Nascimento ECM, Oliva M, widerek K, Martins JBL & Andrés J Binding analysis of some classical acetylcholinesterase inhibitors: Insights for a rational design using free energy perturbation method calculations with QM/ MM MD simulations. *J. Chem. Inf. Model.* 57, 958–976 (2017). doi:10.1021/acs.jcim.7b00037. [PubMed: 28406297]
48. Bryan PN & Orban J Proteins that switch folds. *Curr. Opin. Struct. Biol.* 20, 482–488 (2010). doi:10.1016/j.sbi.2010.06.002. [PubMed: 20591649]
49. Alexander PA, He Y, Chen Y, Orban J & Bryan PN A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci.* 106, 21149–21154 (2009). doi:10.1073/pnas.0906408106. [PubMed: 19923431]
50. Matthews BW In *Advances in Protein Chemistry*. (eds.) Anfinsen CB, Richards FM, Edsall JT & Eisenberg DS. (Academic Press, Vol. 46, pp. 249–278, 1995). [PubMed: 7771320]
51. Xu J, Baase WA, Baldwin E & Matthews BW The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.* 7, 158–177 (1998). doi:10.1002/pro.5560070117. [PubMed: 9514271]
52. Rennell D, Bouvier SE, Hardy LW & Poteete AR Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* 222, 67–88 (1991). doi:10.1016/0022-2836(91)90738-R. [PubMed: 1942069]

53. Shoichet BK, Baase WA, Kuroki R & Matthews BW A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci.* 92, 452–456 (1995). doi:10.1073/pnas.92.2.452. [PubMed: 7831309]
54. Baas e WA, Liu L, Tronrud DE & Matthews BW Lessons from the lysozyme of phage T4. *Protein Sci.* 19, 631–641 (2010). doi:10.1002/pro.344. [PubMed: 20095051]
55. Mooe rs BHM, Tronrud DE & Matthews BW Evaluation at atomic resolution of the role of strain in destabilizing the temperature-sensitive T4 lysozyme mutant Arg 96 → His. *Protein Sci.* 18, 863–870 (2009). doi:10.1002/pro.93. [PubMed: 19384984]
56. Kyte J & Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132 (1982). doi:10.1016/0022-2836(82)90515-0. [PubMed: 7108955]
57. McHaourab HS, Oh KJ, Fang CJ & Hubbell WL Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. *Biochemistry* 36, 307–316 (1997). doi:10.1021/bi962114m. [PubMed: 9003182]
58. Kuroski R, Weaver LH & Matthews BW Structural basis of the conversion of T4 lysozyme into a transglycosidase by reengineering the active site. *Proc. Natl. Acad. Sci. U.S.A.* 96, 8949–8954 (1999). [PubMed: 10430876]
59. Baxa MC, Haddadian EJ, Jumper JM, Freed KF & Sosnick TR. Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci.* 111, 15396–15401 (2014). doi:10.1073/pnas.1407768111. [PubMed: 25313044]
60. Marks AE & van Gunsteren WF Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *J. Mol. Biol.* 240, 167–176 (1994). doi:10.1006/jmbi.1994.1430. [PubMed: 8028000]
61. Dills KA Additivity principles in biochemistry. *J. Biol. Chem.* 272, 701–704 (1997). doi:10.1074/jbc.272.2.701. [PubMed: 8995351]
62. Geschwindner S, Ulander J & Johansson P Ligand binding thermodynamics in drug discovery: Still a hot tip? *J. Med. Chem.* 58, 6321–6335 (2015). doi:10.1021/jm501511fs(2015). [PubMed: 25915439]
63. Clark AJ et al. Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of HIV-1. *J. Mol. Biol.* 429, 930–947 (2017). doi:10.1016/j.jmb.2016.11.021. [PubMed: 27908641]
64. Roy A, Perez A, Dill Ken A. & MacCallum JL Computing the relative stabilities and the per-residue components in protein conformational changes. *Structure* 22, 168–175 (2014). doi:10.1016/j.str.2013.10.015. [PubMed: 24316402]
65. Yu W, Baxa MC, Gagnon I, Freed KF & Sosnick TR Cooperative folding near the downhill limit determined with amino acid resolution by hydrogen exchange. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4747–4752 (2016). doi:10.1073/pnas.1522500113. [PubMed: 27078098]
66. Murpshy RB et al. WScore: A flexible and accurate treatment of explicit water molecules in ligand–receptor docking. *J. Med. Chem.* 59, 4364–4384 (2016). doi:10.1021/acs.jmedchem.6b00131. [PubMed: 27054459]
67. Wood s CJ et al. Rapid decomposition and visualisation of protein–ligand binding free energies by residue and by water. *Faraday Discuss.* 169, 477–499 (2014). doi:10.1039/C3FD00125C. [PubMed: 25340314]
68. Guimarães CRW & Mathiowetz AM Addressing limitations with the MM-GB/SA scoring procedure using the watermap method and free energy perturbation calculations. *J. Chem. Inf. Model.* 50, 547–559 (2010). doi:10.1021/ci900497d. [PubMed: 20235592]
69. Flory PJ & Garrett RR Phase transitions in collagen and gelatin systems. *J. Am. Chem. Soc.* 80, 4836–4845 (1958). doi:10.1021/ja01551a020.
70. Tran HT & Pappu RV Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys. J.* 91, 1868–1886 (2006). doi:10.1529/biophysj.106.086264. [PubMed: 16766618]
71. Zimm BH & Bragg JK Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* 31, 526–535 (1959). doi:10.1063/1.1730390.

72. Lifson S & Roig A On the theory of helix—Coil transition in polypeptides. *J. Chem. Phys.* 34, 1963–1974 (1961). doi:10.1063/1.1731802.
73. Lazaridis T & Karplus M Effective energy function for proteins in solution. *Proteins* 35, 133–152 (1999). doi:10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N. [PubMed: 10223287]
74. Selinger JV Introduction to the Theory of Soft Matter: From Ideal Gases to Liquid Crystals (Springer, 2015).
75. Lazaridis T & Karplus M Thermodynamics of protein folding: A microscopic view. *Biophys. Chem.* 100, 367–395 (2002). doi:10.1016/S0301-4622(02)00293-4.
76. Murza A & Kubelka J Beyond the nearest-neighbor Zimm–Bragg model for helix–coil transition in peptides. *Biopolymers* 91, 120–131 (2009). doi:10.1002/bip.21093. [PubMed: 18814306]
77. Lum K, Chandler D & Weeks JD Hydrophobicity at small and large length scales. *J. Phys. Chem. B* 103, 4570–4577 (1999). doi:10.1021/jp984327m.
78. Berne BJ, Weeks JD & Zhou R Dewetting and hydrophobic interaction in physical and biological systems. *Annu. Rev. Phys. Chem.* 60, 85–103 (2009). doi:10.1146/annurev.physchem.58.032806.104445. [PubMed: 18928403]
79. Chandler D Interfaces and the driving force of hydrophobic assembly. *Nature* 437, 640 (2005). doi:10.1038/nature04162. [PubMed: 16193038]
80. Senet P, Maisuradze GG, Foulie C, Delarue P & Scheraga HA How main-chains of proteins explore the free-energy landscape in native states. *Proc. Natl. Acad. Sci.* 105, 19708–19713 (2008). doi:10.1073/pnas.0810679105. [PubMed: 19073932]
81. Neusius T, Sokolov IM & Smith JC Subdiffusion in time-averaged, confined random walks. *Phys. Rev. E* 80, 011109 (2009).
82. Meroz Y, Ovchinnikov V & Karplus M Coexisting origins of subdiffusion in internal dynamics of proteins. *Phys. Rev. E* 95, 062403 (2017).
83. Gapsys V, Michielssens S, Seeliger D & de Groot BL Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. *Angew. Chem. Int. Ed. Engl.* 55, 7364–7368 (2016). doi:10.1002/anie.201510054. [PubMed: 27122231]
84. Rooklin D et al. Targeting unoccupied surfaces on protein–protein interfaces. *J. Am. Chem. Soc.* 139, 15560–15563 (2017). doi:10.1021/jacs.7b05960. [PubMed: 28759230]
85. Sivaprakasam P et al. Discovery of new acylaminopyridines as GSK-3 inhibitors by a structure guided in-depth exploration of chemical space around a pyrrolopyridinone core. *Bioorg. Med. Chem. Lett.* 25, 1856–1863 (2015). doi:10.1016/j.bmcl.2015.03.046. [PubMed: 25845281]
86. Gumbart JC, Roux B & Chipot C Efficient determination of protein–protein standard binding free energies from first principles. *J. Chem. Theory Comput.* 9, 3789–3798 (2013). doi:10.1021/ct400273t.
87. Duan LL, Zhu T, Li YC, Zhang QG & Zhang JZH Effect of polarization on HIV-1 protease and fluoro-substituted inhibitors binding energies by large scale molecular dynamics simulations. *Sci. Rep.* 7, 42223 (2017). doi:10.1038/srep42223. [PubMed: 28155907]

SIGNIFICANCE

The assessment of protein thermal stability as a function of amino acid sequence enables the testing of molecular models. While most quantitative studies rely on all-atom descriptions and dynamic simulations across many time scales to assess free energy changes, we introduce a model that uses only coarse structural data: protein sequence and a list of side chain–side chain contacts. The model uses amino acid-specific fractional exponents to describe the protein sequence and assigns each residue to one of four conformational states. Despite ignoring all molecular fine structure, including the details of backbone and side chain conformation, and all specific interactions, we find that changes in folding free energy can be estimated with good accuracy and that protein energetics can be interpreted and visualized in terms of per-residue contributions. The key implication of the study is that structure space may be readily converted into energy space.

Figure 1. Toward a reduced complexity description of proteins. **(a)** Lysozyme of T4 bacteriophage (T4L, PDB ID: 3fa0). **(b)** Scale invariant fractional values for the 20 canonical amino acids and with the standard deviation (sdev) determined from the power law: $\alpha_r \propto N^{-\gamma}$ (see Moret and Zebende³¹, text, and Appendix). **(c)** Mutations are classified on the basis of whether the impact of mutation is expected to be primarily on the native state ensemble (NSE) — Mutation class I (MC-I) or on the non-native state ensemble (non-NSE) or on both the NSE and the non-NSE — Mutation class II (MC-II).

Figure 2. σ -Index classification of local protein conformation. Amino acid side chains are coarsened as blobs (circles). The degree of separation between a residue (blue-filled black circle) and the closest-linked nearest neighbor (CLNN, blue-filled orange circle) defines σ . Simplified schematic of each closest-linked nearest neighbor index is shown for $\sigma = 1, 2, 3$, and 4 (two views).

Figure 3.

Computed μ -values of T4L. Per-residue values calculated according to Eq. 1 (blue lines in plot provided for clarity only). Color gradient (right) used to generate heat maps in Fig. 4, where red corresponds to the highest values (hot) and blue to the lowest (cold). Most of the hottest 31 residues make contiguous contacts and 21 aggregate in three clusters. Two of the largest are indicated (dotted lines). The first cluster (green) includes six of the hottest residues in the *N*-terminal domain. The second (black) includes 11 residues and is located in the *C*-terminal domain. The values and specific residue color used to determine the per-residue contributions and for visualization are given in Supplementary Table 4.

Figure 4.

Heat maps of T4 lysozyme. **(a)** The protein interior is predominantly hot (red and salmon interstitial lines) and the exterior predominantly cold (blue and cyan surface lines). Residues M102 (left) and L33 (right) are shown in space filling mode for reference (both residues are hot; hydrogen atoms are shown light grey for clarity). As indicated in Fig. 3, red/salmon correspond to high μ -factors, whereas white indicates midrange values and cyan/blue indicate low range values. **(b)** The patch of hot surface residues corresponds to the substrate binding region (box). **(c)** The *C*-terminal core region (box) is dominated by hot residues; M102 is in contact with many hot core residues. **(d)** The *N*-terminal core region (box), though smaller, is also dominated by hot residues; L33 is in contact with many hot core residues. **(e)** Approximately 25% of the canonical hydrophilic residues are hot, including the key catalytic residue E11. **(f)** Approximately 25% of the canonical hydrophobic residues are cool, including L121, which is buried and makes many contacts with hot residues of the *C*-terminal core. Mutations that increase the μ -factor for L121 (e.g., S117V, as explained later in the text) would be expected to significantly increase the stability of the *C*-terminal core and increase the protein stability

Figure 5.

Mutation impacts the contribution to interaction free energy of multiple residues. **(a)** Mutation of S117 (spheres) changes the μ -factors calculated for θ_m -related residues (sticks). Because serine has a low intrinsic value (γ , Fig. 1b), mutation to most other residues will increase these μ -factors. This set includes hot and cold residues that make many contacts with the hot residues of the *C*-terminal core and will therefore significantly stabilize the protein upon mutation of S in most cases. **(b)** Mutation of M102 (spheres) in most cases will decrease the μ -factors of the θ_m -related residues (sticks). This set of residues, which are hot core residues, will therefore cool down and significantly destabilize the protein.

Figure 6.

Comparison of computed and experimental ΔG of T4L mutants. Differences in thermal stability of well-characterized single point mutants were compared to experimentally determined values (Ref. Baase *et al.*⁵⁴). The calculation is remarkably accurate, considering the only input is sequence and a list of side chain-side chain nearest neighbors. $R^2 = 0.73$ (line); y-intercept = 0.44, average uncertainty in the sums and products of $\mu < 0.01$ kcal/mol; average uncertainty attributable to estimated sidechain and backbone entropy < 0.7 kcal/mol (Ref. Baxa *et al.*⁵⁹); average unassigned error (AUE) = 0.81 kcal/mol (white band); fit with (Eq. 3), $\epsilon = \tau = 1$, $\lambda = 12.5$ kcal/mol.

Figure 7.

Contacts (i,j pairs) for members of the θ_m set of residue 117. **(a)** Interaction energy of each residue is determined by summing the interaction factor products of each residue (i) with its nearest neighbors (j , set to a maximum of 6). **(b)** The impact of a single point mutation requires evaluating all the contacts for each residue of θ_m . Mutation of site 117 is illustrated, and members (i residues) are listed in the top row and neighbors (j residues) are listed in the corresponding columns. Columns with less than six neighbors indicate solvent exposed residues.