HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2017 April 01.

Published in final edited form as:

Proteins. 2016 April ; 84(4): 435–447. doi:10.1002/prot.24989.

The Role of Negative Selection in Protein Evolution Revealed through the Energetics of the Native State Ensemble

Jordan Hoffmann[§], **James O. Wrabl**[§], and **Vincent J. Hilser**[*]

Department of Biology and T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218

Abstract

Knowing the determinants of conformational specificity is essential for understanding protein structure, stability, and fold evolution. To address this issue, a novel statistical measure of energetic compatibility between sequence and structure was developed, using an experimentally validated model of the energetics of the native state ensemble. This approach successfully matched sequences from a diverse subset of the human proteome to their respective folds. Unexpectedly, significant energetic compatibility between ostensibly unrelated sequences and structures was also observed. Interrogation of these matches revealed a general framework for understanding the origins of conformational specificity within a proteome: specificity is a complex function of both the ability of a sequence to adopt folds other than the native, and ability of a fold to accommodate sequences other than the native. The regional variation in energetic compatibility indicates that the compatibility is dominated by incompatibility of sequence for alternative fold segments, suggesting that evolution of protein sequences has involved substantial negative selection, with certain segments serving as “gatekeepers” that presumably prevent alternative structures. Beyond these global trends, a size dependence exists in the degree to which the energetic compatibility is determined from negative selection, with smaller proteins displaying more negative selection. This partially explains how short sequences can adopt unique folds, despite the higher probability in shorter proteins for small numbers of mutations to increase compatibility with other folds. In providing evolutionary ground rules for the thermodynamic relationship between sequence and fold, this framework imparts valuable insight for rational design of unique folds or fold switches.

Keywords

Thermodynamic Environments; Gapless Threading; Metamorphic Proteins; Rational Design; Fold Recognition

[*]Correspondence: hilser@jhu.edu, Tel: 410-516-6072; Fax: 410-516-5213.

[§]Co-first authors.

Present address of J. H. is: Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA 02115. The authors are especially grateful to the Reviewers for their close reading and insightful, constructive comments.

Introduction

Why does an amino acid sequence adopt one particular unique fold and not one of the few thousands of alternatives? How do new folds arise and change during evolution of the proteome? Insight into these essential biological questions will be obtained by understanding the determinants of conformational specificity, the well-known ability of structured proteins to retain a finite population of native fold even under destabilizing conditions. One particularly interesting aspect of this problem is revealed by the repeated observation of "chameleon sequences" [1–4], which can adopt different folds, and the emerging discovery of "metamorphic proteins" [5, 6], which change fold as part of their function. Such extremes of conformational specificity, which have already been shown to be amenable to protein engineering [7–11], may prove to be an evolutionarily important mechanism for both fold change [9, 12–15] and functional versatility (as a prominent sub-class of "moonlighting proteins" [16, 17]). However, current bioinformatics tools and molecular dynamics simulations, using sequence or structure information, fail to reliably identify chameleon or metamorphic proteins [8, 18–20]. Novel information, not entirely based on either sequence or structure alone, may facilitate development of a more effective compatibility measurement between sequence and structure.

Our approach to addressing the problem is rooted in the ensemble nature of proteins [21], leveraging the long-standing realization that fold stability and conformational specificity are both thermodynamic in origin, and partially separable [22, 23]. Proteins in solution sample myriad conformations according to a Boltzmann distribution. Even when a single folded conformation is dominant, alternative structures could be transiently populated, albeit at vanishingly small amounts. Indeed, if protein sequences do obey Boltzmann statistics, each sequence has some probability of adopting every fold. Thus, the question of conformational specificity may be more tractable if rephrased: what is the difference in stability between one sequence adopting each of two alternative folds? Answering this question requires knowledge of the energetics of the compatibility between amino acid sequence and protein structure.

In this work, conformational specificity [22] is addressed from such a thermodynamic standpoint by development of a statistical framework for measurement of the compatibility between sequence and structure. An ensemble-based description of protein thermodynamics [24] is applied to a diverse database of protein folds, for which the positional thermodynamic stability of every residue is estimated. Importantly, this computational stability has been experimentally demonstrated [21, 24–28] to largely capture the cooperativity imparted by both local and global interactions, and from both enthalpic and entropic contributions. Thus, every residue in a protein can be described, not by residue letter or structural type, but instead by a so-called "thermodynamic environment" [29, 30].

Using a previously validated threading algorithm [30, 31], the energetic compatibility of amino acid sequence fragments adopting varied thermodynamic environment contexts was exhaustively computed, exploring general principles for conformational specificity and the organization of protein fold space. The results indicate that there is substantial energetic compatibility between ostensibly unrelated proteins, composed of energetically compatible

and incompatible contributions that are heterogeneously distributed throughout the sequence. We find that conformational specificity, operationally defined as the high energetic compatibility of one sequence with one fold, is a function of both sequence and fold, and that evolution of one fold from another may not be energetically improbable.

Furthermore, because energetic compatibility is correlated with the number of incompatible contributions, negative design appears to be important for conformational specificity, particularly for small, single domain proteins. This finding suggests that negative selection could be an evolutionary strategy to minimize the effects of metamorphic structure, as small proteins are expected to be more susceptible to fold-switching [11].

Materials and Methods

Ensemble-Based Thermodynamic Database of Diverse Human Proteins

This database has been described and used in previous analyses, of note is the presence of diverse secondary structural classes and fold types as curated by the *SCOP* database [32]. Briefly, 122 *H. sapiens* proteins of known structure (Table S1) were taken from the Protein Data Bank (PDB) [33] and native state Boltzmann-weighted thermodynamic ensembles were generated using the *COREX/BEST* algorithm [21, 34]. A summary of the computational procedure used to generate this database is given in Figure 4, below. When present in the PDB coordinates, selenomethionine residues were manually edited to methionine to permit execution of the algorithm. Parameters for the algorithm were: window size of 5 residues, minimum window size of 4 residues, simulated temperature of 25 °C, entropy weighting of 0.5, Monte Carlo sampling of at least 10,000 microstates per partition. Clustering of the *COREX/BEST* thermodynamic parameters ΔG, ΔH_{ap}, ΔH_{pol}, $T\Delta S_{conf}$ to obtain eight thermodynamic environments was performed by partitioning-around-medoids, implemented in *S-PLUS 6* (Insightful Corporation, Seattle, WA), as previously described [30, 31]. Thermodynamic parameters for the 17,801 residues in this database are given in Table S2. Log-odds scores (Figure 1), quantifying the observed to expected ratios of amino acids in thermodynamic environments, were computed from this database as previously described [29–31]. Secondary structure elements were assigned to each residue using *STRIDE* [35] and are listed in Table S2.

Exhaustive Gapless Scoring Between Sequence and Thermodynamic Environments

All amino acid sequences in the database were quantitatively compared with all proteins' thermodynamic environments. This was performed twice, first using complete sequence strings compared with complete environment strings, and second after dividing complete strings into overlapping 13 residue fragments starting at all possible registers. The second procedure was deliberately chosen for three reasons: to reveal regional contributions to energetic compatibility, to avoid possible length-dependent artifacts, and to keep the total amount of computations tractable. (Fragments of lengths 6 and 25 were also explored, with little qualitative change in results, data not shown.) Each comparison of sequence to environments used gapless scoring, popularly referred to as gapless "threading" [36] of an amino acid string against an environment string. A comparison was simply defined as a sum of the log-odds scores given by each residue/environment pair in Figure 1, using custom

scripts written in *Mathematica 9.0* (Wolfram Research, Champaign, IL). For example, the 13 residue amino acid sequence fragment starting at position 155 in the PDB coordinate file 1BYQ is NDDEQYAWESSAG. The threading score of this sequence fragment compared with the 13 residue thermodynamic environments fragment from the PDB file 1GP0 starting at position 1538, *i.e.* 5742211111248, is calculated as the sum of all 13 log-odds scores corresponding to each amino acid/environment pair, as listed in Figure 1. For this example, the sum would be 0.07 − 0.61 − 0.42 − 0.07 − 0.25 − 0.93 + 0.45 − 0.90 − 0.37 − 0.03 + 0.05 + 0.02 − 1.28 = −4.27. These computations were repeated until all sequence fragments were scored against all environments fragments. For comparisons of full-length proteins, the shorter protein was matched in all possible registers against the longer protein, such that the number of terms in the sum for each register was identical to the length of the shorter protein. Then, the maximum score over all registers was taken to be the single final score for that protein pair.

Parameterization of Probability Distributions: Significance of Energetic Compatibility

To assess the quality of these raw summed scores, a mathematical model was developed to estimate the expected chance occurrence of any particular raw score. Proteins of random composition and varying length were created by randomly choosing amino acids according to background frequencies in the Table S1 database (which were similar to background frequencies of amino acids seen in large sequence databases). These random sequences of amino acids were compared to identical length random sequences of similarly chosen thermodynamic environments and the total raw scores computed as described above. 120,000 such random proteins were scored at each chain length to obtain the reported histograms and curve fits (Fig. S1). Random protein creation, scoring, curve and distribution fitting were performed in *Mathematica* using custom scripts.

Empirical distributions of random gapless summed scores between amino acid sequences and thermodynamic environments were discovered to be statistically Gaussian for all lengths tested (Fig. S1). This result allowed the parameterization of a useful probability model for a gapless match of any length protein (Fig. S1b). In this model, as the length of a gapless match increased, it became progressively less likely to obtain a positive log-odds score (Fig. S1a); in other words, a randomly chosen sequence was expected to be energetically incompatible with a randomly chosen structure. In contrast, an extremely high positive score is uncommon in the model, and thus a significantly high score would be consistent with an empirical observation of “conformational specificity”: defined here as the extreme case where one amino acid sequence is energetically compatible with only one unique structure (Fig. S1a).

Computing Compatibility Index of Significant Matches and Principal Components Analysis

The 122 database proteins were exploded into 16,337 overlapping fragments of length 13 residues. Exhaustive all-*vs*-all comparisons of these 16,337 fragments resulted in greater than 266 million raw scores. Each raw score was then treated as a limit of integration in the length 13 Gaussian random score distribution, and the probability of obtaining a score of at most the observed raw score was computed using custom scripts. This list of *p*-values was filtered such that the best (most positive) and worst (most negative) of all comparisons,

defined as those exhibiting $p < 0.01$ or $p > 0.99$, were retained. The resulting filtered comparisons were then mapped back on to the positions of amino acid sequence or thermodynamic environments in the full-length proteins from which they originally came. Counts at each position were tabulated to produce a density of significant best, or worst, comparisons with regard to either sequence or structure. Thus, this analysis resulted in a total of four new attributes measured at every position in every protein: most significant matches of amino acid sequence against all other thermodynamic environments, least significant matches of amino acid sequence against all other thermodynamic environments, most significant matches of thermodynamic environments against all other amino acid sequences, and least significant matches of thermodynamic environments against all other amino acid sequences. These four attributes were, respectively, named "positive compatibility index (PCI with respect to sequence)", "negative compatibility index (NCI with respect to sequence)", "positive compatibility index (PCI with respect to structure)", and "negative compatibility index (NCI with respect to structure)" throughout the rest of this paper. To minimize possible end effects, the N-terminal 12 and C-terminal 13 values for each protein were ignored, resulting in a total of 14,751 residue positions, with four density counts at each position. These data were treated as a four-dimensional space and were subjected to standard eigenvalue decomposition [37] using an in-house *C* program (Figure 2). "Aggregate Negative Compatibility Indices" with respect to sequence or structure of an individual protein were defined as the integrated area along the entire protein of these respective densities (*i.e.* the area under the blue curves in Figs. 7a and 7b, respectively).

Provisional Classification of Energetic Compatibility: Susceptibility to Fold Switch

The median PCI and NCI within each protein was used to classify residue positions according to the following definitions. Figure 3 is a visual representation of this classification that may be referenced when the various categories are discussed later in the text. "Gatekeeper" positions exhibited an NCI greater than median and a PCI less than median; the term "Gatekeeper" was meant to capture the intuitive notion of a protein fragment being energetically unlikely to adopt any known conformation. "Permissive" positions exhibited an NCI less than median and a PCI greater than median; the term "Permissive" was meant to capture the intuitive notion of a protein fragment being energetically likely to adopt many conformations. "Selective" positions exhibited NCI and PCI both greater than median; the term "Selective" was meant to capture the intuitive notion of a protein fragment being energetically likely to adopt multiple conformations but simultaneously being unlikely to adopt others. In other words, "Selective" positions could indicate regions of a protein more susceptible to fold switching. "Inactive" positions exhibited NCI and PCI both less than median; the term "Inactive" was meant to capture the intuitive notion of no strong conformational preference. Since NCI and PCI were separate attributes of both sequence and structure, each residue position was assigned two classifications, one in terms of sequence and one in terms of structure. These classifications are listed in Table S2 for the proteins analyzed in this work.

Results

Proteins Represented in Energetic Terms

Previous work has established that proteins can be represented in energetic rather than in structural terms [30]. The conceptual basis of this energetic representation is that the positional thermodynamic stability of a folded protein can be computationally estimated, by treating the protein as a Boltzmann-weighted ensemble of partially folded microstates [24]. This process, algorithmically named COREX/BEST [34], can be summarized as follows (Figure 4). The experimental coordinates (*i.e.* crystallographic or NMR structure) are the input for COREX/BEST (Fig. 4, Step 1). A large number, typically millions, of partially folded microstates involving all regions of the protein are generated based on the input (Fig. 4, Step 2); a key simplification here are the assumptions that any folded conformation is native-like and any unfolded conformation is expressed by average amounts of newly exposed polar and apolar surface area, relative to the PDB structure. [21, 24, 38] Each microstate is assigned a Gibbs free energy from a surface-area based function, and statistical weights and populations are calculated for every microstate in the ensemble (Fig. 4, Steps 2 & 3). For every residue position *j* in the protein, the entire ensemble is partitioned into sub-ensembles in which the position is either in a folded conformation or an unfolded conformation (Fig. 4, Step 4), thus defining a position-specific equilibrium constant, $\kappa_{f,j}$, between folded and unfolded. This equilibrium constant can be converted (Fig. 4, Step 5) to a position-specific stability, ΔG_j, which quantitatively matches experimental position-specific stabilities measured from hydrogen exchange (Fig. 4, Step 6). Statistical analysis of the COREX/BEST output from a large number of diverse proteins results in a meaningful simplification of all position-specific stabilities into a small number (*i.e.* eight [30]) of clusters that share similar average values of stability.

Using our structure-based model of the native state ensemble (*i.e.* COREX/BEST) [21], it has been shown that these eight different “thermodynamic environments” [29] exist within any protein [30]. Furthermore, the propensities of amino acids to appear in these environments could be used as the basis of a fold recognition algorithm, much in the same way that helical sequences can be predicted from known helix propensities. Figure 5 shows an example protein color-coded according to the ensemble-based thermodynamic description of proteins, which is represented as eight color-coded environments [30]. Each energetic environment has a characteristic average stability resulting from enthalpic and entropic contributions associated with the computed change in solvent accessible polar and apolar surface upon locally unfolding each segment (Fig. 5, bottom) [21]. Importantly, these environments report on the energetics observed at a particular position rather than the contribution of the individual amino acid occupying that position, thus revealing how homologous proteins with marginal sequence identity can nonetheless share common thermodynamic signatures, and thus identical folds [29, 39]. As demonstrated, this representation has recapitulated numerous experimental observations that ground-state structures of proteins have regions of relatively high and low thermodynamic stability, and that these regions are not always intuitive upon visual inspection of the structure [28, 40].

As noted previously, several key features of this representation are exemplified in the Hsp90 protein (Fig. 5). First, the most stable regions are often in the core of the protein, which is true of this Hsp90. Second, elements of secondary structure, even those located in the core, are not uniformly stable: it is often observed that the middle residue positions of elements are more stable than the termini [41]. Third, although the most unstable regions are loops and turns, not all loops and turns are necessarily unstable, a counterintuitive result that has been borne out by experiment [42]. Although there are at least two low stability turns in this example (purple or blue), there is a prominent higher stability (orange) turn between strands 4 and 5 (upper left, Fig. 5), and the apparently coil-like linker (dark red) between strand 3 and helix 3 is among the highest stability regions of any protein in the database.

This energetic representation of proteins alone has formed the basis of an effective fold recognition algorithm, whereby sequences could be matched with their respective folds [29–31], even if the secondary structure information of the fold was not present in the training set [43]. This last result, that the energetic information of entirely alpha-helical proteins permitted recognition of entirely beta sheet proteins, compellingly established the universality of this energetic representation with regard to protein structure classification [44].

Quantifying Energetic Compatibility between Homologous and Non-Homologous Proteins

To test whether structured full-length proteins exhibit significant energetic compatibility with their respective sequences using the probability model described in Methods, we applied the model to the scores of all amino acid sequences in the database against all sets of thermodynamic environments (Fig. 6). Because the log-odds scores (Fig. 1) are dependent on both amino acid and thermodynamic environment, an all-*vs.*-all plot is necessarily separated into scoring of sequences against a structure (rows in Fig. 6), and structures against a sequence (columns in Fig. 6). Unlike scoring derived from symmetric amino acid substitution matrices, this analysis is not symmetric and thus may reveal differential scoring contributions from either a sequence or a structure perspective.

There are several noteworthy observations in Fig 6. First, the diagonal of this plot, representing “self” matches of an amino acid sequence to its known correct fold, was clearly populated by substantial and significant scores, indicating that the algorithm works. These correct matches were highly specific: except for known homologous proteins (as classified in the *SCOP* database), no non-self match exhibited a *p*-value more significant than approximately 0.001. Although expected conformational specificities were thus recapitulated by the significant energetic compatibilities, no obvious relationship was observed that differentiated conformational specificity with respect to sequence or environments (the median correlation coefficients between rows and columns of Fig. 6 was $r = +0.5$, data not shown). Also not observed was any general pattern between fold type (*e.g.* all-alpha or all-beta, Fig. 6 braces) and energetic compatibility. For example, the mixed alpha + beta proteins 1BYQ and 1MWP did not exhibit increased energetic compatibility to other mixed alpha + beta proteins (boxed vertical columns in Fig. 6).

Unexpectedly, however, there was a large amount of marginal, yet significant, energetic compatibility between otherwise unrelated proteins: more than half of the non-self matches

were significant at the $0.01 < p < 0.001$ level (blue dots in Fig. 6). To investigate the source of this unexpected observation, the energetic compatibility between regions of individual proteins and the rest of the sequence or fold space was quantified.

Negative Contributions Dominate Energetic Compatibility between Sequence and Structure

The most statistically significant best and worst matches of 13 residue fragments were mapped to their locations on the full-length protein, and the densities of the matches were tabulated, as described in Methods. These densities were recorded in two ways: 1) mapping structure fragments to the full-length sequence, and 2) mapping sequence fragments to full-length structure. Thus, the highs and lows of density approximated the average energetic compatibility of a protein's sequence or structure with a representative sample of the entire sequence or structure space. Since these densities were composed of the most extreme energetically compatible and incompatible matches between arbitrary sequences and arbitrary structures of globular proteins, they are referred to as "positive" and "negative" compatibility indices, respectively. In short, the fragment matches revealed regions of full-length proteins likely (or unlikely) to exhibit non-self conformational specificity, due to energetic characteristics shared between other globular proteins.

One example of these compatibility indices is displayed from the perspective of sequence (*i.e.* how a sequence scored in other fold fragments - Fig. 7a) and from the perspective of structure (*i.e.*, how other sequence fragments scored in its fold – Fig. 7b). The variability of indices within an individual protein suggests that energetic compatibility is not uniformly distributed. Also clear is that sequence and structural compatibility indices are asymmetric. In other words, at a given position within an individual protein, the amino acid sequence at that position could have a very different compatibility for other environments than does the environments at that position for other sequences. For example, in labeled regions A, B, and C (Fig. 7), the negative compatibility index between the 1BYQ structure and all other sequences was relatively high, while the negative compatibility index between the sequence at this position and all other structures was low. This means that while the structure at that position does not accommodate many sequences, the sequence that is there, is compatible with many folds. A third observation is that the magnitude of the negative compatibility index is, in general, much greater than the magnitude of the positive compatibility index. In other words, the blue curves in Fig. 7, and in most other proteins, are larger in magnitude than the red curves, consistent with the higher likelihood of obtaining negative random scores in the probability model. No obvious relations between fold type, secondary structure type, location of secondary structure, and the compatibility indices were seen.

It was hypothesized that these indices contained detailed information about energetic compatibility with multiple structures, and thus would provide insight into conformational specificity. To explore this hypothesis, eigenvalue decomposition (principal components analysis) was used to simplify these four-dimensional compatibility indices (Fig. 2). As expected, the first two principal components of the decomposition were dominated by the sequence and structure negative compatibility indices (red circles in Fig. 2), and constituted almost the entire information content (60% + 35% = 95%). Unexpectedly, the

decomposition also revealed a secondary, but substantial, correlation in the patterns of positive and negative compatibility indices, as the coefficients of these quantities are of the same sign and order of magnitude (Fig. 2). Thus, the locations of the largest negative compatibility indices with respect to structure are also often the locations of the largest positive compatibility indices with respect to structure. Examples of this phenomenon can be seen in Fig. 7b, boxes A and B, where the peaks and valleys of both red and blue curves (positive and negative indices, respectively) roughly track each other. In summary, 95% of the information about positive and negative energetic compatibility could be retained by considering only the first two principal components, which are largely due to negative compatibility. Therefore, despite the necessity of a high positive score for one sequence to be conformationally specific for one structure, thermodynamically incompatible regions of sequence and structure largely organize the energetic compatibility, and thus possibly the conformational specificity, of this representative sample of protein fold space.

The trends in Figure 7 were used as the basis for provisionally classifying the susceptibility of a sequence to switch fold (Fig. 7a) or the ability of a fold to accommodate other sequences. (Fig. 7b). Four types of sequence segments were defined (Fig. 3); “permissive”, “selective”, “inactive”, and “gatekeeper” (Fig. 7 a&b – upper bar). Permissive sequence, which accounts for 15% of the total sequence space, is so named because it is highly compatible with other folds, but rarely is it highly incompatible with other folds. In other words, these sequence segments may contribute to stabilizing a fold, but do little to select against other folds. Selective sequences, which at 35% of sequence space, constitutes one of the highest fractions, are those that score very highly in, and are thus highly compatible with, many folds, but are also highly incompatible with other folds. These sequence segments contribute to stabilizing the native fold, but also significantly select against other folds. Inactive sequences are those that appear to not contribute significantly to determining any particular fold and do little to select against any fold. Finally, there are so-called “gatekeeper” sequences that are not compatible with most other folds, and indeed significantly select against many folds, these comprise approximately 15% of sequence space. A similar analysis was performed to categorize the compatibility of fold segments; fractions of gatekeeper and permissive structure were each found to be approximately 11% and fractions of inactive and selective structure were each found to be approximately 39%.

Importantly, all proteins in this representative subset of the human proteome contained variable sized segments of each type of sequence (Fig. 7c) and fold (Fig. 7d) revealing an overall architecture, which indicates that sequence and fold contributions to energetic compatibility are heterogeneously distributed throughout individual proteins. Indeed, the relatively large fractions of inactive sequence and fold segments suggests that the specific folds, which some sequence segments adopt, may be context dependent, lacking significant intrinsic propensity. Ideas such as context dependent sequence propensities have been discussed for the particular case of beta strands, [7] although for these proteins we find no significant correlation between beta sheet and inactive sequence (data not shown).

Protein Size Dependence of Negative Energetic Compatibility

Although the magnitudes of negative *(i.e.* incompatible) and positive (*i.e.* compatible) scoring sequences were comparable, the amount of the incompatibilities was observed to be significantly higher than the amount of compatibilities. The dominance of energetic incompatibility is consistent with the idea of "negative selection" [23, 45–48], *i.e.* that through evolution most other competing folds become thermodynamically incompatible with a particular amino acid sequence. It was hypothesized that the total amount of negative energetic compatibility (*i.e.* the integrated area of the blue curves in Fig. 7) exhibited by a protein, from either sequence or structure, could be related to the widespread non-self energetic compatibility seen in Fig. 6. In other words, does the amount of negative selection scale with the overall ability of a sequence to adopt other folds? To address this question, the *p*-value of the optimal non-self scores in Fig. 6 were plotted against the aggregate energetic incompatibility of each protein (Fig. 8). Significant, though modest, correlations were indeed observed between aggregate negative compatibility and the energetic compatibility between sequence and structure, suggesting that negative selection exerts a significant influence on conformational specificity.

However, an unexpected pattern was observed in these correlations: the relationship between aggregate negative compatibility and energetic compatibility changed sign as a function of protein size (Fig. 9). Longer proteins, such as the 228-residue Hsp90 1BYQ, exhibited a negative correlation between negative compatibility and energetic compatibility (Fig. 8a), while shorter proteins, such as the 96-residue N-terminal domain of amyloid precursor protein 1MWP, exhibited a positive correlation (Fig. 8b). In other words, shorter proteins exhibited increased conformational specificity towards an alternative fold when that alternative fold exhibited increased energetic incompatibility with fragments from all other proteins. The positive correlation reached a maximum value at a protein size of approximately 100 residues (Fig. 9). The fact that the correlation changed sign indicates that both compatibility and incompatibility influence the specificity of proteins of all sizes, but that the relative contribution of incompatibility monotonically decreases with protein length. In other words, the requirement for negative selection appears to be released as sequence length increases.

Discussion

Energetic Incompatibility Influences Protein Conformational Specificity

Two significant insights emerge from these studies, accepting the hypothesis that energetic compatibility is a measure of the degree of conformational specificity. First, conformational specificity of a representative sample of the human proteome, and presumably the entire protein fold space, appears to be organized by energetic incompatibility. This key insight could not be obtained by inspection of the amino acid sequences or ground state structures. Second, protein conformational specificity is a complex function of both the sequence and the fold, with both positive and negative contributions. Just as not all sequence and structure segments contribute equally to protein stability, neither do they contribute equally to conformational specificity. Importantly, although a weak trend exists for more stable regions to exhibit higher negative compatibility index with respect to structure, there is imperfect

correlation between stability and specificity; the most stable regions are not always the most specific nor are the least stable regions always the least specific. Thus, gatekeeper residues located in high-stability regions may be informative determinants of conformational specificity and potential targets for fold-switch engineering. Conversely, mutation or removal of permissive residues may permit increased specificity for a desired fold.

For all proteins, regardless of size, “designing-in” favorable interactions is important for adopting stable structure, a conclusion that can be drawn directly from the highly significant diagonal scores in Fig. 6. Indeed, structure-guided protein engineering has repeatedly employed this idea of “positive design” with much success [49–51]. However, the present analysis suggests that in natural proteins extremely unfavorable interactions in alternative folds (energetic incompatibility) dominate conformational specificity. Furthermore, the intriguing sign switch observed in Fig. 9 reveals that negative selection has an even more pronounced influence for proteins of small size.

Local Structure and Sequence Contributions to Negative Selection

The regions of highest NCI are enriched in proline (Pro) and glycine (Gly) residues (with respect to sequence, Fig. S2a) and are enriched in high-stability (with respect to structure, Fig. S2c). These enrichments suggest that the mechanisms for negative selection can be localized to individual positions of a protein’s structure and sequence: high stability environments, and the amino acids Gly and Pro. Examples include Box A of Fig. 7b, which is a high stability region of 1BYQ that exhibits high NCI with respect to structure, and Box D of Fig. 7a, a region enriched in Gly and Pro that exhibits high NCI with respect to sequence. Conformational restriction and freedom afforded by Pro and Gly side chains, respectively, are likely to be two physical mechanisms for mediating negative selection [52].

Localization of negative selection could guide protein engineering efforts to promote desired, or alternative, structure using targeted Pro or Gly substitutions [53] and core destabilization. The similarity in environmental propensities of negative compatibility and gatekeeper positions (Fig. S2c&d) suggests co-localization of gatekeeper positions, high structural stability, and negative selection. Bearing in mind that mutational effects of sequence and stability changes could be opposing (*e.g.* Fig. 1 indicates that a Pro substitution in a high stability region, intentioned to increase NCI with respect to sequence, is unfavorable and could be destabilizing with respect to structure), such changes might afford a crude tool to introduce or remove specificity. In any event, the analysis presented here provides the locations on each protein where such efforts should be targeted to increase chances for success.

Negative Selection Mediates Protein Domain Evolution

The average domain size of structured proteins is approximately 100 residues [54], and 90% of all known domains are less than 200 residues [55], spanning the size range of proteins sampled here. One implication for protein evolution is that single domain proteins, usually thought of as modular “building blocks” in the organization of larger proteins [56], might be particularly susceptible to sampling alternative structures. Fold-switching is expected to be more prevalent in smaller size proteins [11] and is expected to be less prevalent in larger

size proteins [15]. These expectations are supported by strong positive correlations between experimental thermodynamic stability and size [38, 57], as well as from lattice models that exhibit a larger fraction of alternative minimum energy compact structures as the chain length decreases [58, 59]. Thus, increased negative design would be important for small proteins to preserve structure and function when faced with the constraints of small stability and large numbers of alternative folds.

Consistent with this conjecture is the *SCOP* classification of “small proteins”, whose membership consists of proteins that explicitly require disulfide bonds or metal ions for increased stability [32]. In this scenario, primordial small proteins would have had the tendency for metamorphic behavior, thus negative evolutionary selection would have been a necessary adaptation for dependable metabolic processes mediated by such molecules. Figure 9 suggests that as proteins increase in length they gradually lose the requirement for negative selection. Perhaps this implies the conformational space for a large protein is so vast that preservation of fold is energetically “easier”, as long as aggregation is avoided [60]. Alternatively, larger proteins, which are sometimes composed of several smaller domains, contain functionally important intra-domain interfaces that alter the energetic landscape relative to the individual domains. The role of negative design in larger multi-domain proteins remains to be investigated.

Protein Design Strategy Based on Negative Selection

Aside from effects of negative selection for small proteins, we believe that thermodynamic environments data (Fig. 5) could be practically used as a template for fold design. The log-odds scores (Fig. 1) may be used to generate reasonable amino acid choices for site directed mutagenesis and/or *de novo* design of full-length proteins. A possible advantage of using thermodynamic environments as a design template, as opposed to structural coordinates, is that environments avoid the difficulty of a “frozen approximation” of the backbone [51]. Instead, thermodynamic environments intrinsically incorporate a range of small conformational adjustments that are approximately isoenergetic within the average stability, enthalpy, and entropy of the environment. One possible design strategy, leveraging both the theory and techniques in this paper, would be to simultaneously maximize the positive score of sequence choices for a desired fold target using the log-odds scores while maximizing the negative score of the same sequence against a large library of alternative folds. This strategy could mimic the “energy gap” between desired and alternative structures that has been demonstrated to be useful in protein design [61]. Such a strategy would be computationally fast to implement using the thermodynamics environment data in Table S2. The pursuit of such avenues is currently under way [62].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant R01-GM63747 and NSF grant MCB0446050. Additional support from the Woodrow Wilson Fellowship of Johns Hopkins University to J. H. is gratefully acknowledged. The authors wish to thank Alex Chin for critical reading of the manuscript and for helpful discussions.

References

1. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proceedings of the National Academy of Sciences of the United States of America. 1984; 81:1075–1078. [PubMed: 6422466]
2. Sudarsanam S. Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. Proteins: Structure, Function, and Genetics. 1998; 30:228–231.
3. Guo JT, Jaromczyk JW, Xu Y. Analysis of chameleon sequences and their implications in biological processes. Proteins: Structure, Function, and Bioinformatics. 2007; 67:548–558.
4. Li W, et al. ChSeq: A database of chameleon sequences. Protein Science. 2015; 24(7):1075–1086. [PubMed: 25970262]
5. Murzin AG. Metamorphic proteins. Science. 2008; 320:1725–1726. [PubMed: 18583598]
6. Chang YG, et al. Circadian rhythms. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. Science. 2015; 349(6245):324–8. [PubMed: 26113641]
7. Minor DL Jr, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature. 1996; 380(6576):730–4. [PubMed: 8614471]
8. Alexander PA, et al. From the Cover: A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A. 2009; 106(50):21149–54. [PubMed: 19923431]
9. Cordes MH, et al. Evolution of a protein fold in vitro. Science. 1999; 284(5412):325–8. [PubMed: 10195898]
10. Chen SH, Meller J, Elber R. Comprehensive analysis of sequences of a protein switch. Protein Science. 2015 p. Epub ahead of print. 10.1002/pro2723
11. Porter LL, et al. Subdomain interactions foster the design of two protein pairs with ~80% sequence identity but different folds. Biophysical Journal. 2015; 108(1):154–162. [PubMed: 25564862]
12. Bryan PN, Orban J. Implications of protein fold switching. Current Opinion in Structural Biology. 2013; 23:314–316. [PubMed: 23518177]
13. He Y, et al. Mutational tipping points for switching protein folds and functions. Structure. 2012; 20(2):283–291. [PubMed: 22325777]
14. Eaton KV, et al. Studying protein evolution with hybrids of differently folded homologs. Protein Engineering, Design, and Selection. 2015; 28(8):241–250.
15. Meyerguz L, Kleinberg J, Elber R. The network of sequence flow between protein structures. Proceedings of the National Academy of Sciences USA. 2007; 104(28):11627–11632.
16. Jeffery CJ. Why study moonlighting proteins? Front Genet. 2015; 6:211. [PubMed: 26150826]
17. Copley SD. Moonlighting is mainstream: paradigm adjustment required. Bioessays. 2012; 34:578–588. [PubMed: 22696112]
18. Cao B, Elber R. Computational exploration of the network of sequence flow between protein structures. Proteins: Structure, Function, and Bioinformatics. 2010; 78:985–1003.
19. Allison JR, et al. Current computer modeling cannot explain why two highly similar sequences fold into different structures. Biochemistry. 2011; 50(50):10965–10973. [PubMed: 22082195]
20. Shen Y, et al. De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Science. 2010; 19(2):349–56. [PubMed: 19998407]
21. Hilser VJ, et al. A statistical thermodynamic model of the protein ensemble. Chem Rev. 2006; 106(5):1545–58. [PubMed: 16683744]
22. Lattman EE, Rose GD. Protein folding - what's the question? Proceedings of the National Academy of Sciences, USA. 1993; 90:439–441.
23. Bolon DN, et al. Specificity vs. stability in computational protein design. Proceedings of the National Academy of Sciences, USA. 2005; 102(36):12724–12729.
24. Hilser VJ, Freire E. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J Mol Biol. 1996; 262(5):756–72. [PubMed: 8876652]

25. Pan H, Lee JC, Hilser VJ. Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. Proc Natl Acad Sci U S A. 2000; 97(22):12020–5. [PubMed: 11035796]
26. Babu CR V, Hilser J, Wand AJ. Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. Nat Struct Mol Biol. 2004; 11(4):352–7. [PubMed: 14990997]
27. Whitten ST, Garcia-Moreno EB, Hilser VJ. Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins. Proc Natl Acad Sci U S A. 2005; 102(12):4282–7. [PubMed: 15767576]
28. Liu T, et al. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. Journal of the American Society for Mass Spectrometry. 2012; 23:43–56. [PubMed: 22012689]
29. Wrabl JO, Larson SA, Hilser VJ. Thermodynamic environments in proteins: fundamental determinants of fold specificity. Protein Sci. 2002; 11(8):1945–57. [PubMed: 12142449]
30. Larson SA V, Hilser J. Analysis of the “thermodynamic information content” of a Homo sapiens structural database reveals hierarchical thermodynamic organization. Protein Sci. 2004; 13(7): 1787–801. [PubMed: 15215522]
31. Wang S, et al. Denatured-state energy landscapes of a protein structural database reveal the energetic determinants of a framework model for folding. J Mol Biol. 2008; 381(5):1184–201. [PubMed: 18616947]
32. Andreeva A, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36(Database issue):D419–25. [PubMed: 18000004]
33. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–42. [PubMed: 10592235]
34. Vertrees J, et al. COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures. Bioinformatics. 2005; 21(15):3318–9. [PubMed: 15923205]
35. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins. 1995; 23(4):566–79. [PubMed: 8749853]
36. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? Folding and Design. 1998; 3:141–147. [PubMed: 9565758]
37. Press, WH., et al. Numerical recipes in C: the art of scientific computing. 2. New York: Cambridge University Press; 1992.
38. Robertson AD, Murphy KP. Protein structure and the energetics of protein stability. Chemical Reviews. 1997; 97:1251–1267. [PubMed: 11851450]
39. Wrabl JO V, Hilser J. Investigating homology between proteins using energetic profiles. PLoS Comput Biol. 2010; 6(3):e1000722. [PubMed: 20361049]
40. Bai Y, et al. Thermodynamic parameters from hydrogen exchange measurements. Methods Enzymol. 1995; 259:344–56. [PubMed: 8538461]
41. Munoz V, Serrano L. Helix design, prediction, and stability. Current Opinion in Biotechnology. 1995; 6:382–386. [PubMed: 7579647]
42. Wang Y, Shortle D. Residual helical and turn structure in the denatured state of staphylococcal nuclease: analysis of peptide fragments. Folding and Design. 1997; 2(2):93–100. [PubMed: 9135981]
43. Wrabl JO, Larson SA, Hilser VJ. Thermodynamic propensities of amino acids in the native state ensemble: implications for fold recognition. Protein Sci. 2001; 10(5):1032–45. [PubMed: 11316884]
44. Vertrees J, Wrabl JO, Hilser VJ. An energetic representation of protein architecture that is independent of primary and secondary structure. Biophys J. 2009; 97(5):1461–70. [PubMed: 19720035]
45. Leaver-Fay A, et al. A generic program for multistate protein design. PLoS One. 2011; 6(7):e20937. [PubMed: 21754981]

46. Minning J, Porto M, Bastolla U. Detecting selection for negative design in proteins through an improved model of the misfolded state. Proteins: Structure, Function, and Bioinformatics. 2013; 81(7):1102–1112.
47. Noivert-Brik O, Horovitz A, Unger R. Trade-off between positive and negative design of protein stability: from lattice models to real proteins. PloS Computational Biology. 2009; 5(12):e1000592. [PubMed: 20011105]
48. Berezovsky I, Zeldovich KB, Shakhnovich EI. Positive and negative design in stability and thermal adaptation of natural proteins. PloS Computational Biology. 2007; 3(3):e52. [PubMed: 17381236]
49. Kuhlman B, et al. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003; 302(5649):1364–1368. [PubMed: 14631033]
50. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science. 1997; 278(82–87):82. [PubMed: 9311930]
51. Murphy GS, et al. Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. Structure. 2012; 20(6):1086–1096. [PubMed: 22632833]
52. Creighton, TL. Proteins: Structures and Molecular Properties. 2. New York: W.H. Freeman and Company; 1993.
53. Matthews BW, Nicholson H, Becktel WJ. Enhanced protein thermostability from site-directed mutations that descrease the entropy of unfolding. Proceedings of the National Academy of Sciences of the United States of America. 1987; 84(19):6663–6667. [PubMed: 3477797]
54. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics. 2000; 16(7):613–618. [PubMed: 11038331]
55. Islam SA, Luo J, Sternberg MJ. Identification and analysis of domains in proteins. Protein Engineering. 1995; 8(6):513–525. [PubMed: 8532675]
56. Cesareni, G., et al., editors. Modular Protein Domains. Wiley-VCH; Weinheim, FRG: 2005.
57. Ghosh K, Dill KA. Computing protein stabilities from their chain lengths. Proceedings of the National Academy of Sciences, USA. 2009; 106(26):10649–10654.
58. Camacho CJ, Thirumalai D. Minimum energy compact structures of random sequences of heteropolymers. Physical Review Letters. 1993; 71(15):2505–2508. [PubMed: 10054697]
59. Dill KA, et al. Principles of protein folding - a perspective from simple exact models. Protein Science. 1995; 4:561–602. [PubMed: 7613459]
60. Doye JP, Louis AA, Vendruscolo M. Inhibition of protein crystallization by negative design. Physical Biology. 2004; 1(1–2):9–13.
61. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chemical Reviews. 2006; 106(6):1559–1588. [PubMed: 16683745]
62. Hoffmann J, Wrabl JO, Hilser VJ. Towards the design of metamorphic proteins using ensemble-based energetic information. Biophysical Journal: 2013 Biophysical Society Meeting Abstracts. 2013; (Supplement):2897-Pos.

	W	F	Y	M	L	I	V	A	C	G	P	T	S	N	Q	D	E	H	R	K
TE 1	-0.90	-1.53	-0.93	-0.28	-0.23	0.12	0.20	0.45	-0.29	0.37	0.99	0.15	-0.03	-0.27	-0.19	-0.22	-0.37	-0.18	-0.53	-0.11
TE 2	-0.82	-1.05	-1.09	-0.39	-0.20	-0.24	-0.03	0.05	-0.18	0.66	0.63	0.17	0.05	0.10	-0.25	0.11	-0.07	-0.24	-0.42	0.05
TE 3	-0.83	-1.05	-0.95	-0.11	-0.85	-0.77	-0.60	-0.29	0.36	0.65	-0.42	0.19	0.04	0.54	0.15	0.47	0.41	-0.26	-0.55	0.39
TE 4	-0.03	0.18	0.05	-0.57	0.39	0.56	0.51	0.02	-1.62	-0.25	0.38	0.24	-0.39	-0.47	-0.89	-0.42	-0.33	-0.19	-0.29	-0.12
TE 5	-1.04	-0.50	-0.55	0.07	-0.03	-0.29	-0.10	0.39	-0.09	-0.27	-0.14	-0.16	0.37	0.07	0.22	0.15	-0.12	0.04	0.28	0.01
TE 6	-0.79	-0.66	-0.09	0.18	-0.44	-0.75	-0.63	-0.45	0.66	-0.37	-1.36	-0.11	0.40	0.45	0.44	0.32	0.39	0.08	0.40	0.20
TE 7	0.72	0.98	0.42	0.12	0.55	0.57	0.38	0.15	-0.56	-1.08	-0.49	-0.16	-0.58	-1.11	-0.59	-0.61	-0.47	0.10	-0.44	-0.43
TE 8	1.09	0.58	1.06	0.57	-0.09	-0.21	-0.44	-0.93	0.39	-1.28	-2.00	-0.51	-0.37	-0.26	0.44	-0.41	0.13	0.45	0.66	-0.25

Figure 1. Log-odds compatibility scores relating amino acids to native state ensemble-based thermodynamic environments

These scores were computed as previously described [29–31] using the amino acids and thermodynamic environments data given in Table S2. A positive value indicates that the amino acid is found more often than expected in a particular thermodynamic environment within globular proteins, while a negative value indicates occurrence less often than expected. Colors are identical to those used in Figures 4 and 5, *i.e.* violet, blue, green are lower predicted stability and yellow, orange red are higher stability.

	Positive Compatibility Index with respect to Sequence	Negative Compatibility Index with respect to Sequence	Positive Compatibility Index with respect to Structure	Negative Compatibility Index with respect to Structure	Sqrt(Eigenvalue)	Information Content
Principal Component 1	-0.272	-0.954	-0.001	0.123	3280	60%
Principal Component 2	-0.163	-0.079	-0.120	-0.976	2462	35%
Principal Component 3	-0.948	0.288	-0.011	0.136	976	5%
Principal Component 4	-0.030	-0.007	0.993	-0.116	314	<1%

Figure 2. Principal components analysis of positive and negative energetic compatibilities demonstrates the dominance of incompatibility in a representative sample of globular proteins

Values in the last four columns of Table S2 were subjected to standard eigenvalue decomposition. [37] The vast majority of the information content of the four-dimensional data can be described by the first two principal components, dominated by energetically incompatible sequence and structure indices, respectively, interpreted as effects of negative selection in the organization of protein fold space. Red circles indicate the indices contributing the most to the information content.

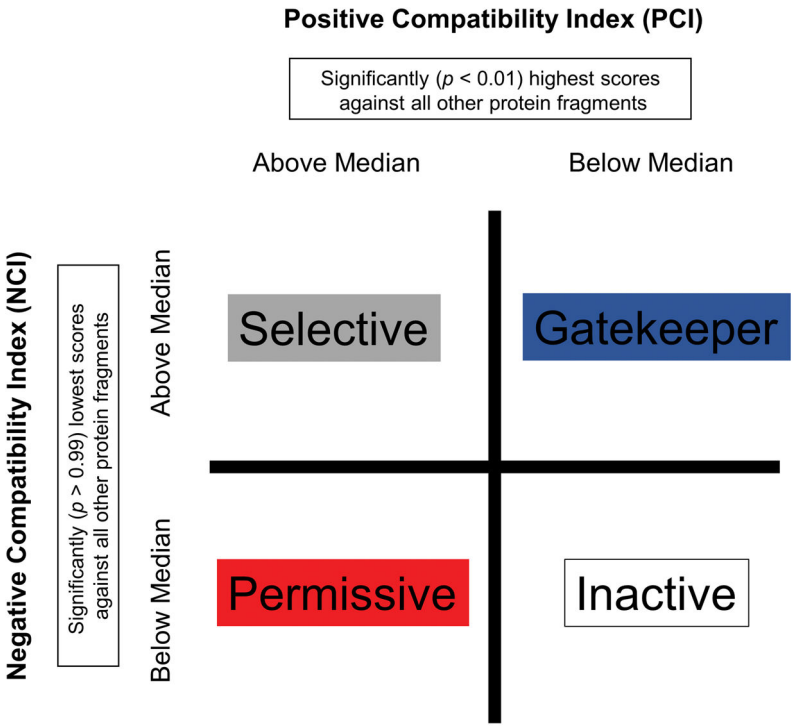


Figure 3. Provisional classification scheme for energetic compatibility indices within proteins

The scheme is a simple contingency table wherein categories are defined based on the median Positive Compatibility Index (PCI) and median Negative Compatibility Index (NCI) for an individual protein. Attributes for the category labels are described in Methods, and the colors correspond to those in Figure 6.

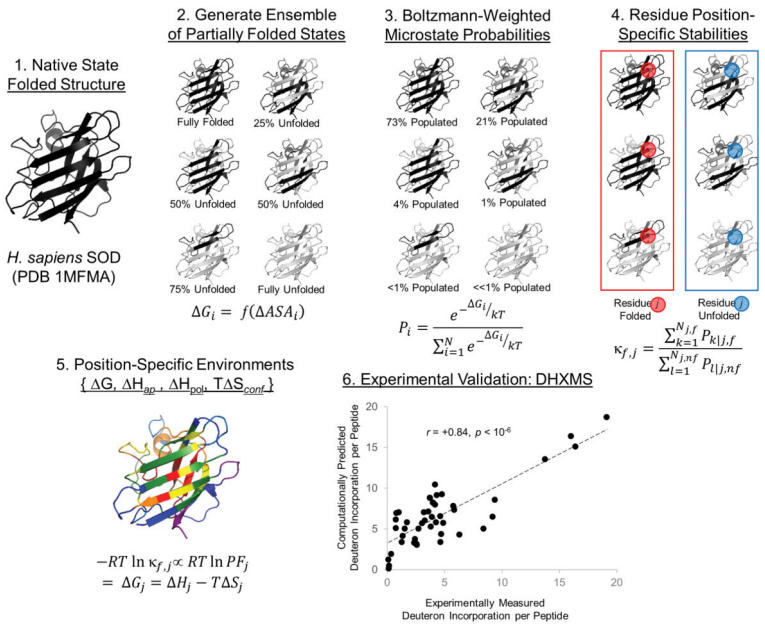
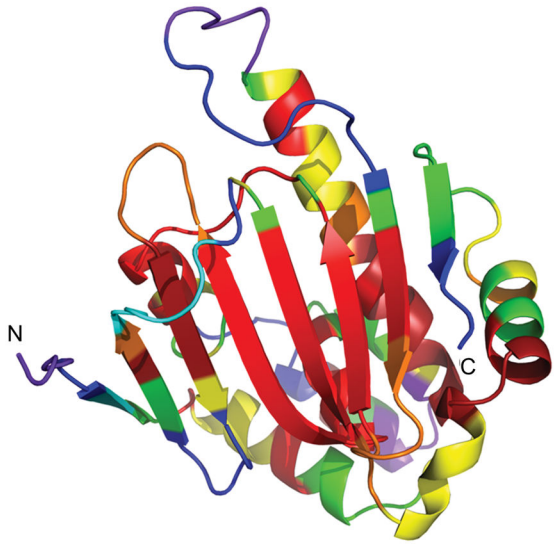


Figure 4. Conceptual basis for native state ensemble-based thermodynamic environments

The human superoxide dismutase (SOD) protein (Step 1) is used as an example for the COREX/BEST algorithm, briefly explained in the main text. An experimentally validated positional thermodynamic stability ΔG_j measured at a residue position *j* in the protein (Steps 4 and 6), is obtained from the Boltzmann-weighted ensemble of partially folded microstates (Steps 2 and 3). Clustering of a large number of positional stabilities from diverse proteins, with respect to the relative contributions of enthalpy and entropy to those stabilities, results in eight colored "thermodynamic environments". These colors correspond to the average Gibbs free energy of the position: purple/blue colors are less stable and orange/red colors are more stable (as displayed in Figure 5). Black regions of the molecular cartoon represent folded, native-like conformations in a greatly simplified COREX ensemble, and gray represents regions of unfolded conformations. Experimental data was obtained from Liu, *et al.* [28]. Abbreviations: ASA = solvent accessible surface area, ap = apolar surface area, pol = polar surface area, conf = conformational, PF = hydrogen exchange protection factor, DHXMS = deuterium – hydrogen exchange mass spectrometry. The thermodynamic environments for this protein are listed in Table S2.



H. sapiens Hsp90
(PDB 1BYQA)

Low Stability → High Stability

	TE 1	TE 2	TE 3	TE 4	TE 5	TE 6	TE 7	TE 8
ΔG	-3.5	-4.4	-6.4	-7.5	-8.5	-9.8	-10.5	-12.4
ΔH_{ap}	4.9	6.3	6.5	10.9	8.8	9.1	14.0	14.2
ΔH_{pol}	-6.1	-8.6	-11.5	-9.4	-12.4	-15.1	-12.5	-16.3
$T\Delta S_{conf}$	-3.1	-4.1	-4.9	-4.1	-4.4	-5.2	-4.6	-5.6

Figure 5. Representation of protein structure in terms of native state ensemble-based thermodynamic environments

Example protein Hsp90 1BYQ from the thermodynamic environments database (top). Residue cartoon color coding corresponds to the average thermodynamic quantities in the environments table (bottom). Values in the table are in units of kcal/mol under simulated folding conditions: 25 °C, pH = 7.0. Rainbow coloring follows the order of average thermodynamic stability: purple, blue, green exhibit lowest stability (least negative ΔG), yellow, orange, red exhibit highest stability (most negative ΔG). The beta-strand core of this protein contains most (but not all) of the highest stability regions, while some (but not all) of the loops and turns are lower in stability.

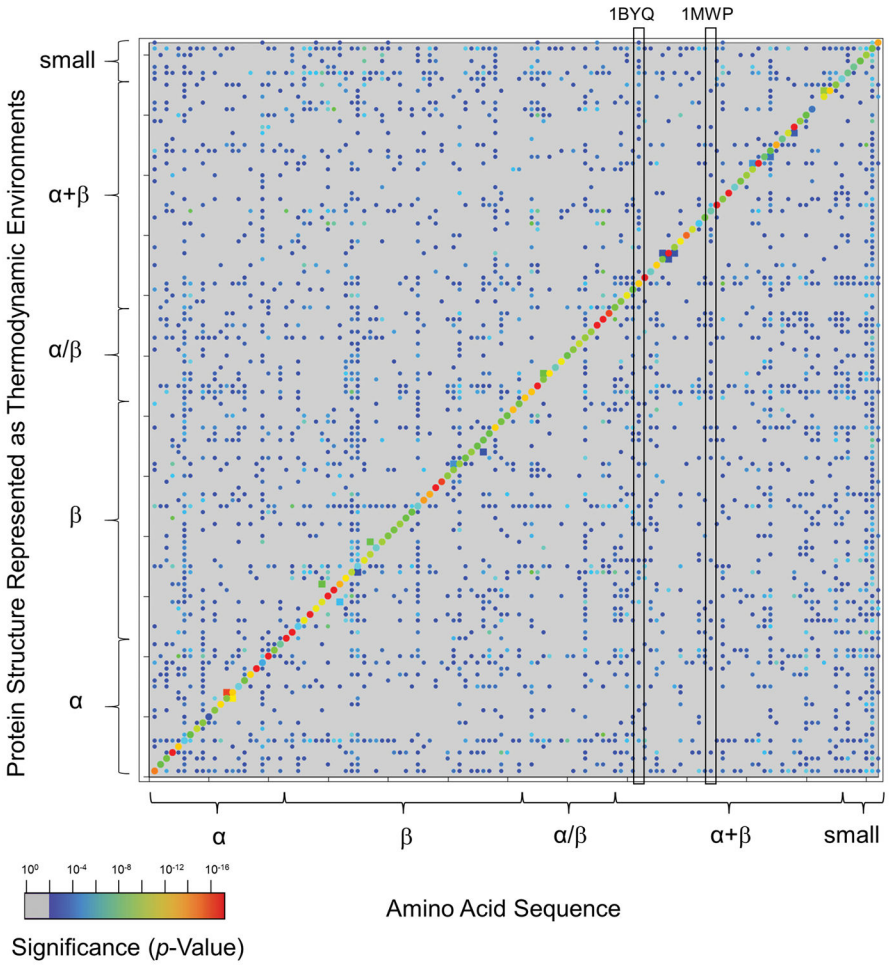


Figure 6. Parameterized random model recapitulates expected sequence-structure conformational specificity as statistically significant

122 *H. sapiens* proteins are listed on each axis in the order given in Table S1. *SCOP* secondary structure classes [32] of each protein are indicated by braces. Dots represent significance levels of either sequence-environment or environment-sequence energetic compatibilities of full length proteins of $p < 0.01$. Rainbow coloring indicates the statistical significance of the energetic scores, with dark blue corresponding to $p \sim 0.01$ and red corresponding to $p \sim 10^{-15}$. The most significant scores are located along the diagonal, corresponding to sequences that are conformationally specific for known structures. Homologous proteins, displayed as squares, also display significant sequence-environment scores. Gray areas, largely off-diagonal, indicate insignificant scores of $p > 0.01$. Unexpectedly, approximately one-half of the off-diagonal points are significant to at least $p = 0.01$. The column locations of two proteins discussed in the text, 1BYQ and 1MWP, are indicated by vertical boxes: the values within these column vectors are plotted as the x-axes in Figs. 8a and 8b.

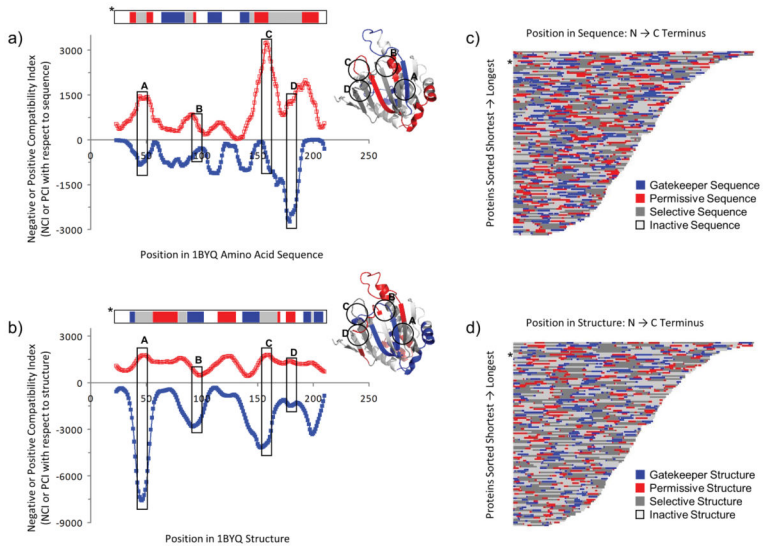


Figure 7. Energetic scoring varies with sequence and structure position, as energetically "compatible" and "incompatible" regions ubiquitous within the proteome

The y-axes in panels a) and b) indicate the number of times any 13-residue fragment from any other protein was significantly compatible with the 1BYQ protein at the residue positions located on the x-axes. Panel a) displays compatible structure fragments with 1BYQ sequence, and panel b) displays compatible sequence fragments with 1BYQ structure. "Significantly" was defined as exhibiting an energetic compatibility of at least $p < 0.01$ (red open squares) or $p > 0.99$ (blue filled squares). For most proteins analyzed, the density of incompatible matches dominated the most compatible matches, suggesting the importance of energetic incompatibility in conformational specificity. Horizontal colored bar above the chart indicates regions of compatibility defined in the text and in Figure 3: "gatekeeper" (blue), "permissive" (red), "selective" (gray), and "inactive" (white); these regions are colored on the molecular cartoon. Labeled vertical boxes A – D denote regions of interest discussed in the text. Panels c) and d) summarize the energetic compatibilities of a representative subset of 122 human proteome amino acid sequences and structures, respectively. Colors are identical to those in panels a) and b) and the locations of the data for the protein displayed in panels a) and b) are indicated by asterisks in panels c) and d), respectively. Panels c) and d) indicate that, for both sequence and structure, total amounts of gatekeeper and permissive regions are less than amounts of inactive and selective regions. Within the sequence and structure of any particular protein, gatekeeper and permissive regions, thought to be important for conformational specificity, are located at different positions.

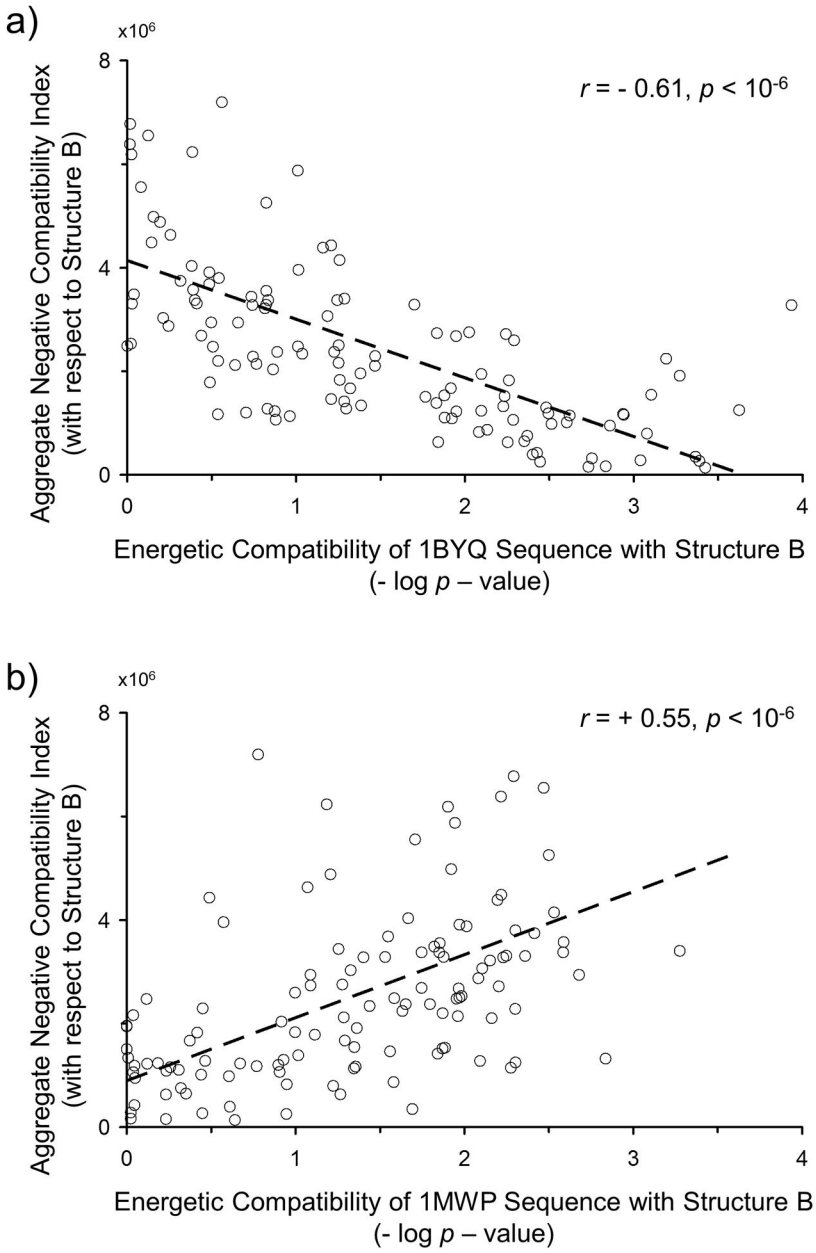


Figure 8. Aggregate negative energetic compatibility of a structure correlates with energetic compatibility of a sequence for that structure

Two protein sequences, 1BYQ (Fig. 8a) and 1MWP (Fig. 8b), are compared with each of 122 structures, the latter represented as native state ensemble-based thermodynamic environments. The *p*-value of the optimal gapless match, computed by the random model described in Fig. S1, is displayed as a log value on the x-axis, negated so that increased energetic compatibility between sequence and structure is represented by a more positive value. The y-axis represents the aggregate negative compatibility of a second protein, examples of which are displayed by the blue curves in Fig. 7. For many proteins studied, modest but significant correlations are observed (Pearson correlation coefficient *r* shown

[37]). Across the entire database of studied proteins, these correlations trend with length: length inversely varies with correlation coefficient: longer proteins such as 1BYQ exhibit negative correlations (Fig. 8a) while shorter proteins such as 1MWP exhibit positive correlations (Fig. 8b). This trend, displayed in Fig. 9, is interpreted as increased importance of negative selection in the conformational specificity of smaller, single-domain proteins.

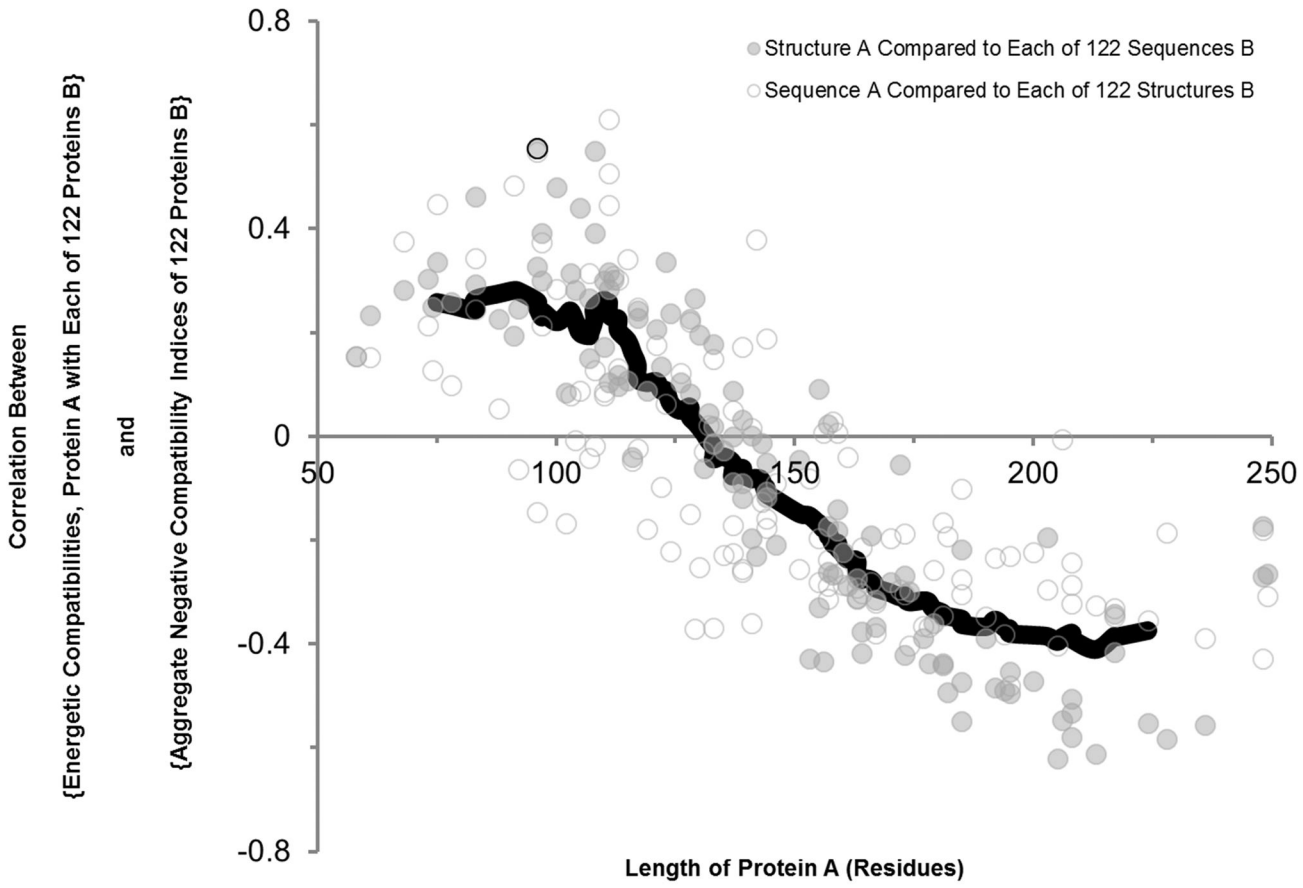


Figure 9. Relationship between energetic compatibility and negative compatibility depends on protein size

Small, single domain proteins exhibit a positive Pearson correlation [37] between negative energetic compatibility and energetic compatibility of sequence with structure. This relationship is interpreted as evidence of the effect of negative selection on conformational specificity. Examples of such correlations are shown in Fig. 8. Open circles indicate aggregate negative compatibility index with respect to structure (as displayed in Fig. 7b), and filled circles indicate aggregate negative compatibility index with respect to amino acid sequence (as displayed in Fig. 7a). The solid dark curve is to guide the eye, a window size 11 moving average over all the data. Energetic compatibilities are expressed as negative log *p*-value, as shown on the x-axes of Fig. 8. The correlation coefficients for proteins 1BYQ and 1MWP shown in Fig. 8 are labeled for reference.