Positive and Negative Design in Stability and Thermal Adaptation of Natural Proteins

Igor N. Berezovsky, Konstantin B. Zeldovich, Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America

The aim of this work is to elucidate how physical principles of protein design are reflected in natural sequences that evolved in response to the thermal conditions of the environment. Using an exactly solvable lattice model, we design sequences with selected thermal properties. Compositional analysis of designed model sequences and natural proteomes reveals a specific trend in amino acid compositions in response to the requirement of stability at elevated environmental temperature: the increase of fractions of hydrophobic and charged amino acid residues at the expense of polar ones. We show that this "from both ends of the hydrophobicity scale" trend is due to positive (to stabilize the native state) and negative (to destabilize misfolded states) components of protein design. Negative design strengthens specific repulsive non-native interactions that appear in misfolded structures. A pressure to preserve specific repulsive interactions in non-native conformations may result in correlated mutations between amino acids that are far apart in the native state but may be in contact in misfolded conformations. Such correlated mutations are indeed found in TIM barrel and other proteins.

Citation: Berezovsky IN, Zeldovich KB, Shakhnovich EI (2007) Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput Biol 3(3): e52. doi:10.1371/journal.pcbi.0030052

Introduction

Despite recent advances in computational protein design [1], there is no complete understanding of basic principles that govern design and selection of naturally occurring proteins [2]. In particular, the physical basis for the ability of proteins to achieve an adaptation to a wide variety of external conditions is still poorly understood. While several attempts to design proteins with a desired fold were successful [1,3], rational design of proteins with desired thermal properties is still an elusive goal. However, Nature apparently succeeds in doing so by "designing" proteins in hyperthermophiles that are stable and functional up to 110 °C. Thus, in the absence of the complete solution of the protein design problem, it is tempting to get clues from Nature as to how thermal properties of proteins can be modulated by proper sequence selection and which physical factors play a role in this process.

A clear manifestation of thermophilic adaptation can be found in a highly statistically significant variation of amino acid compositions of proteomes between meso- and thermophilic organisms [4–7]. Recently, we showed that the total concentration of seven amino acids, I, V, Y, W, R, E, and L, is highly correlated with optimal growth temperature (OGT) of an organism ($R = 0.93$) [8]. The total concentration of IVYWREL combination of amino acids serves as a predictor of OGT with mean accuracy of 8.9 °C [8]. In this work we seek a fundamental theoretical explanation as to why Nature requires an elevated concentration of both hydrophobic and charged amino acids to design hyperthermostable proteins.

Our first goal here is to develop a minimalistic physical model of protein design that could help us to rationalize comparative proteomic analysis of thermo- and mesophiles. A crucial question is how to incorporate the environmental temperature in the model of protein design. Two factors may play a role. The first effect is due to fundamental statistical mechanics of proteins that posit that stable and foldable proteins should have an "energy gap" [9–12]. Specifically, the stability of the native state of a protein is determined by the Boltzmann factor $\exp(-\Delta E/k_B T)$, where ΔE is the energy gap between the native state and lowest energy completely misfolded structures [11–15]. Therefore, to maintain their stability at elevated temperature, the thermophilic proteins should have a greater energy gap. In principle, the increase of the energy gap can be achieved by lowering the energy of the native state (positive design), raising the energy of misfolds (negative design), or both. Another factor that may affect protein thermostability is a possible dependence of fundamental interactions (e.g., hydrophobic forces) on temperature. However, the temperature dependence of different types of interactions may be very complex, and it remains a subject of controversy as to how and to what extent it influences the stability of proteins [16–19]. Our approach to this complex issue is simple: consider first how far one can go based on purely statistical–mechanical analysis of protein thermostability without resorting to explanations based on temperature dependence of various interactions. Specifically, here we use the 27-mer cubic lattice model of proteins [20,21]. The model features 20 types of amino acids that interact when they are nearest neighbors on the lattice; interaction energy depends on types of amino acids involved.

Editor: Andrej Sali, University of California San Francisco, United States of America

Received November 17, 2006; **Accepted** January 31, 2007; **Published** March 23, 2007

A previous version of this article appeared as an Early Online Release on February 1, 2007 (doi:10.1371/journal.pcbi.0030052.eor).

Abbreviations: OGT, optimal growth temperature

* To whom correspondence should be addressed. E-mail: eugene@belok.harvard.edu

Author Summary

What mechanisms does Nature use in her quest for thermophilic proteins? It is known that stability of a protein is mainly determined by the energy gap, or the difference in energy, between native state and a set of incorrectly folded (misfolded) conformations. Here we show that Nature makes thermophilic proteins by widening this gap *from both ends*. The energy of the native state of a protein is decreased by selecting strongly attractive amino acids at positions that are in contact in the native state (positive design). Simultaneously, energies of the misfolded conformations are increased by selection of strongly repulsive amino acids at positions that are distant in native structure; however, these amino acids will interact repulsively in the misfolded conformations (negative design). These fundamental principles of protein design are manifested in the "from both ends of the hydrophobicity scale" trend observed in thermophilic adaptation, whereby proteomes of thermophilic proteins are enriched in extreme amino acids—hydrophobic and charged—at the expense of polar ones. Hydrophobic amino acids contribute mostly to the positive design, while charged amino acids that repel each other in non-native conformations of proteins contribute to negative design. Our results provide guidance in rational design of proteins with selected thermal properties.

The potential is derived from known protein structures and is temperature-independent [22]. For this lattice model all compact conformations can be enumerated [20] and, therefore, exact statistical–mechanical analysis is possible. Previously, protein thermodynamics [9,23], folding [24,25], and evolution [11,26,27] were extensively studied by using this model. We simulate the process of thermal adaptation by the design of 27-mer sequences with selected (at a given environmental temperature T_{env}) thermal properties [14,15]. The algorithm of design (see Methods) carries out simultaneous unrestricted search in conformational, sequence, and amino acid composition spaces. In our analysis we will focus on the amino acid composition of designed sequences as a function of the environmental temperature and we will compare the model findings with amino acid trends in real proteomes. Our main result is that thermal adaptation utilizes both positive and negative design. We show that by increasing the content of amino acids from both extremes of the hydrophobicity scale, thermostable proteins achieve exactly that goal: hydrophobic residues help with positive design while elevated concentration of charged residues helps to achieve stronger negative design. Further, we find an interesting and potentially important aspect of negative design: similar to positive design that strengthens certain native interactions, negative design can make specific non-native interactions strongly repulsive. This, in turn, may lead to emergence of correlated mutations between amino acids that are not in contact in native structure.

Results

We design lattice model proteins with selected thermostability as a first step toward modeling thermal adaptation of organisms. There is a direct connection between OGT (environmental temperature, or T_{env}) of an organism and the melting temperature of its proteins [28,29]. We used the P–design procedure to create model 27-mer sequences that are stable at selected T_{env} (see Methods). We designed sets of 5,000 model proteins for each T_{env} in the range $0.3 < T_{env} < 0.8$ in Miyazawa–Jernigan dimensionless units. The average melting temperature $<T_{melt}>$ of lattice proteins is strongly correlated with T_{env} (see Figure S1) suggesting that the P-design procedure does work. It provides model proteins with desired stability in response to the increase of environmental temperature. The dependence of $<T_{melt}>$ on T_{env} is close to linear and qualitatively matches the empirical linear relationship, $T_{melt} = 24.4 + 0.93\ T_{env}$, between the average living temperature of the organism and melting temperature of its proteins [28].

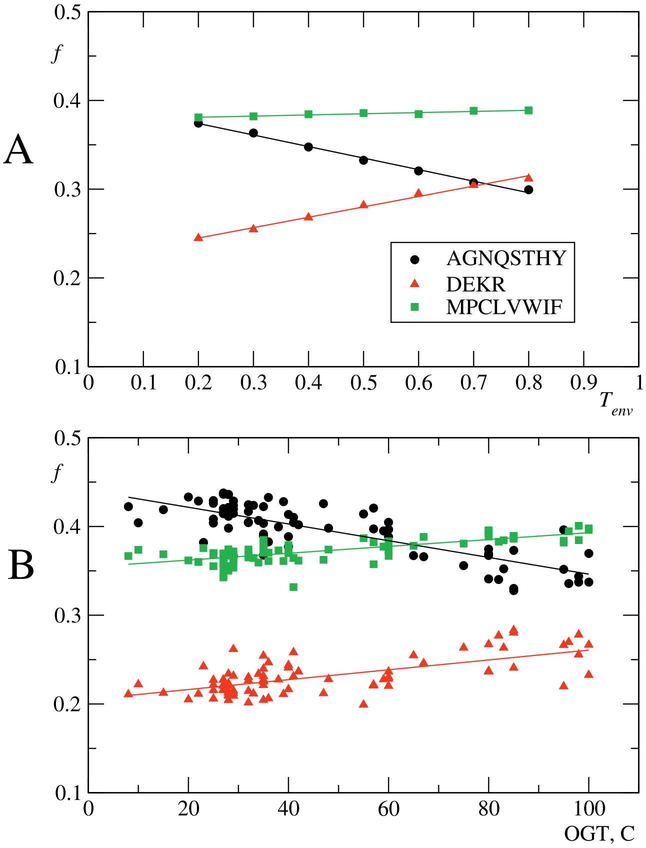


Figure 1. Temperature Dependences of the Fractions of Hydrophobic (LVWIFMPC), Weak Hydrophophobic and Polar (AGNQSTHY), and Charged (DEKR) Amino Acids Plotted Against Temperature T_{env} in the Design Experiment (A) and for Real Proteomes (B)

The data received in P-design procedure applied to sets of 5,000 27-mer sequences with random amino acid composition, using a different value of T_{env} for each set, $0.3 < T_{env} < 0.8$ (T_{env} is measured Miyazawa–Jernigan dimensionless units). Temperature dependences of the fractions of amino acids in natural prokaryotic proteomes is plotted in (B) against OGT of the organism. There are a total of 83 natural proteomes with optimal growth temperatures spanning the interval from −10 to +110 °C.
doi:10.1371/journal.pcbi.0030052.g001

As expected, the amino acid composition of designed proteins does depend on T_{env} for which they were designed. To quantify the differences between "low-temperature" and "high-temperature" amino acid compositions, we plotted temperature dependencies of the fractions of hydrophobic (LVWIFMPC), weak hydrophobic and polar (AGNQSTHY), and charged (DEKR) amino acids for designed lattice proteins (Figure 1A) and natural (Figure 1B) proteomes. Figure 1A shows a significant increase in the amount of charged residues (red triangles) and a slight increase in hydrophobic amino acids (green squares) at the expense of

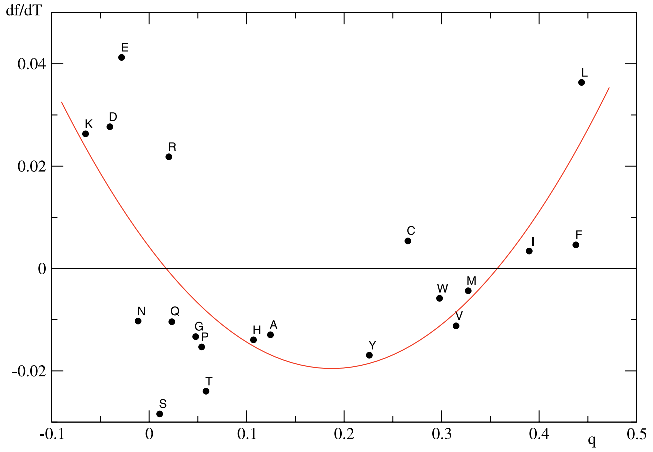


Figure 2. Temperature Derivatives of the Fractions of Amino Acids Plotted against Their Hydrophobicity q (See Methods for Definition of Hydrophobicity Parameter q in Miyazawa–Jernigan Parameter Set)

The temperature derivatives were obtained as slopes from linear regression between the frequency of every one of the 20 amino acids in the designed proteome and T_{env} for which these sequences were designed. The parabolic fit (in red) is to guide the eye.

doi:10.1371/journal.pcbi.0030052.g002

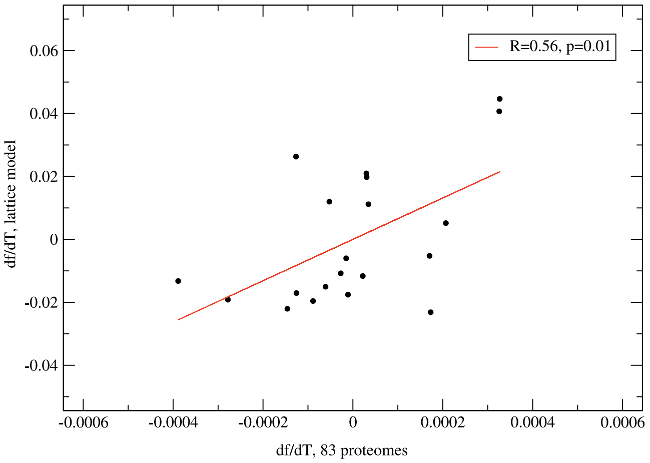


Figure 3. The Scatter Plot Between the Temperature Derivatives of the Fraction of Each of 20 Amino Acids in Designed Lattice Proteins (y-Axis) against the Corresponding Temperature Derivative Calculated over the 83 Natural Proteomes (x-Axis)

The correlation coefficient is $R = 0.56$.

doi:10.1371/journal.pcbi.0030052.g003

polar ones (black dots). Remarkably, the results shown in Figure 1 suggest that increase of thermostability is accompanied by growth of amino acid content from both extremes of the hydrophobicity scale, adding both charged and hydrophobic residues. This observation is further highlighted in Figure 2, which shows—amino acid by amino acid—how compositions of model proteins with T_{env} in designed model proteomes for all 20 amino acids are ranked by their hydrophobicity according to the Miyazawa–Jernigan set of interaction parameters (see Methods and Figure S2A and S2B for more detailed explanation). Figure 2 clearly shows that addition of amino acids to thermophilic model proteomes occurs from the extremes of the hydrophobicity scale while the middle is depressed. The content of charged (Asp, Glu, Lys, Arg; DEKR) and four of the hydrophobic (Ile, Leu, Phe, Cys; ILFC) residues is increased with temperature at the expense of other residues, mostly polar ones. This observation shows that combining amino acids with maximum variance in their hydrophobicity is crucial for creating hyperthermostable model proteins. We refer to this effect as the "from both ends of hydrophobicity scale" trend.

For comparison, we analyzed the variation of amino acid composition in fully sequenced bacterial proteomes (83 species in total, see complete list, Table S1) of psycho-, meso-, thermo-, and hyperthermophilic prokaryotes (habitat temperatures from −10 to +110 °C, see Table S1). Importantly, amino acid composition of 83 natural prokaryotic proteomes reveals similar trends, an increase of the contents of hydrophobic and charged residues, and a decrease of the content of polar ones (Figure 1B). For a more direct comparison of the predictions of our model with the properties of natural proteomes, in Figure 3 we plotted the temperature derivative of the fraction of each of the amino acids in designed lattice proteins against the corresponding temperature derivative calculated over the 83 natural proteomes. The observed positive significant correlation ($R = 0.56$, $p = 0.01$) suggests that generic physical factors captured by this simple statistical–mechanical model played a major role in shaping the amino acid composition patterns across a wide range of habitat temperatures.

We hypothesize that the generic character of the "from both ends" trend that is universally observed in the model and in natural proteins is related to the positive and negative elements of design. In this case, one (hydrophobic) end of the scale is responsible for positive design while another (hydrophilic) end provides negative design. To test this hypothesis, we first studied how the energy gap between the energy of the native state and that of misfolded conformations for the designed model proteins depends on T_{env} (Figure 4). Positive design is the major contributor to the effect (the slope of the temperature-dependent energy decrease of the native state with growth of T_{env} is −5.22; Figure 4, black line), while the increase of the average energy of decoys with T_{env} (slope +1.64; Figure 4, orange line) is pronounced, but less significant. Nevertheless, the results presented in Figure 4 provide clear evidence that negative design works, along with positive design, in the selection of thermostable model proteins.

The findings shown in Figure 4 demonstrate that indeed both positive and negative design act in enhancing thermostability of model proteins. However, the question remains as to how positive and negative design is related to the "from both ends" trend in amino acid compositions, as shown in Figure 2. To address this question, we plot the number of contacts between amino acids whose content grows with T_{env}, according to Figure 2. Figure 5 shows how the average number of contacts (per structure) within both groups of amino acids, FILC and DEKR, in native conformations and in misfolded decoys, depends on T_{env}. Remarkably, we see that in decoy structures the growth of the number of contacts occurs only within the "charged" group DEKR, some of which—according to the Miyazawa–Jernigan potential—repel one another. On the other hand, the number of contacts in the hydrophobic group in decoys do not change despite an overall increase of concentration of these amino acids in sequences designed at higher T_{env}. This result shows that while strongly mutually attractive hydrophobic groups provide

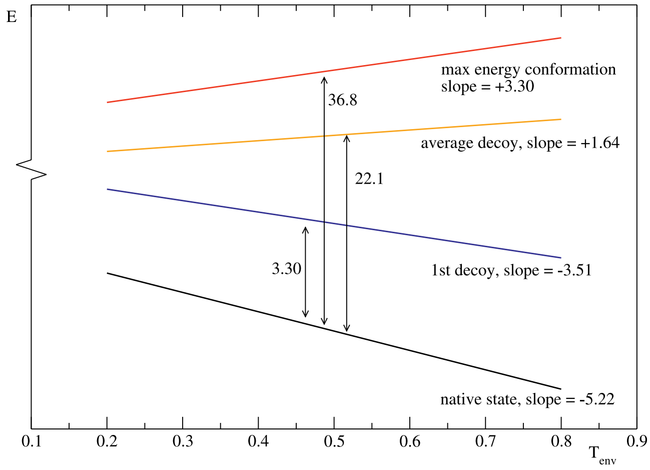


Figure 4. Temperature Dependence of the Contributions of Positive and Negative Design to the Gap in Simulations of Lattice Model Sequences Designed To Be Stable at Various Temperatures

T_{env}, 5,000 sequences were designed at each T_{env}. Positive design results in significant lowering of the native state energy (energy decrease in the interval of temperatures $T_{env} = 0.2/0.8$ has a slope equal to –5.22, black line). Temperature-dependent increase of the decoys' energy is less pronounced, pointing to the specific nature of the negative design. Slopes of the temperature dependences are: –3.57 (blue line) for the first decoy, 1.64 (orange line) for the interaction energies averaged over all decoy structures, and 3.30 (red line) for the maximal energy structure.

doi:10.1371/journal.pcbi.0030052.g004

lower energies of native states for hyperthermophilic model proteins, the growth in concentration of "charged" (DEKR) groups mainly contributes to the negative design factor by raising average energy of misfolded conformations. Remarkably, the average number of contacts between hydrophobic groups (FILC) in misfolded conformations remain roughly the same in meso- and hyperthermophilic model proteins despite significant growth in overall concentration of these groups in hyperthermophiles. Therefore, the data shown in Figure 5 indicate that the "from both ends" trend in amino acid composition is directly related to positive and negative design in stabilization of hyperthermophilic model proteins.

The data presented so far provide insight into averaged (over many model proteins) contributions to the energies of native conformation and decoys. However, a question arises whether negative design works by increasing "average" non-native interactions or by strengthening certain specific repulsive non-native interactions. Indeed, negative design may be based on introducing a few energetically disadvantageous non-native contacts that are persistent in many decoy structures, increasing their energy [2,30]. Therefore, non-native contacts responsible for negative design may well be specific for each sequence, making this effect more detectable if individual proteins are considered.

The exact nature of the lattice model makes a detailed residue-by-residue analysis of the action of both positive and negative design possible. To this end, it is instructive to identify interactions, native and non-native contacts, between residues that play especially important roles in stabilization of the native state and destabilization of decoys. The key idea here is that such important interactions should be conserved in all sequences that fold into a given structure. While identities of amino acids that form such a contact may vary from sequence to sequence, the strength (or energy) of key native or non-native contacts will be preserved: it will be either strongly repulsive or strongly attractive for all sequences that fold into a given structure [31]. Therefore, to identify such key contacts, distributions of energies of native and non-native contacts in multiple sequences that fold into the same native structure should be considered. Such analysis can reveal not only conserved strong native contacts but also possible conserved strong repulsive non-native contacts. To investigate such possibility, we designed 5,000 lattice proteins that all fold into the same (randomly chosen) native structure. To achieve that, we used the design algorithm similar to P-design (see Methods), but for a fixed native structure, and checked a posteriori that the target structure is indeed the native state for all 5,000 sequences. We designed a set of 5,000 mesophilic sequences at $T_{env} = 0.2$ and 5,000 hyperthermophilic sequences that fold to the same structure but are much more stable ($T_{env} = 0.8$).

The concept of native and non-native contacts for our lattice model is illustrated in Figure 6A. It is a cartoon with a zoom-in into the contact matrix of the lattice structure used in simulations. The contact matrix of any compact lattice conformation contains all native (green, total 28 in any structure) and all possible non-native (blue, total 128) contacts.

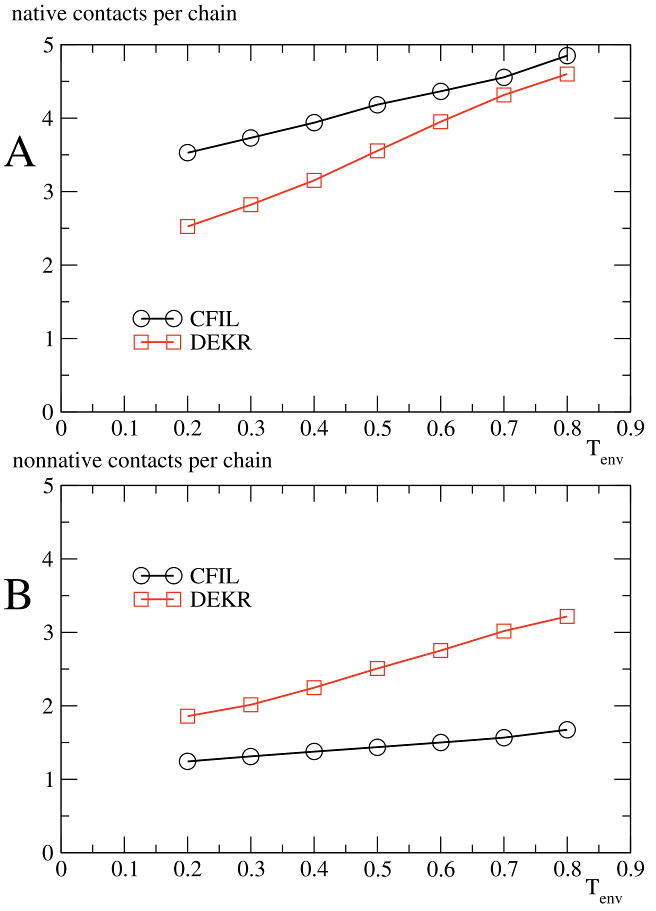
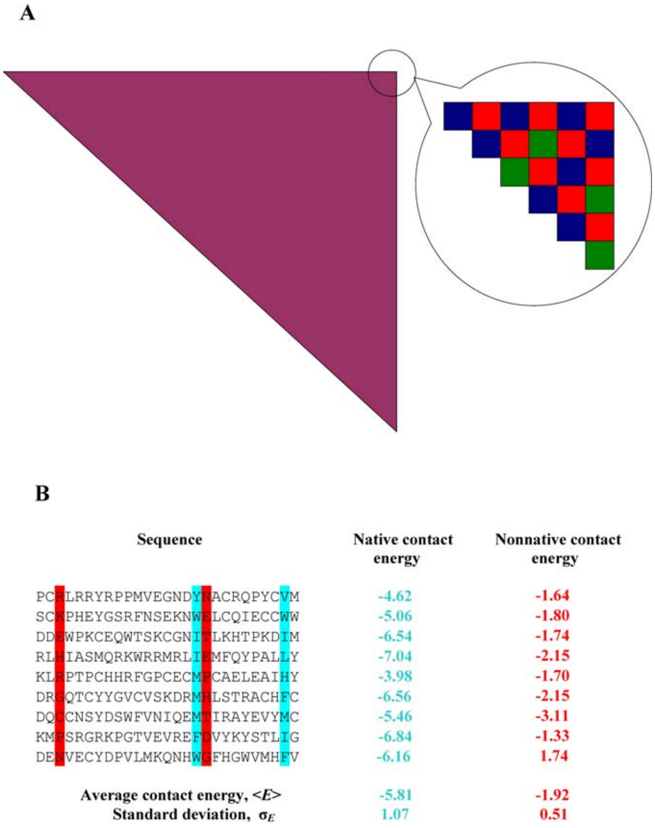


Figure 5. Average Number of Contacts between Amino Acids Whose Fraction Increases in Thermophilic Model Sequences (See Figure 1)

(A) Contacts in native structures. Contacts between "charged" (DEKR) residues are shown in red and contacts between "hydrophobic" (CFIL) residues are shown in black.

(B) Misfolded structures. Color coding is the same as in (A).

doi:10.1371/journal.pcbi.0030052.g005



Sequence	Native contact energy	Nonnative contact energy
PCRLRRYRPPMVEGNDYNACRQPYCVM	-4.62	-1.64
SCKPHEYGSRFNSEKNWELCQIECCWW	-5.06	-1.80
DDEWPKCEQWTSKCGNITLKHTPKDIM	-6.54	-1.74
RLHIASMQRKWRRMRLIEMFQYPALLY	-7.04	-2.15
KLRPTPCHHRFGPCECMPCAELEAIHY	-3.98	-1.70
DRGQTCYYGVCVSKDRMHLSTRACHFC	-6.56	-2.15
DQCCNSYDSWFVNIQEMTIRAYEVYMC	-5.46	-3.11
KMPSRGRKPGTVEVREFDVYKYSTLIG	-6.84	-1.33
DENVECYDPVLMKQNHWGFHGWVMHFV	-6.16	1.74
Average contact energy, <*E*>	-5.81	-1.92
Standard deviation, σ_E	1.07	0.51

Figure 6. Illustration of the Calculation of Energy Dispersion of Native and Non-Native Contacts for the Lattice Model

(A) Zoom-in into contact matrix of one of the 103,346 compact lattice conformations, the cartoon. Even/even, odd/odd, diagonal, (i,i + 1), (i,i + 2) contacts do not exist in a 3 × 3 × 3 lattice (red). Every compact conformation has 28 contacts considered native for this conformation (green). There are also a total of 128 contacts that may appear in alternative conformations; they serve as non-native contacts for this conformation (shown in blue).

(B) Calculation of average energies and dispersion of all native and non-native contacts; illustrative example. Ten aligned sequences all folding into the same shown structure are presented. An example of residues making native (cyan) and non-native (red) contacts is shown. As an illustration, energy of the selected native and non-native contacts are shown for each sequence (this energy is calculated according to the identities of amino acids forming a contact using Miyazawa–Jernigan knowledge-based potential). Average energy of all ten aligned sequences of shown native and non-native contacts and its standard deviation are calculated for illustration here.

doi:10.1371/journal.pcbi.0030052.g006

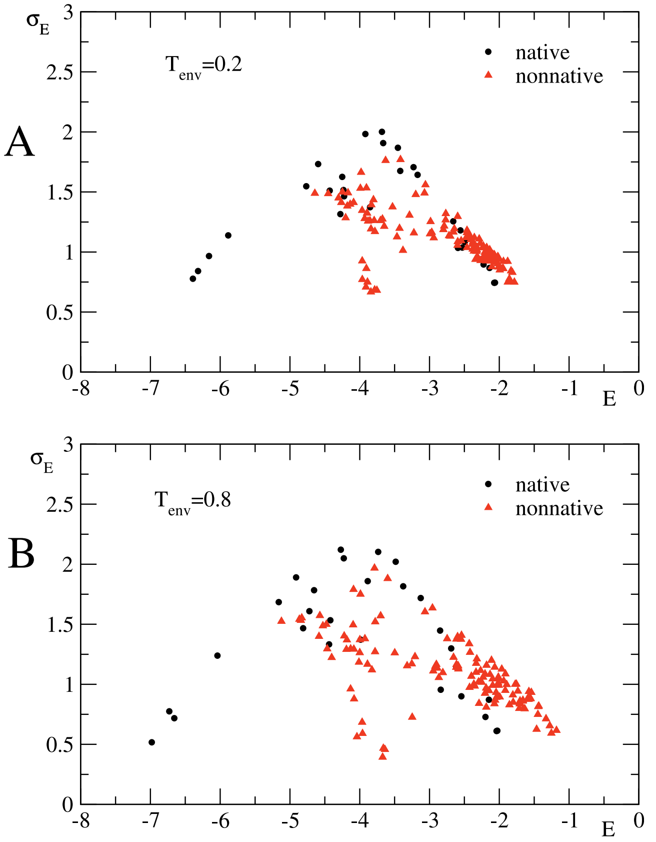


Figure 7. Average Interaction Energies and Their Standard Deviations for All Native (Black Dots) and Non-Native (Red Triangles) Contacts Calculated over 5,000 Sequences Having the Same Native State

(A) Design of "mesophilic" lattice proteins, $T_{env} = 0.2$

(B) Design of "hyperthermophilic" lattice proteins, $T_{env} = 0.8$.

doi:10.1371/journal.pcbi.0030052.g007

All other contacts (red) are prohibited according to the properties of the cubic lattice. To identify important native and non-native contacts whose energies are conserved, we applied the following procedure. First, for each of the 5,000 sequences that fold into selected structure, we calculated energies of 28 native and 128 possible non-native contacts in this structure (using the identities of residues and Myazawa–Jernigan potentials that were employed to design sequences). Next, for each contact we calculated average energy and its standard deviation over all 5,000 designed sequences (see Figure 6B for illustration of this calculation). Contacts whose energy shows a very low standard deviation over all designed sequences are apparently the ones that are most important for stability. This procedure was carried out both for mesophilic sequences ($T_{env} = 0.2$) and for thermophilic sequences ($T_{env} = 0.8$). The results are shown in Figure 7, which presents standard deviation of interaction energies of each native and non-native contact over all 5,000 designed mesophilic sequences (A) and hyperthermophilic sequences (B), plotted against the average (over 5,000 designed sequences) energy of that contact. The plot consists of 28+128=156 points, covering all native and all possible non-native interactions. The native state clearly defines conserved low- and high-energy native contacts (shown in black) in most of the sequences, as the standard deviation is the lowest at the extreme values of the energy. Conserved attractive interactions are in the protein interior, corresponding to the lattice analog of the hydrophobic core; apparently, they emerge due to the action of positive design. The non-native contacts (red dots) follow a different pattern, with only a few conserved attractive interactions, suggesting the diversity of decoy structures. What is surprising to see, however, is that energies of certain most-repulsive (high-energy) non-native contacts show a very low standard deviation, indicating that such contacts may be as important for protein stability as conserved native ones. Comparison of meso- and hyperthermophilic sequences shows clearly that emergence of strong and conserved attractive and repulsive interactions in key native and non-native contacts is directly related to sequence design that generates stable sequences: design of hyperthermostable sequences (Figure 7B) results in stronger and more conserved (lower dispersion of energy) attractive and repulsive specific native and non-native interactions. The only

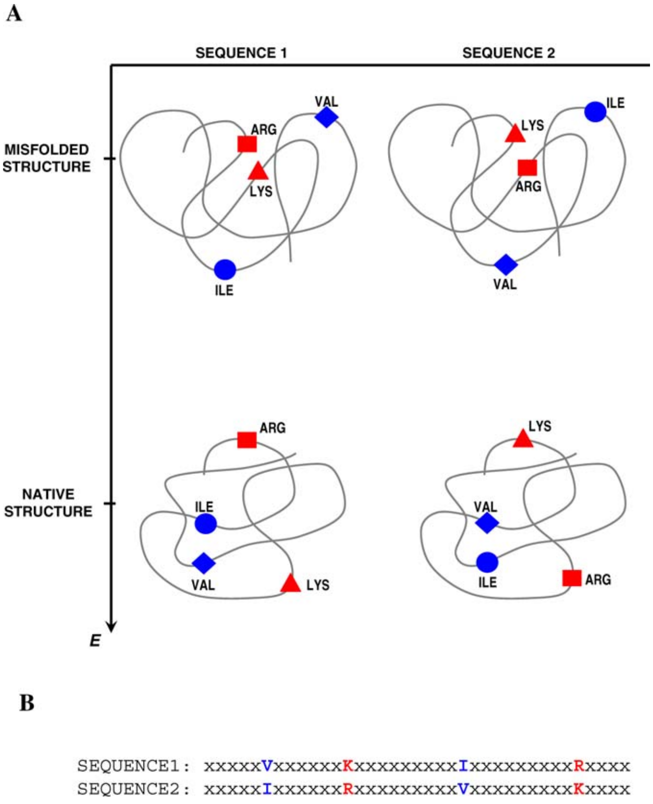

A
SEQUENCE 1
SEQUENCE 2
MISFOLDED STRUCTURE
NATIVE STRUCTURE
E
ARG
LYS
VAL
ILE
B
SEQUENCE1: xxxxxVxxxxxxKxxxxxxxxxIxxxxxxxxxRxxxx
SEQUENCE2: xxxxxIxxxxxxRxxxxxxxxxVxxxxxxxxxKxxxx

Figure 8. Schematic Illustration of the Concept of Mutations by Swaps

(A) A cartoon schematically shows how mutations by swaps preserve the contact energy for some important native (blue) and non-native (red) contacts in Sequence 2, which folds to the same native state as Sequence 1. Higher-energy misfolded structures are also shown schematically for both sequences. Swap of ILE and VAL in Sequence 2 does not change the energy of the native contact between these amino acids in native structure. Repulsion between ARG and LYS residues is also preserved in decoy structure in Sequence 2 if ARG and LYS swap their positions in this sequence as compared with Sequence 1.

(B) Implication of swaps for multiple sequence alignments. Residues that swap in structure appear as substitutions in a multiple sequence alignment.

doi:10.1371/journal.pcbi.0030052.g008

reason that repulsive energies of non-native contacts are conserved is that such contacts persist in certain frequent decoy structures and contribute to the widening of the gap between the native state and decoys. Such repulsive contacts are indirectly (via the sequence) related to a particular native state and are not numerous. Their role may be completely obscured in a "high-throughput" analysis where sequences with different native states are considered together, as in Figure 4. Therefore, we conclude that negative design involves a very specific strategic placement of repulsive contacts in certain decoy structures.

The results and analysis presented in Figure 7 have very important implications for real proteins. The requirement to conserve the energy of key contacts in multiple sequences that fold into the same structure implies that amino acids forming such contacts can mutate in a correlated way, for example by swaps. The observation that mutations may occur often as swaps to preserve specific attractive native and specific repulsive non-native interactions leads to a prediction of a peculiar dependence between frequency of amino acid substitutions (as in, e.g., BLOSUM matrices [32]) and interaction energy between amino acids. Indeed, as illustrated in Figure 8, a correlated mutation in the form of a swap can manifest itself in sequence alignment as a substitution between amino acids that are making the swap. The implication is that frequent substitutions will be observed between amino acids that strongly attract each other (to preserve specific stabilizing native contacts). More interesting and perhaps more surprising, frequent substitutions are also predicted between amino acids that strongly repel each other (to preserve specific non-native repulsive contacts). In other words, we predict that the scatter plot between elements of amino acid interaction energy matrix and substitution matrix will be non-monotonic with maxima at both extremes.

We tested this prediction by plotting the dependence of elements of the substitution matrix BLOSUM62 [32] for 190 pairs of amino acids (synonymous substitutions are excluded) versus their interaction energy as approximated by the knowledge-based Miyazawa–Jernigan potential [22] (Figure 9). This analysis indeed reveals a non-monotonic shape: the parabolic fit in Figure 9A highlights the highly significant non-monotonic nature of the dependence. The striking feature of this dependence is that most frequent substitutions are observed not only between the most attractive amino acids but also between the most repulsive ones. One could argue, however, that the high frequency of substitutions between amino acids that repel each other may be a trivial consequence of conserved substitutions that preserve the charge (R to K and E to D). However, a detailed inspection of the upper right part of the plot in Figure 9A shows that this is not the case (Figure 9B). Indeed, frequent substitutions are observed between mutually repulsive amino acids with vastly different physical–chemical properties and encoded by very dissimilar codons, such as Serine to Asparagine, Glutamin to Arginine, etc. Several highly nonconservative substitutions show about "random"frequencies (element of BLOSUM matrix close to zero, e.g., for Asn to Lys), but this may be due to compensation of two opposite effects: suppression of highly nonconservative substitutions (e.g., that change charge) and facilitation of correlated substitutions such as the ones in the form of swaps as illustrated here.

Use of correlated mutations as predictors of spatial proximity of amino acids in the native structure has been proposed by many authors [33–36]. Indeed, statistical analysis shows that overall correlation between distance between amino acids and degree of correlation in multiple sequence alignments does exist [37]. However sometimes correlated mutations are observed between amino acids that are distant in native structure [33,38]. While sometimes such observations are discarded as false positives in the prediction algorithm [33], our analysis predicts that indeed residues that are distant in structure but may form important repulsive contacts in misfolded conformations may exhibit correlated mutations as illustrated in Figures 8 and 9.

As an illustration of the significance of correlated mutations between amino acids that are far apart in structure, we consider a TIM–barrel fold protein triosephosphate isomerase. Guided by the results of statistical analysis shown in Figure 9, we looked for pairs of residues with strong repulsion according to the Miyazawa–Jernigan potential, random or higher substitution rates between these residues according to the BLOSUM matrix, and highly correlated substitutions of these residues in two positions of the protein

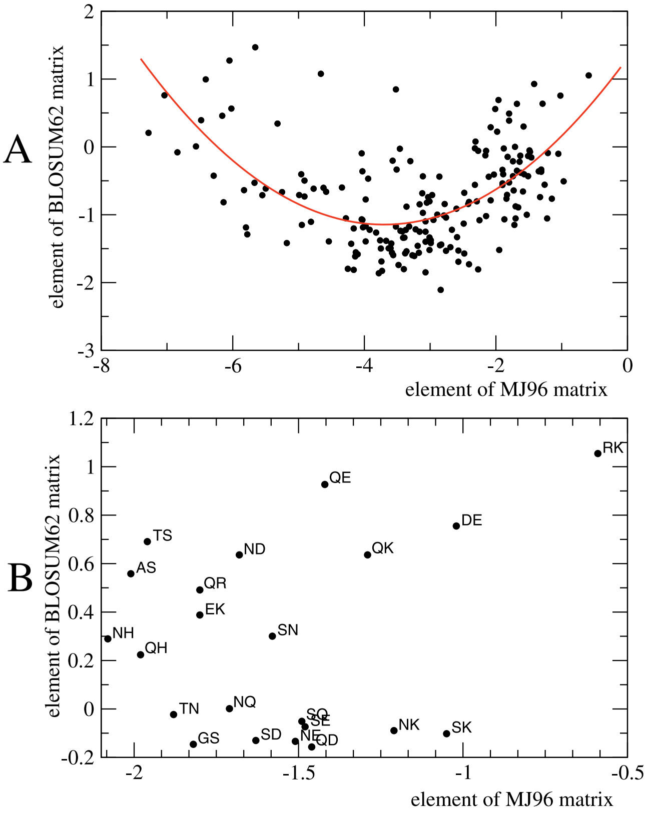


Figure 9. Scatter Plots Showing the Dependence of the Elements of the BLOSUM62 Substitution Matrix on Interaction Energies between Amino Acid Residues (Approximated by Miyazawa–Jernigan Parameters [22])

Only nonsynonymous substitutions are presented.

(A) A complete plot showing all 190 possible nonsynonymous pairs. Red lines represent parabolic fit to highlight nonmonotonic nature of the plot. $R^2 = 0.36$, $p < 10^{-4}$ for the fit and the coefficient at X^2 is 0.18 ± 0.02. An alternative linear fit (unpublished data) is highly insignificant: $R^2 = 0.01$, $p = 0.89$.

(B) Blowup of the right upper corner of (A) with amino acid pairs labeled.

doi:10.1371/journal.pcbi.0030052.g009

sequence in multiple sequence alignment for 7tim (see Methods). To distinguish the effect that we seek from functional conservation, these residues should not be in contact in the native state and should not be involved in the functional site or in the protein–protein interactions

We found correlated substitutions in the sequence of TIM-barrel fold (7tim, chain a), according to the physicochemical characteristic hydropathy [39], by using the CRASP program, which "estimates the contribution of the coordinated substitutions to invariance or variability of integral protein physicochemical characteristics" [40]. Four pairs of residues (Table 1) that have highly coordinated substitutions and repel each other according to the Miyazawa–Jernigan energy matrix were selected. None of those residues belong to the functional site of triosephosphate isomerase, and the protein itself is a single-domain protein not involved into protein–protein interactions [41]. Four pairs of polar and charged residues and one pair of charged residues were identified (see Figure 10A and Table 1). The shortest contact distance (Cα – Cα, 8.7 Å) is between charged Lys(84) and polar Gln(119), which excludes the stabilizing interaction between them in the native structure. We found that correlated mutations between some of these amino acids occur as swaps in the TIM–barrel fold, possibly accompanied by conservative mutation, e.g., surface Lys 120 and Gln 85 in 7tim swap to Gln 120 Lys 85 in *Thermotoga maritima* thermophilic ortholog of triosephosphate isomerase. Even more striking, the Gln85, Lys 213 pair in 7tim (distance in native structure 30 Å) is replaced by Lys 85, Asn 213 in 1b9b (Figure 10B). This pair of residues shows a highly correlated substitution pattern in TIM–barrel multiple sequence alignment despite the fact that these are very distinct amino acids.

Table 1. Most Significantly Correlated Substitutions in Triosephosphate Isomerase (7tim, Chain A)

Pair of Amino Acid Residues	Correlation Coefficient, R	Distance in the Structure, Å
Lys(84)–Gln(119)	0.33	8.7
Lys(84)–Asn(213)	0.51	29.3
Gln(119)–Glu(133)	0.42	33.9
Lys(134)–Ser(202)	0.32	26.1

doi:10.1371/journal.pcbi.0030052.t001

Discussion

Stabilization of thermophilic proteins is achieved by negative and positive design working together, i.e., the gap "opens" from both sides, decreasing energy of the native state and at the same time increasing the energy of misfolded conformations. This factor is responsible for the "from both ends of hydrophobicity scale" trend observed in model and real [8] thermophilic proteomes. In particular, our recent analysis of complete bacterial proteomes [8] revealed that proteomes of thermophilic bacteria are enriched in both hydrophobic residues (IVYLW) and charged ones (ER), while all polar residues are suppressed. Discrepancies between different hydrophobicity scales [42], the statistical nature of knowledge-based Miyazawa–Jernigan potential [22], and limitations of the lattice model make it impossible to quantitatively compare the content of individual amino acids in lattice and natural proteomes or exactly predict the amino acid composition of thermophilic proteomes with very high accuracy from lattice model calculations. Nevertheless our lattice calculations are in semiquantitative agreement with data on natural proteomes, (see Figures 1 and 3) and exhibit the same "from both ends of hydrophobicity scale" trend in amino acid composition adaptation in response to elevated habitat temperature.

The knowledge-based Miyazawa–Jernigan potentials, derived from native structures of proteins, are certainly a crude approximation to real protein energetics [43]. A question arises as to whether our observations are generic or are due to the specific potential used to design model proteins. A detailed comparison of several potentials—all atom and group-based derived by different methods—was carried out recently in our lab [44]. Remarkably, we found that despite differences in detail all these potentials reflect the same dominant contributions to protein stabilization. It appears that dominant contributions to energy gaps in proteins come principally from two types of interactions: hydrophobic interactions and electrostatics [44]. Further, it was found that knowledge-based potentials derived using structures of

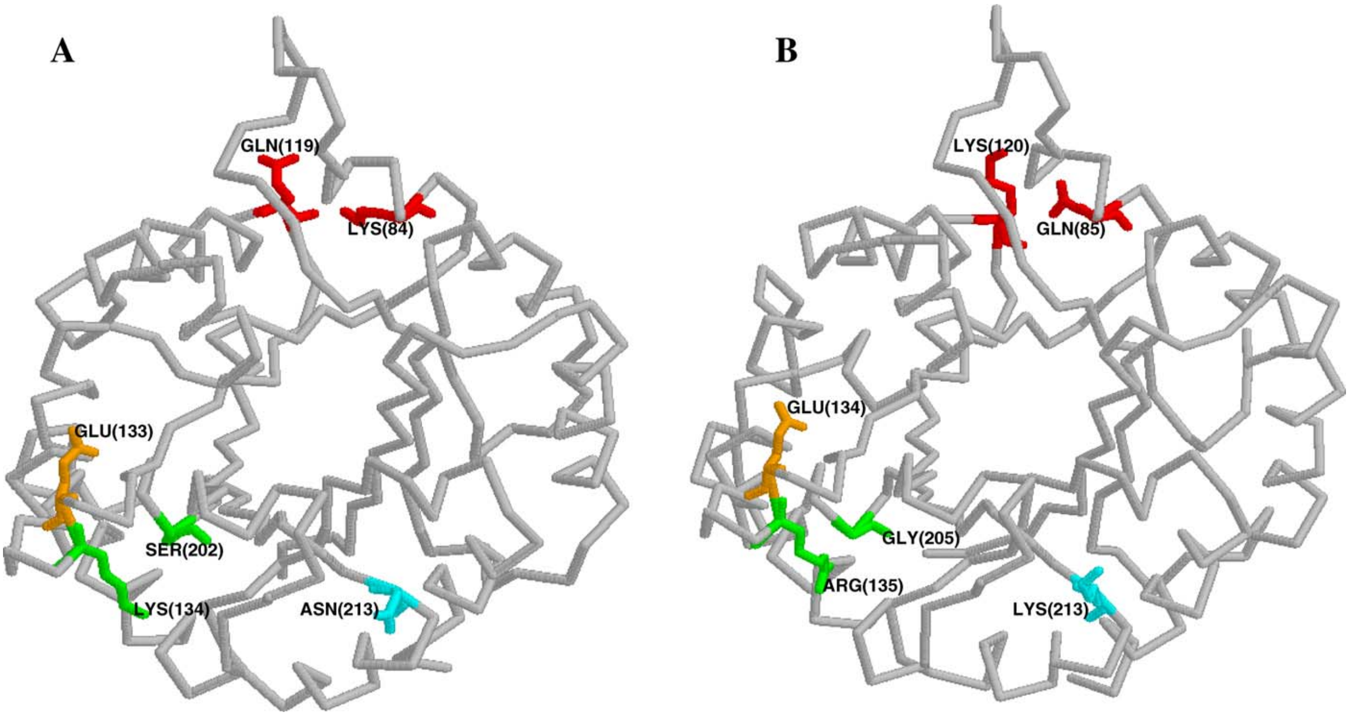


Figure 10. Correlated Mutations and Swaps in Representatives of the TIM–Barrel Fold

(A) Four pairs of surface residues exhibiting correlated mutations in triosephosphate isomerases from *Saccharomyces cerevisiae* (7tim, chain a): Lys84(red)–Gln119(red), Lys84(red)–Asn213(cyan), Gln119(red)–Glu133(orange), and Lys134(green)–Ser202(green).

(B) Comparison of *S. cerevisiae* triosephosphate isomerase with triosephosphate isomerase from *T. maritima* (shown here: 1b9b, chain a) reveals two swaps of surface amino acid involved into correlated mutations in triosephosphate isomerase: Gln85(red)–Lys120(red) and Gln85(red)–Lys(213). These amino acids do not interact in the native structure of either molecule. Positions are numbered according to the 1b9b sequence aligned with 7tim.

doi:10.1371/journal.pcbi.0030052.g010

meso- and thermophilic proteins are virtually indistinguishable (KZ and ES, unpublished data). These observations suggest that the "from both ends of hydrophobicity scale" trend observed in model calculations and in real proteome is a robust phenomenon, reflecting basic physical principles of protein design, rather than a consequence of a specific potential set used in calculations.

While positive design [45] is universally used in experiments, the role and omnipresence of negative design are still under discussion [46]. The main challenge in the study of negative design stems from the difficulties in the modeling of relevant misfolded conformations and energetic effects of mutations that destabilize them [46]. It was shown that charged residues can be effectively used in negative design [30]. Another indirect evidence of the contribution of charged residues to negative design emerges from site-directed mutagenesis, where mutations of polar groups to charged ones on the surface of a protein lead to protein stabilization even in the absence of salt–bridge partners of the mutated group [47–49]. In a series of experiments [47,48,50], surface electrostatic interactions were shown to provide a marginal contribution to stability of the native structure, hence their possible importance for making unfavorable high-energy contact in decoys. An alternative view, proposed recently by Makhatadze et al., suggests that long-range electrostatic interactions may contribute to stability of the native state [51]. However, at normal physiological conditions the range of electrostatic interactions is limited due to Debye screening and hardly exceeds 8 Å. Our simulations and proteomic analysis point to a possible role of some surface charged residues as contributing to destabilization of misfolded structures through a negative design mechanism.

Positive and negative elements of design affect the evolution of protein sequences. The dependence of substitution rates in sequences of natural proteins (BLOSUM62 substitution matrix) on interaction energies according to knowledge-based Miyazawa–Jernigan potential has a peculiar nonmonotonic shape showing elevated substitution rates between residues that attract each other as well as between residues that repel each other. The physical reason for this phenomenon is the same as for the "both ends of hydrophobicity scale" trend: simultaneous action of positive and negative design. Upon substitutions, energy of attractive contacts in native states should be preserved as well as energies of specific repulsive contacts in misfolded conformations. Apparently both these factors act in concert to preserve the energy gap in proteins.

Our study deepens an understanding of correlated mutations in proteins. With regard to native contacts, the fact that amino acids making strongly attractive native interactions should exhibit correlated mutations had been realized long ago. Several authors proposed to use correlated mutations as a tool to determine possible native contacts from multiple sequence alignment [33–36]. However, this suggestion is complicated by the observation that correlated mutations are often found between residues that have no obvious functional role and are distant in structure [33,38,52,53]. Using the double mutant technique, Horovitz et al. [33] suggested a relation between correlated mutations and energetic connectivity (i.e., nonadditivity of stability effects in double mutation cycles) between corresponding amino acids. Green and Shortle [54] showed that amino acids that are distant in structure may indeed be "energetically coupled," attributing this effect to influence of mutations on the unfolded state of proteins, consistent with our findings. Lockless and Ranganathan [38] suggested that a "pathway of energetic connectivity" exists between distant residues that exhibit correlated mutations. Fodor and Aldrich [37], however, examined several other proteins and argued against the "general principle of isolated pathways of evolutionarily conserved energetic connectivity in proteins." Here we show that negative design that destabilizes misfolded conformations of proteins may be responsible for correlated mutations between residues that are far apart in native structures.

In this work, we developed a simple exact model of thermophilic adaptation and discovered fundamental statistical–mechanical rules that Nature uses in her quest to enhance protein stability. While many other factors, including dependence of hydrophobic and other interactions on temperature, certainly play a role in protein stabilization, the action of positive and negative design found and described here in a minimalistic model appears to be a basic universal principle determining evolution of sequences of thermostable proteins. A better understanding of fundamental principles of protein design and stability makes it possible to decipher peculiar signals that emerge in the analysis of meso- and thermophilic genomes and proteomes [8] and in many studies of correlated mutations in proteins [33,35,53].

Materials and Methods

We use the standard lattice model of proteins as compact 27-unit polymers on a $3 \times 3 \times 3$ lattice [20]. The residues interact with each other via the Miyazawa–Jernigan pairwise contact potential [22]. It is possible to calculate the energy of a sequence in each of the 103,346 compact conformations allowed by the $3 \times 3 \times 3$ lattice, and the Boltzmann probability of being in the lowest energy (native) conformation,

$$P_{nat}(T_{env}) = \frac{e^{-E_0/T_{env}}}{\sum\limits_{i=0}^{103345} e^{-E_i/T_{env}}},$$

where E_0 is the lowest energy among the 103,346 conformations, and T_{env} is the environmental temperature. The melting temperature T_{melt} is found numerically from the condition $P_{nat}(T_{melt}) = 0.5$. Note that if the energy spectrum E_i is sparse enough at low energies, the value of P_{nat} is determined chiefly by the energy gap $E_1 - E_0$ between the native state and the closest decoy structure that has no structural relation to the native state.

To design lattice proteins, we use here a Monte-Carlo procedure (P-design, [14,15]) that maximizes the Boltzmann probability $\mathrm{P_{nat}}$ of the native state by introducing mutations in the amino acid sequence and accepting or rejecting them according to the Metropolis criterion. As this procedure takes the environmental temperature $\mathrm{T_{env}}$ as an input physical parameter, and generates amino acid sequences designed to be stable at $\mathrm{T_{env}}$, it is an obvious choice for modeling the thermophilic adaptation.

Initially, the sequence is chosen at random; the frequencies of all amino acid residues in the initial sequences are equal to 5%. At each Monte-Carlo step, a random mutation of one amino acid in a sequence is attempted and P_{nat} of the mutated protein is determined. The native structure is determined at every step of the simulation; generally, the native state changes upon mutation of the sequence. If the value of P_{nat} increased, the mutation is always accepted; if P_{nat} decreased, the mutation is accepted with the probability $\exp[-(P_{nat}(\text{old}) - P_{nat}(\text{new}))/p]$, with $p = 0.05$ (a Metropolis-like criterion). We chose $p = 0.05$ so that the average melting temperature of designed proteins is higher than the environmental temperature (see Figure S2), in agreement with experimental observations [29,55]. The design procedure is stopped after 2,000 Monte-Carlo iterations. Such length of design runs is sufficient to overcome any possible effects of the initial composition of the sequences, so the amino acid composition of the designed sequences depends only on the environmental temperature T_{env}.

To relate the trends in amino acid composition with the physical properties and interaction energies of individual amino acids, we use hydrophobicity as a generic parameter characterizing an amino acid [42]. To characterize the hydrophobicity of amino acids in the simulations, we make use of the fact that the Miyazawa–Jernigan interaction energy matrix is very well approximated by its spectral decomposition [43]. Interestingly, it is sufficient to use only one eigenvector q, corresponding to the largest eigenvalue, so the interaction (contact) energy E_{ij} between amino acids of types i and j reads $E_{ij} \approx E_0 + \lambda q_i q_j$ [43]. In this representation, hydrophobic residues have the largest values of q, while hydrophilic (charged) residues correspond to small q.

All sequences of TIM–barrel folds with length less than 300 amino acid residues were extracted according to the SCOP database description [56]. Identical sequences were excluded from further consideration. Remaining sequences (total 39) were aligned against the sequence of the triosephosphate isomerase (7tim.pdb, chain a) by using Kalign Web-server for multiple alignment of protein sequences (http://msa.cgb.ki.se/cgi-bin/msa.cgi, [57]).

Correlated substitutions in the multiple alignments were determined by using the CRASP program (http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp, [40]). The CRASP program gives the correlation coefficient between the values of physicochemical parameters at a pair of positions of sequence alignment. We chose hydropathy [39] as a physicochemical characteristic appropriate for establishing correlated mutations of interest. Only significant correlations, with the correlation coefficient higher than the critical threshold (0.311), were considered.

The complete genomes were downloaded from the National Center for Biotechnology Information Genome database at http://www.ncbi.nih.gov/entrez/query.fcgi?db=Genome (see Table S1).

Supporting Information

Figure S1. Average Melting Temperature of Designed Lattice Proteomes (5,000 Sequences Each) Depending on the Environmental Temperature T_{env} Entering the P-Design Procedure (2,000 Monte-Carlo Mutation Steps to Generate Each Sequence)

Found at doi:10.1371/journal.pcbi.0030052.sg001 (82 KB DOC).

Figure S2. Temperature Dependences of Amino Acid Fraction for Val (A) and Glu (B) in 204 Natural Psycho-, Meso-, Thermo-, and Hyperthermophilic Proteomes (Habitat Temperatures from –10 °C to +110 °C

See Table S1 for optimal growth temperatures and references.

Found at doi:10.1371/journal.pcbi.0030052.sg002 (294 KB DOC).

Table S1. Prokaryotes with Completely Sequenced Genomes and Their Optimal Growth Temperatures

The columns are: NN, number; Organism, name of the organism; OGT, optimal growth temperature, °C; Source, source of the optimal growth temperature.

Found at doi:10.1371/journal.pcbi.0030052.st001 (203 KB DOC).

Accession Numbers

The accession numbers from the Protein Data Bank (http://www.rcsb.org/pdb/) used in this paper are: TIM–barrel fold protein triosephosphate isomerase (7tim); *T. maritima* thermophilic ortholog of triosephosphate isomerase (1b9b).

Acknowledgments

INB was supported by the Merck Fellowship for Genome-related research. We thank George Makhatadze for useful correspondence.

Author contributions. INB and EIS conceived and designed the experiments and wrote the paper. INB and KBZ performed the experiments. INB, KBZ, and EIS analyzed the data.

Funding. This work is supported by the US National Institutes of Health.

Competing interests. The authors have declared that no competing interests exist.

References

1. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302: 1364–1368.
2. Butterfoss GL, Kuhlman B (2006) Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 35: 49–65.
3. Dahiyat BI, Mayo SL (1997) De novo protein design: Fully automated sequence selection. Science 278: 82–87.
4. Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem 275: 32383–32386.
5. Chakravarty S, Varadarajan R (2002) Elucidation of factors responsible for enhanced thermal stability of proteins: A structural genomics based study. Biochemistry 41: 8152–8161.
6. Das R, Gerstein M (2000) The stability of thermophilic proteins: A study based on comprehensive genome comparison. Funct Integr Genomics 1: 76–88.
7. Kreil DP, Ouzounis CA (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 29: 1608–1615.
8. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3: e5.
9. Shakhnovich EI, Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346: 773–775.
10. Sali A, Shakhnovich E, Karplus M (1994) How does a protein fold? Nature 369: 248–251.
11. Shakhnovich E (2006) Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. Chem Rev 106: 1559–1588.
12. Shakhnovich EI, Gutin AM (1993) Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci U S A 90: 7195–7199.
13. Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1992) Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci U S A 89: 4918–4922.
14. Morrissey MP, Shakhnovich EI (1996) Design of proteins with selected thermal properties. Fold Des 1: 391–405.
15. Seno F, Vendruscolo M, Maritan A, Banavar JR (1996) Optimal protein design procedure. Phys Rev Lett 77: 1901–1904.
16. Makhatadze GI, Privalov PL (1990) Heat capacity of proteins. I. Partial molar heat capacity of individual amino acid residues in aqueous solution: Hydration effect. J Mol Biol 213: 375–384.
17. Makhatadze GI, Privalov PL (1993) Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration. J Mol Biol 232: 639–659.
18. Prabhu NV, Sharp KA (2005) Heat capacity in proteins. Annu Rev Phys Chem 56: 521–548.
19. Privalov PL, Makhatadze GI (1993) Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. J Mol Biol 232: 660–679.
20. Shakhnovich E, Gutin A (1990) Enumeration of all compact conformations of copolymers with random sequence of links. J Chem Phys 93: 5967–5971.
21. Sali A, Shakhnovich E, Karplus M (1994) Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J Mol Biol 235: 1614–1636.
22. Miyazawa S, Jernigan RL (1996) Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256: 623–644.
23. Pande VS, Grosberg AY, Tanaka T (1994) Thermodynamic procedure to synthesize heteropolymers that can renature to recognize a given target molecule. Proc Natl Acad Sci U S A 91: 12976–12979.
24. Klimov DK, Thirumalai D (1996) Criterion that determines the foldability of proteins. Phys Rev Lett 76: 4070–4073.
25. Shakhnovich EI (1994) Proteins with selected sequences fold into unique native conformation. Phys Rev Lett 72: 3907–3910.
26. Govindarajan S, Goldstein RA (1996) Why are some proteins structures so common? Proc Natl Acad Sci U S A 93: 3341–3345.
27. Taverna DM, Goldstein RA (2000) The distribution of structures in evolving protein populations. Biopolymers 53: 1–8.
28. Gromiha MM, Oobatake M, Sarai A (1999) Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 82: 51–67.
29. McFall-Ngai MJ, Horwitz J (1990) A comparative study of the thermal stability of the vertebrate eye lens: Antarctic ice fish to the desert iguana. *Exp Eye Res* 50: 703–709.
30. Summa CM, Rosenblatt MM, Hong JK, Lear JD, DeGrado WF (2002) Computational de novo design, and characterization of an A(2)B(2) diiron protein. J Mol Biol 321: 923–938.
31. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics, and function. J Mol Biol 291: 177–196.
32. Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. Proteins 17: 49–61.
33. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309–317.
34. Horovitz A, Bochkareva ES, Yifrach O, Girshovich AS (1994) Prediction of an inter-residue interaction in the chaperonin GroEL from multiple sequence alignment is confirmed by double-mutant cycle analysis. J Mol Biol 238: 133–138.
35. Noivirt O, Eisenstein M, Horovitz A (2005) Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel 18: 247–253.
36. Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2: S25–S32.
37. Fodor AA, Aldrich RW (2004) On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 279: 19046–19050.
38. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.
39. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
40. Afonnikov DA, Kolchanov NA (2004) CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. Nucleic Acids Res 32: W64–W68.
41. Davenport RC, Bash PA, Seaton BA, Karplus M, Petsko GA, et al. (1991) Structure of the triosephosphate isomerase–phosphoglycolohydroxamate complex: An analogue of the intermediate on the reaction pathway. Biochemistry 30: 5821–5826.
42. Bastolla U, Porto M, Roman HE, Vendruscolo M (2005) Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles. Gene 347: 219–230.
43. Li H, Tang C, Wingreen NS (1997). Nature of driving force for protein folding: A result from analyzing the statistical potential Phys Rev Lett 79: 765–768.
44. Chen WW, Shakhnovich EI (2005) Lessons from the design of a novel atomic potential for protein folding. Protein Sci 14: 1741–1752.
45. Shifman JM, Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc Natl Acad Sci U S A 100: 13274–13279.
46. Bolon DN, Grant RA, Baker TA, Sauer RT (2005) Specificity versus stability in computational protein design. Proc Natl Acad Sci U S A 102: 12724–12729.
47. Pjura P, Matsumura M, Baase WA, Matthews BW (1993) Development of an in vivo method to identify mutants of phage T4 lysozyme of enhanced thermostability. Protein Sci 2: 2217–2225.
48. Sali D, Bycroft M, Fersht AR (1991) Surface electrostatic interactions contribute little of stability of barnase. J Mol Biol 220: 779–788.
49. Zhang XJ, Baase WA, Shoichet BK, Wilson KP, Matthews BW (1995) Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. Protein Eng 8: 1017–1022.
50. Perez-Jimenez R, Godoy-Ruiz R, Ibarra-Molero B, Sanchez-Ruiz JM (2005) The effect of charge–introduction mutations on *E. coli* thioredoxin stability. Biophys Chem 115: 105–107.
51. Strickler SS, Gribenko AV, Gribenko AV, Keiffer TR, Tomlinson J, et al. (2006) Protein stability and surface electrostatics: A charged relationship. Biochemistry 45: 2761–2766.
52. Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW (2002) Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. J Mol Biol 316: 225–233.
53. Larson SM, Di Nardo AA, Davidson AR (2000) Analysis of covariation in an SH3 domain sequence alignment: Applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 303: 433–446.
54. Green SM, Shortle D (1993) Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. Biochemistry 32: 10131–10139.
55. Hochachka P, Somero G (2002) Biochemical adaptation. Mechanism and process in physiological evolution. New York: Oxford University Press. 480 p.
56. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
57. Lassmann T, Sonnhammer EL (2005) Kalign: An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298.