CHAPTER TEN

Evolution-Based Design of Proteins

Kimberly A. Reynolds, William P. Russ, Michael Socolich, Rama Ranganathan[1]
Green Center for Systems Biology, Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[1]Corresponding author: e-mail address: rama.ranganathan@utsouthwestern.edu

Contents

Abstract

Statistical analysis of protein sequences indicates an architecture for natural proteins in which amino acids are engaged in a sparse, hierarchical pattern of interactions in the tertiary structure. This architecture might be a key and distinguishing feature of evolved proteins—a design principle providing not only for foldability and high-performance function but also for robustness to perturbation and the capacity for rapid adaptation to new selection pressures. Here, we describe an approach for systematically testing this design principle for natural-like proteins by (1) computational design of synthetic sequences that gradually add or remove constraints along the hierarchy of interacting residues and (2) experimental testing of the designed sequences for folding and biochemical function. By this process, we hope to understand how the constraints on fold, function, and other aspects of fitness are organized within natural proteins, a first step in understanding the process of "design" by evolution.

Methods in Enzymology, Volume 523
ISSN 0076-6879
http://dx.doi.org/10.1016/B978-0-12-394292-0.00010-2

1. INTRODUCTION

Natural proteins can fold under physiological conditions into compact three-dimensional structures and are capable of remarkably complex and high-performance biochemical functions. Because these properties require great accuracy in the position and dynamics of certain amino acids, it is tempting to think of proteins as precisely engineered systems in which interactions between the components (amino acids) are finely tuned and exactly arranged throughout the structure. However, other aspects of natural proteins are inconsistent with this view and demand a deeper examination of the basic underlying design principles through the process of evolution. For example, proteins are typically robust to random mutation; that is, they tolerate perturbations in many amino acid positions without much alteration in function (Bowie, Reidhaar-Olson, Lim, & Sauer, 1990; McLaughlin, Poelwijk, Gosal, & Ranganathan, 2012; Reidhaar-Olson & Sauer, 1990). In addition, they are plastic; that is, they have the ability to adapt to changing selection pressures by allowing specific variation of a few residues to profoundly alter function (McLaughlin et al., 2012; Orencia, Yoon, Ness, Stemmer, & Stevens, 2001). This curious combination of robustness to random mutation and yet sensitivity to targeted perturbation is interesting because it suggests that despite the appearance of precise construction throughout, strong heterogeneity exists in the design of proteins such that some residues and interactions between residues (the "core" machinery) are much more important than others. A major current goal in protein biology is to define and then mechanistically understand this heterogeneous architecture of natural proteins.

What is a good approach for this problem? The first step is to systematically map the energetic value of amino acid interactions globally in proteins, a nontrivial task by computational or experimental approaches. The main reasons are well known: (1) the relationship between the observed structural features of amino acid interactions and their net energetic value is extremely subtle and (2) amino acids can interact through complex high-order cooperative groups that can induce functionally significant energetic couplings between noncontacting amino acids. In general, physics-based approaches to computing protein energetics are based on making simplifying approximations for these two issues that result in quasi-empirical potential functions that consider only local interactions in protein structure. These approximations are necessary to make the calculations feasible and have led to remarkable successes in the engineering of protein folds (Dahiyat & Mayo, 1997; Dantas, Kuhlman, Callender, Wong, & Baker, 2003; Harbury, Plecs, Tidor, Alber, & Kim, 1998; Kuhlman et al., 2003). Nevertheless, it is important to realize that the design principles imposed by the approximations made can, in principle, deviate significantly from the natural evolutionary design of proteins.

The remarkable recent advances in genome sequencing efforts suggest an alternative *statistical* approach to this problem. The basic idea is that extant sequences have been selected through a long process of random mutation and selection and that a protein family that shares an overall fold and basic aspects of function should reveal the architecture of key amino acid interactions in the pattern of statistical constraints on and between amino acid positions. This idea is nothing more than a quantitative generalization of the widely accepted principle of sequence conservation as a metric of structural or functional importance. The conjectures are twofold: (1) positions that are important should experience an evolutionary constraint and should show a degree of conservation that reflects this constraint and (2) positions that energetically interact (whether through direct structural interactions or through indirect pathways of amino acid interactions) should experience a joint evolutionary constraint and should show correlated conservation or coevolution. Below, we discuss primarily one quantitative approach based on these conjectures called the statistical coupling analysis (SCA). For a given protein family, this analysis yields a mechanistically unbiased global map of amino acid interactions that encapsulates evolutionary constraints over all biochemical and biophysical properties that contribute to fitness.

Application of SCA in many protein families has led to a general model for the architecture of natural proteins—protein structure and function are hierarchically encoded by a subset of residues (termed the sector) embedded within the protein. In this chapter, we first give a brief description of our mathematical approach to sequence analysis and summarize the basic findings of protein sectors. We then describe a design method for testing this model for natural proteins by the creation of synthetic proteins that explore the hierarchy of statistical constraints. It is our intent to provide a general recipe for experiments to investigate how the pattern of amino acid correlations specifies folding, stability, and function in natural proteins.

2. SCA: THE PATTERN OF EVOLUTIONARY CONSTRAINT IN PROTEINS

The details of SCA have been described elsewhere (Halabi, Rivoire, Leibler, & Ranganathan, 2009; Smock et al., 2010), but here we give an overview as a preliminary to describing our design methodology. Matlab scripts for performing these calculations in full as well as tutorials illustrating SCA for several representative protein families are provided on our website (http://systems.swmed.edu/rr_lab).

2.1. The basic calculations

The process of SCA begins with assembly of a large and diverse multiple sequence alignment (MSA) for a particular protein family. For example, consider an MSA comprising 240 sequences of the WW domain family of small protein interaction modules that bind to proline-containing target peptides (Fig. 10.1A). The suitability of an MSA for SCA depends on multiple factors, such as the number of sequences, the sampling of phylogenetic space, and the general quality of the alignment (lack of large gapped regions, correct alignment of key functional residues). However, in practice, a general (though not strict) guideline for alignment construction is the inclusion of more than 100 sequences with a mean sequence identity between sequence pairs in the range of 15–50%.

From the MSA, the first-order analysis is to compute the conservation of each amino acid a at position i considered independently of other positions. In SCA, conservation is measured by $D_i^{(a)}$, an information-theoretic quantity called the Kullback–Leibler (K–L) relative entropy. This quantity indicates the deviation in the frequency of amino acid a at position i $\left(f_i^{(a)}\right)$ from the background probability of amino acid a $(q^{(a)})$ estimated from the nonredundant protein database. In the limit of large sampling (number of sequences > 80; Halabi et al., 2009), this calculation reduces to

$$D_i^{(a)} = f_i^{(a)} \ln\frac{f_i^{(a)}}{q^{(a)}} + \left(1 - f_i^{(a)}\right) \ln\frac{1 - f_i^{(a)}}{1 - q^{(a)}}. \quad [10.1]$$

The K–L entropy basically describes how unexpected the observed frequency $f_i^{(a)}$is, given an expected probability of $q^{(a)}$, and has the following two properties: (1) $D_i^{(a)} = 0$ if $f_i^{(a)} = q^{(a)}$ and (2) $D_i^{(a)}$ increases nonlinearly more and more steeply as f deviates from q. An overall positional K–L entropy D_i can also be computed that takes into account all the amino acids per position (Fig. 10.1B, bar graph):

$$D_i = \sum_{a=0}^{20} f_i^{(a)} \ln\left(\frac{f_i^{(a)}}{\bar{q}^{(a)}}\right), \quad [10.2]$$

where $\bar{q}^{(a)}$ represents the background frequencies including gaps (Halabi et al., 2009). For the WW domain, the conservation pattern is as expected;

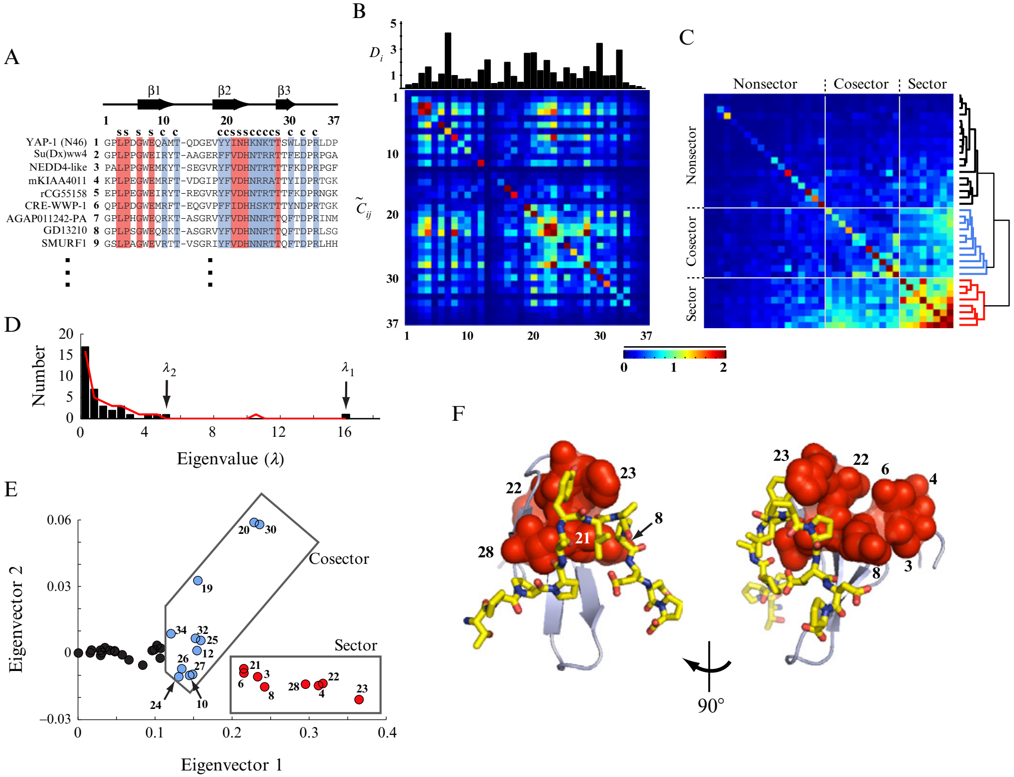


Figure 10.1 Statistical coupling analysis (SCA). (A) A portion of the alignment for the WW domain family. Sector (s), cosector (c), or nonsector (unmarked) positions (defined below) show no obvious arrangement in primary or secondary structure. (B) The site-independent conservation (D_i, bar graph) and the SCA matrix of coevolution between all pairs of amino acids $\left(\tilde{C}_{ij}\right)$. Values in the matrix are as indicated by the color bar. (C) Clustering in the matrix reveals three main groups of residues: sector, cosector, and nonsector. (D) Comparison of the eigenspectrum of the SCA matrix generated from the natural alignment (bars) to eigenspectra for randomized versions of the alignment (line) indicates that just the top two eigenvalues, λ_1 and λ_2, are distinguished from noise. (E) The corresponding eigenvectors reveal the positions that contribute the most to the top eigenvalues and define the sector positions (red) and cosector positions (blue). (F) The sector shown as red space-filling spheres on a representative WW domain structure (PDB ID: 2LAW, gray cartoon) in complex with a peptide ligand (stick bonds).

the two positions with the eponymous tryptophan residues (7 and 30) are the most conserved, together with a proline at position 33 that is a key part of the protein core. In general, prior work shows that the pattern of positional conservation is an effective predictor of residue burial in the tertiary structure (Halabi et al., 2009).

The second-order analysis is to compute a conservation-weighted correlation matrix $\left(\tilde{C}_{ij}\right)$ that represents the coevolution of each pair of positions in the MSA. To do this, we compute the weighted correlation tensor $\tilde{C}_{ij}{}^{(ab)}$:

$$\tilde{C}_{ij}{}^{(ab)} = \phi_i{}^{(a)}\phi_j{}^{(b)}C_{ij}{}^{(ab)}, \qquad [10.3]$$

where $C_{ij}{}^{(ab)} = f_{ij}{}^{(ab)} - f_i{}^{(a)}f_j{}^{(b)}$ represents the raw frequency-based correlations between each pair of amino acids (a,b) at each pair of positions (i, j), and ϕ represents a conservation-based weighting function. In the current implementation of SCA, $\phi_i{}^{(a)} = \left|\partial D_i{}^{(a)}/\partial f_i{}^{(a)}\right|$, the gradient of relative entropy. Just as $D_i{}^{(a)}$ represents positional conservation by the significance of observing a frequency $f_i{}^{(a)}$ given a background expectation, $\tilde{C}_{ij}{}^{(ab)}$ represents coevolution by the significance of observing a raw correlation in $C_{ij}{}^{(ab)}$ as judged by the weighting functions $\phi_i{}^{(a)}$ and $\phi_j{}^{(b)}$. Thus, $\tilde{C}_{ij}{}^{(ab)}$ up-weights correlations between conserved positions and damps correlations between less conserved positions. This conservation weighting serves to minimize the contribution of purely phylogenetic correlations between weakly conserved positions that are expected to emerge from small clades of sequences that have not had sufficient time to decorrelate unconstrained pairs of sequence positions. Other weighting functions are possible and are the subject of ongoing studies, and they will not be discussed further here. Regardless, the salient concept is that SCA considers conservation-weighted correlations between amino acids.

The result of this calculation is a four-dimensional tensor, $\tilde{C}_{ij}{}^{(ab)}$, that contains the correlation of every amino acid pair (a, b) for every position pair (i, j). For a protein with N positions, the dimensions of $\tilde{C}_{ij}{}^{(ab)}$ are $N \times N \times 20 \times 20$. We then reduce this tensor to an $N \times N$ matrix of positional correlations $\left(\tilde{C}_{ij}\right)$ by taking the Frobenius norm of each 20×20 amino acid correlation matrix for each amino acid pair:

$$\tilde{C}_{ij} = \sqrt{\sum_{(ab)}\left(\tilde{C}_{ij}^{(ab)}\right)^2}. \qquad [10.4]$$

This matrix norm gives the overall magnitude of correlation between each pair of positions (i, j) arising through all possible amino acid pairs. $\tilde{C}_{ij}$ is also referred to, in short, as the SCA matrix—a global examination of statistical coupling between all pairs of positions in the long-term evolutionary record of a protein family. Figure 10.1B shows $\widetilde{C}_{ij}$ for the WW family; in this matrix, diagonal elements are related to the intrinsic conservation of each position, and each off-diagonal element indicates the coevolution between a pair of amino acid positions.

2.2. Analysis of the SCA positional coevolution matrix

How can we analyze the pattern(s) of amino acid coevolution in the SCA matrix? Visual examination of $\widetilde{C}_{ij}$ for many different protein families leads to two main observations. First, the pattern of correlations is not obviously organized with respect to proximity in primary structure or to the pattern of contacts between secondary structure elements (Fig. 10.1A). Second, the matrix is remarkably sparse—the majority of amino acids appear to evolve relatively independently (as indicated by the large number of weak correlations in Fig. 10.1B), while a few show strong indications of coevolution.

Clustering of the SCA matrix makes this result more obvious—a subset of amino acids located in the bottom right corner are strongly coevolving while the majority of amino acids are more weakly coupled to one another (Fig. 10.1C). A closer inspection of the clustered matrix suggests a hierarchical organization of amino acid interactions. The bottom cluster (Fig. 10.1C) constitutes a small group of residues that collectively coevolve with one another and the group contains the majority of the signal in the matrix; we define such a group of residues as a "protein sector." The middle cluster (Fig. 10.1C) comprises residues that show little direct coupling to each other but that show systematic coevolution with sector positions (Fig. 10.1C). As these positions cluster by association to the sector, we call them the "cosector." The third and largest set (the nonsector, Fig. 10.1C) shows very little coupling at all.

Clustering provides one means of examining the pattern of evolutionary constraints within the matrix, but a more rigorous approach derives from the principles of spectral decomposition and random matrix theory (Halabi et al., 2009). The spectral decomposition mathematically transforms a correlation matrix between initial variables (e.g., the SCA matrix of amino acid correlations) into eigenmodes, which are described by a set of eigenvectors and eigenvalues. In this representation, each eigenvector contains the weights for linearly combining the initial variables (e.g., the amino acid positions) and each associated eigenvalue indicates the quantity of overall variance captured by that eigenmode. For the SCA matrix, each eigenmode represents a group of residues that share a similar pattern of coevolution, and the eigenvalue spectrum—the histogram of ($\lambda_1 > \lambda_2 > \lambda_3 > \ldots$)—reveals how the coevolutionary signal is quantitatively distributed among the eigenmodes. Comparison of this distribution with the eigenvalue spectra of correlation matrices derived from randomized alignments (shown as a line in Fig. 10.1D) shows that most of the lowest modes are indistinguishable from noise, while the top few modes capture statistically significant correlations.

For the WW domain, examination of the eigenvectors associated with the top two eigenmodes confirms the findings from clustering; a small set of amino acid positions contribute to the majority of the covariation in the matrix and emerge along the first eigenmode (the sector) (Fig. 10.1D). Consistent with coevolution with the sector, the cosector is evident as positions with weaker weights on the first eigenmode and projection along the second eigenmode (Fig. 10.1E). Indeed, the sector identified by eigendecomposition corresponds exactly, in this case, to the set of residues identified by clustering the matrix, though an exact match between the two methods need not always be true. In general, the spectral decomposition (rather than clustering) is a more quantitative approach for sector identification and is valuable in the process of SCA-based protein design, described later.

What is the structural interpretation of the sector? In the WW domain, the sector forms a sparse, distributed, and physically contiguous network that is distinct from known classifications of proteins based on primary, secondary, and tertiary structure (Russ, Lowery, Mishra, Yaffe, & Ranganathan, 2005; Fig 10.1F). Mutational studies show that sector residues, whether near or far from the ligand, contribute cooperatively to binding affinity—an extended network underlying the WW domain function (Russ et al., 2005). In the PDZ family of protein interaction modules, the sector connects the ligand-binding site to an allosteric site located on the opposite face (Fig. 10.2A; Halabi et al., 2009; Lockless & Ranganathan, 1999; McLaughlin et al., 2012). Sectors have been found in all protein families studied to date, and like in WW and PDZ domains, are empirically observed to share three properties: they are (1) sparse (comprising ~20% of the protein structure), (2) they are physically contiguous, and (3) they connect the active site or ligand-binding site to distant surfaces distributed throughout the structure (Ferguson et al., 2007; Halabi et al., 2009; Hatley, Lockless, Gibson, Gilman, & Ranganathan, 2003; Lee et al., 2008; Lockless & Ranganathan, 1999; Shulman, Larson, Mangelsdorf, & Ranganathan, 2004; Smock et al., 2010; Suel, Lockless, Wall, & Ranganathan, 2003).

Interestingly, the mapping of sectors to domains is not necessarily one to one. Using more advanced extensions of eigendecomposition (Smock et al.,

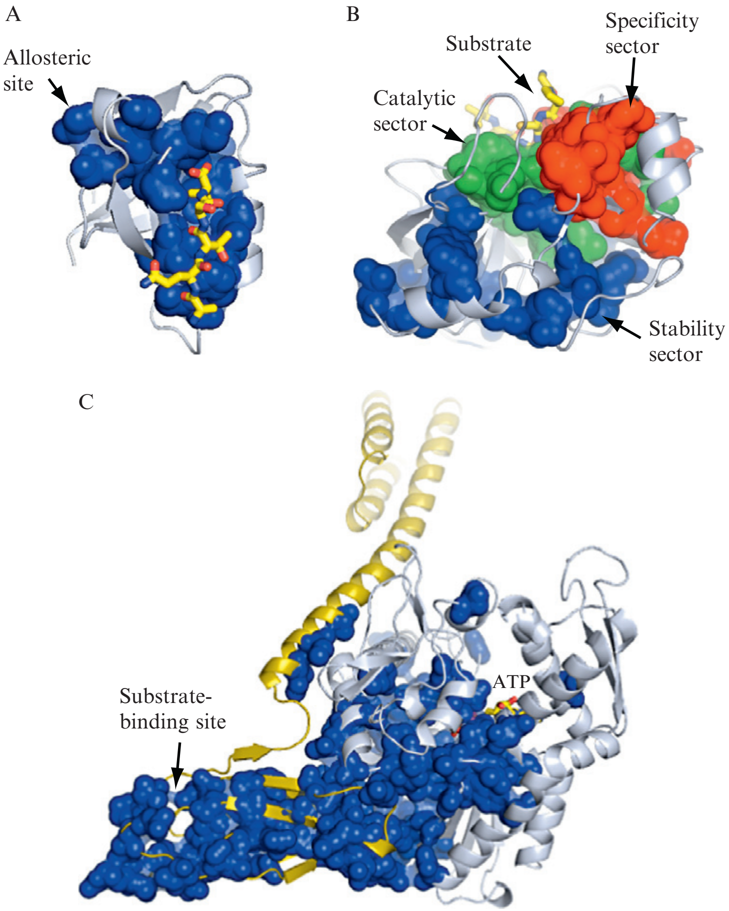


Figure 10.2 Sectors in three protein families. A single sector in the PDZ domain connects the ligand-binding pocket to a distant allosteric site (A), three quasi-independent sectors occur in the S1A family of serine proteases (B), and a single interdomain sector functionally connects the ATP-binding site in the nucleotide-binding domain (white) to the ligand-binding site in the substrate-binding domain (gold) in the Hsp70 family of molecular chaperones.

2010), it is possible to find multiple independent sectors within a single-domain protein. For example, in the S1A serine proteases, three near-independent sectors are evident, each of which comprises a distinct but physically contiguous subnetwork within the tertiary structure (Fig. 10.2B). Studies in one member of the S1A family (rat trypsin) show that each sector corresponds to a distinct biochemical property—catalytic mechanism, substrate specificity, and stability—indicating that this decomposition of the protein by patterns of coevolution is functionally relevant (Halabi et al., 2009). Conversely, it is also possible to find a single sector spanning two distinct domains of a protein (e.g., the Hsp70 chaperone, Fig. 10.2C; Smock et al., 2010). In Hsp70, the single interdomain sector links the ATP-binding site in the nucleotide-binding domain with the ligand-binding site in the substrate-binding domain through the interdomain interface, a feature that reflects the fact that the conserved functional activity of Hsp70 proteins is allosteric communication between the two domains. Thus, sectors expose the architecture of fitness constraints on or between proteins, independent of structural basis or mechanistic detail.

3. SCA-BASED PROTEIN DESIGN

How can we test the sufficiency of the sector model for protein structure, function, stability, and adaptability—the basic features of natural proteins? Targeted mutational analyses provide a first-order test that sectors specify aspects of protein function. But a more global and complete test comes through synthetic protein design. The idea is to carry out computational simulations that start with random sequences and evolve (*in silico*) synthetic sequences that are constrained by the observed evolutionary statistics. Experimental study of libraries of the designed sequences represents a deep test of the sufficiency of the applied constraints for recapitulating the properties of natural proteins.

3.1. Defining an objective function

The approach in SCA-based protein design is to use the Metropolis Monte Carlo simulated annealing (MCSA) algorithm to explore the sequence space consistent with a set of applied constraints between amino acids. The MCSA algorithm is an iterative numerical method for searching for the global minimum energy configuration of a system starting from any arbitrary state and is especially useful when the number of possible states is very large and the energy landscape is rugged and characterized by many local minima (Kirkpatrick, Gelatt, & Vecchi, 1983; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The energy function (or “objective function”) to be minimized can, in general, depend on many parameters of the system and represents the constraints that define the size and shape of the final solution space. In essence, the objective function can be thought of as the hypothesis being tested—the set of applied constraints that we wish to test for specifying folding, thermodynamic stability, function, and any other aspects of protein fitness.

For SCA-based protein design, the system under consideration is a MSA (rather than a single sequence), and the objective function (E) is the summed difference between the correlation tensor for a MSA of protein sequences during iterations of the design process and the target correlation matrix deduced from the natural MSA:

$$E = \sum_{ijab} \left| \tilde{C}^{(ab)}_{ij(\text{design})} - \tilde{C}^{(ab)}_{ij(\text{natural})} \right|. \qquad [10.5]$$

Thus, the lowest energy configuration for the designed MSA is the set of sequences that gives a pattern of correlations in the designed sequences $\left(\tilde{C}^{(ab)}_{ij(\text{design})}\right)$ that most closely reproduces that of the natural MSA $\left(\tilde{C}^{(ab)}_{ij(\text{natural})}\right)$. At the limit of large numbers of sequences, this result is tantamount to drawing sequences from a maximum entropy probability distribution consistent with the applied set of observed correlations (Bialek & Ranganathan, 2007).

What correlations should be included in the objective function? At the extreme limit, the objective function could involve the full correlation tensor in which $\tilde{C}_{ij}^{(ab)}$ has indices i and j that run over all positions in the MSA and a and b that run over all 20 amino acids; this is a trivial simulation because the only ensemble of sequences that lies at the global minimum of the objective function is the same sequences that comprise the natural MSA. However, a large number of (weak) correlations in the full $\tilde{C}_{ij}^{(ab)}$ are indistinguishable from noise due to finite sampling or phylogeny and are therefore proposed to be functionally insignificant. In addition, even the statistically significant correlations are not likely to be all equally important; indeed, there is a hierarchy of correlations within the sector such that some amino acids are more strongly coevolving and are surrounded by residues making lesser contributions (Fig. 10.1C). The key goal in SCA design is then is to find appropriate "reduced" objective functions that comprise a hypothesis for the "relevant" constraints and then to test these for sufficiency with regard to protein structure and function.

Reduced objective functions can be obtained by two general approaches: (1) elimination (or masking) of correlations in the $\tilde{C}_{ij}^{(ab)}$ tensor based on heuristic knowledge or on statistical cutoffs on the distribution of correlations and (2) partial convergence on the full correlation tensor. It is important to say that these techniques are not entirely different from each other; indeed, the collective group of correlations defining sectors tend to be larger in magnitude, and as we will describe below, partial convergence by the MCSA algorithm has the property of building in the collective modes defining sectors at the expense of the weaker, more idiosyncratic correlations. Nevertheless, these strategies represent different practical ways of posing hypotheses for testing the information content of protein sequences through design.

For example, in an initial study on SCA-based design of WW domains, the objective function involved a subset of $\widetilde{C}_{ij}^{(ab)}$ corresponding to the correlations for only five sector residues—a test that the correlations between just these few specific positions and all other sites is sufficient for protein folding and function (Russ et al., 2005; Socolich et al., 2005). However, other objective functions might be imagined; for example, for multisector proteins such as the S1A serine proteases, one can envision designs that target only the correlations that define a single protein sector—a design that should cause variation in one functional property while leaving the roles of other sectors relatively unperturbed. More generally, it would be interesting to test objective functions that sample different eigenmodes of the SCA correlation matrix to experimentally understand if and how the spectral decomposition of information content in protein sequences corresponds to a decomposition of the different biochemical properties of proteins that contribute to fitness.

3.2. The simulated annealing algorithm

Given an objective function, the strategy of the MCSA algorithm underlying SCA-based design follows the standard simulated annealing process. We initiate the simulation with an alignment that has been randomized by a process we term "vertical shuffling"—randomly permuting each column of the MSA independently. Vertical shuffling removes all nonrandom correlations between positions, but by nature, preserves the frequency distribution (i.e., the conservation) of amino acids at positions exactly. The MCSA method then involves many iterations of a two-step process to converge on the set of constraints specified by the objective function. In step one, we choose one position and two sequences from the MSA at random and swap the corresponding amino acids (Fig. 10.3A). Since the swap is always done within one position, this perturbation never influences the independent conservation of positions, but it could introduce a change to the pattern of correlations. In step 2, we evaluate the objective function (E) to give the impact of the swap on the overall set of included correlations and compute $\Delta E = E_{\text{current}} - E_{\text{previous}}$, the change in the objective function from the previous iteration. If $\Delta E \leq 0$, the perturbation is favorable (i.e., the pattern of correlations has become more natural-like), and we always accept the swap.

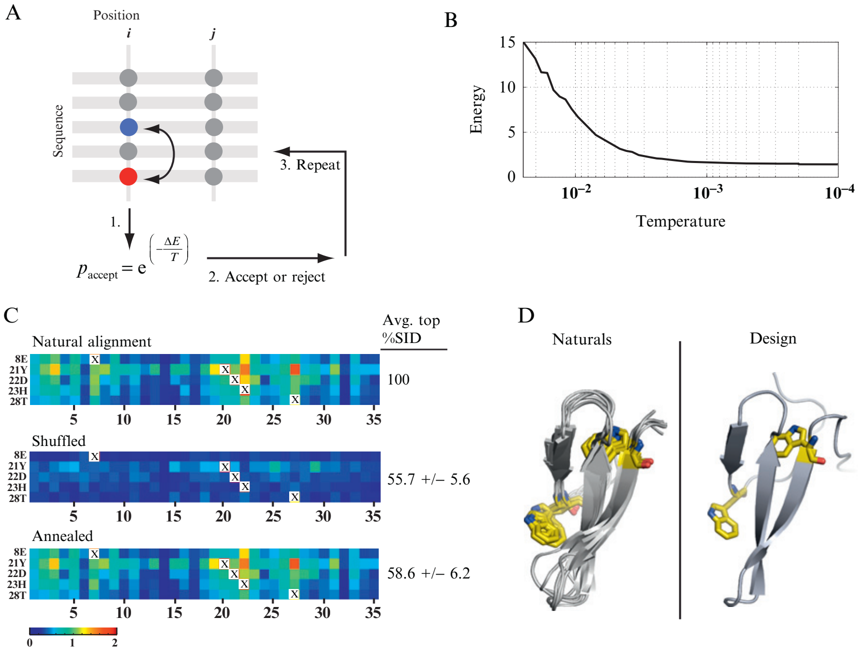


Figure 10.3 Evolution-based design of the WW domain. (A) The Monte Carlo design process. Each iteration involves two steps: (1) a swap of amino acids between two randomly chosen sequences at one randomly chosen position and (2) a decision to either accept or reject the swap based on the difference in the objective function (ΔE, see text) and a computational "temperature" (T). If $\Delta E > 0$, the swap is accepted with a probability determined by the Boltzmann distribution. (B) A MCSA trajectory for the WW domain family. The process starts with a high temperature and exponentially cools the MSA, converging toward a minimum for the objective function. (C) In Socolich et al. (2005), the objective function involved correlations for five sector positions. This portion of the SCA matrix (with self-correlations blanked) is shown for the natural WW alignment, the vertically shuffled (high temperature) alignment, and the annealed alignment. At right is indicated the average sequence identity (± SD) to the closest sequence in the natural WW domain MSA. (D) Comparison of experimental structures for a collection of natural WW domains and one of the synthetic WW domains. The eponymous tryptophan residues are indicated in stick bonds.

If $\Delta E > 0$, the perturbation is unfavorable, and we accept the swap with a probability given by the Boltzmann distribution:

$$p = e^{\frac{-\Delta E}{T}}. \qquad [10.6]$$

This property of MCSA—in which unfavorable swaps are probabilistically accepted—is a key feature that prevents the search algorithm from becoming trapped in local minima. The "temperature" factor T is a purely computational term that controls the probability of accepting unfavorable swaps. At high temperatures, swaps causing even significant perturbation to correlations are likely, while at low temperatures such swaps become exponentially less probable. The basic idea of the MCSA is to gradually cool the MSA from a high temperature along a near-equilibrium path until the simulation converges on a set of synthetic sequences that reproduces the correlations included in the objective function (Fig. 10.3B). It can be intuitively seen that in the path of an MCSA simulation, the strongest collectively evolving modes of the correlation matrix (that define sectors) will anneal first and cooperatively over a narrow range of temperatures, with the remainder of the weaker and less collective correlations converging gradually as the temperature cools further. The simulation exits when the temperature cools sufficiently that no further swaps are accepted.

3.3. SCA-based design of WW domains

The first application of protein design using evolutionary correlations was carried out for the WW family of protein interaction modules (Russ et al., 2005; Socolich et al., 2005). The WW domain adopts a curved, three-stranded antiparallel β-sheet configuration and binds to proline-rich peptide ligands along one face of the sheet (Fig. 10.1F). As mentioned earlier, the objective function in this initial design experiment involved a matrix comprising just the correlations between one dominant amino acid at five sector positions and all amino acids at all other positions—a heuristic choice based on the fact that these correlations capture much of the total information content in the SCA matrix for the WW family. Starting from the vertically shuffled MSA as the initial state (IC, or site-*i*ndependent *c*onservation sequences), the MCSA algorithm was used to converge on new MSA of sequences (CC, or "*c*oupled *c*onservation" sequences) that recapitulate the applied constraints (i.e., minimizing the objective function, Fig. 10.3B–C).

Four libraries of synthetic genes were constructed and analyzed to experimentally test the likelihood of native folding and function in SCA-based design: (1) 48 natural WW domains drawn randomly from the natural MSA (as a positive control), (2) 48 IC sequences that represent the input for MCSA, (3) 48 CC sequences that represent the output of MCSA, and (4) 23 completely random sequences (R) with amino acids at each position chosen from the mean frequency of each amino acid in the WW MSA (as a negative control). Statistically, the natural, IC, and CC sequences showed a mean amino acid identity to natural WW domains of ~36%, an expected result given the constraint to preserve the conservation of amino acids at sites. Also expected, the random sequences show a much lower mean identity to natural WW domains (~6%). However, the IC and CC sequences show a similar "top-hit" identity on average with natural sequences—the percent identity to their closest counterpart in the natural world (Fig. 10.3C). Thus, the CC sequences are statistically indistinguishable in sequence divergence from IC domains indicating that, by this measure, the number of extra constraints from correlations is small. In essence, the difference between IC and CC sequences is not the magnitude of similarity to natural sequences but the pattern by which they are similar.

Analysis of solubility, folding thermodynamics, and ligand-binding specificity for all soluble domains showed a clear result: no random or IC sequences were folded, but a significant fraction of CC sequences showed both native folding and biochemical function that quantitatively recapitulated the behavior of natural WW proteins (Socolich et al., 2005). In addition, structure determination of one synthetic CC domain demonstrated recapitulation of the characteristic tertiary structure of the WW domain, with an atomic-level accuracy that is within the variance of known natural WW structures (Fig. 10.3D). Thus, for this domain, fold and function can be recapitulated by the information contained in the portion of the SCA matrix that contains the pairwise correlations for a set of sector positions. This experiment provides a first test that the pattern of amino acid coevolution in the SCA matrix represents one solution for specifying natively folded and functional proteins. The small fraction of total correlations used in the objective function implies a surprising simplicity in the evolutionary design of this protein domain.

4. SCA-BASED PARSING OF PROTEIN STABILITY AND FUNCTION

The objective function used in the initial design of WW domains included the correlations for five sector positions over all other positions. The near-sufficiency of this design is interesting, but this experiment does not decompose the contribution of sector and nonsector positions. Indeed, does the hierarchy of correlations differentially encode properties of protein folding and function? To examine this, let us revisit the clustered matrix for the WW domain family (Fig. 10.1C). This matrix shows a hierarchical pattern of organization with three groups of residues identified by clustering:

sector, cosector, and nonsector. To understand how this organization encodes protein stability and function, we conducted a simple "positional shuffling" experiment on the WW domain sequence alignment (Fig. 10.4A–E). In this experiment, we designed synthetic variants of one WW domain sequence within the overall MSA (the second WW domain from the Yes-kinase associated protein 1 (YAP-1), referred to as N46; Socolich et al., 2005) in which statistical couplings are systematically eliminated for sector, cosector, and nonsector groups, either alone or in combination. To do this, we simply vertically shuffle the amino acids independently at each position comprising a group and select the N46 variant. As described in Section 3.2, this process removes all nonrandom correlations between positions comprising a selected group(s) while preserving the conservation of amino acids at individual sites (Fig. 10.4E).

In total, we generated four sets of 20 shuffled N46 sequences each (80 total proteins), which are named according to the residue clusters in which correlations were retained: N46(1)—sector intact, all other positions shuffled (Fig 10.4A); N46(1 + 2)—sector/cosector intact, nonsector positions shuffled (Fig 10.4B); N46(2 + 3)—cosector/nonsector intact, sector positions shuffled (Fig 10.4C); and N46(3)—nonsector intact, sector, and cosector positions shuffled (Fig 10.4D). All the synthetic shuffled sequences were characterized for three properties: solubility upon expression in *Escherichia coli*, presence of a cooperative unfolding transition by thermal denaturation (Socolich et al., 2005), and function as assessed by class-specific peptide binding (Russ et al., 2005). N46 is a class I WW domain, recognizing PPxY containing target peptides; accordingly, functional synthetic WW domains were scored as those that recapitulate this same binding specificity.

The results of these experiments are summarized in Fig. 10.4F and G. First, the data show that for all four sequence sets only a small number of domains are insoluble (gray wedges), leaving the majority for analysis of fold stability and function. Second, both the necessity and near-sufficiency of the sector/cosector for N46-like class I binding specificity are evident. Shuffling the sector positions (N46(2 + 3)) results in only one of 16 tested domains with class I specificity, and shuffling of both the sector and cosector positions (N46(3)) results in a total loss of function for all domains (Fig. 10.4G). In contrast, preserving only the eight sector residues and shuffling the remainder (N46(1)) results in 5 of 16 functional domains, and retaining both the sector and the cosector (N46(1 + 2)) results in 16 of 18 class I domains (Fig. 10.4G). Thus, protein function in the WW domain largely emerges from the sparse network of amino acid positions that define the sector.

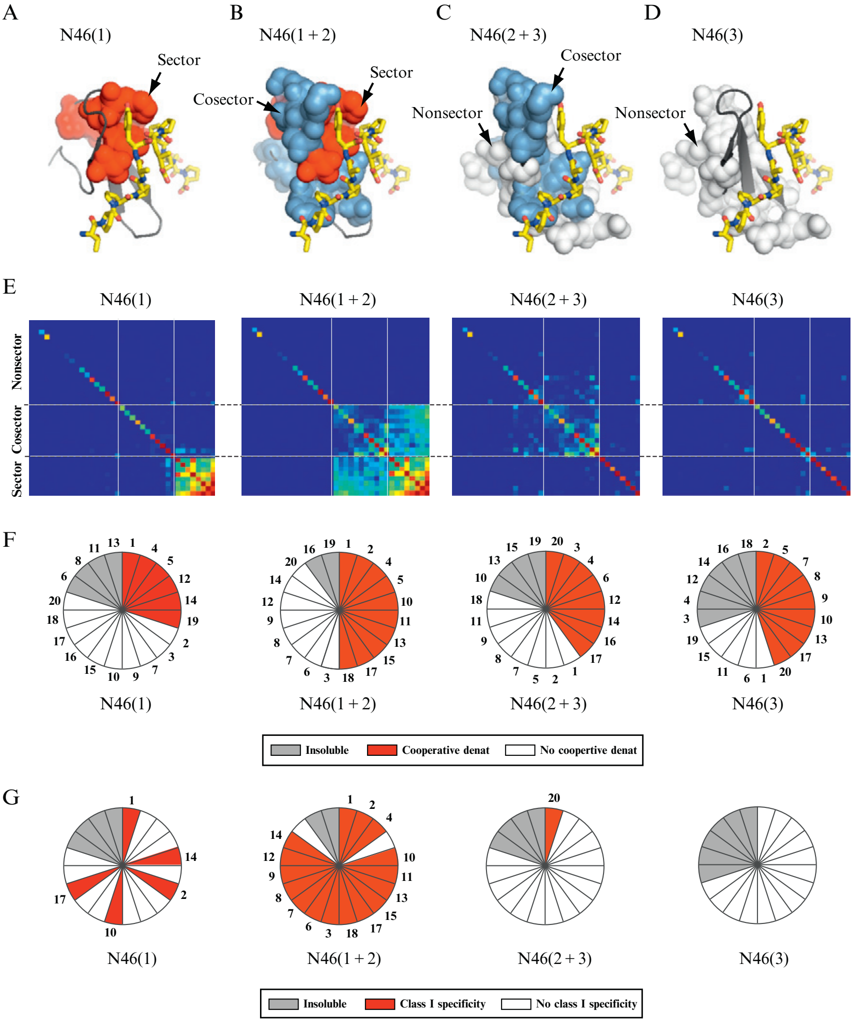


Figure 10.4 SCA-based design of shuffled WW domains. (A–D) In each panel, the group of residues for which couplings were preserved (nonshuffled positions) are shown in space-filling spheres on the structure of the "N46" WW domain in complex with a Class I peptide ligand (PDB 2LAW). Sector, cosector, and nonsector positions are labeled. (E) SCA matrices for each set of shuffled N46 sequences, visually illustrating the correlations that are scrambled. (F–G) Ordered pie charts showing the outcome of thermal denaturation (F) and Class I peptide binding specificity (G) experiments for each of the four groups of synthetic WW domains. The order of slices is the same in the two panels to facilitate comparison.

Interestingly, fold stability seems to obey a different rule. Regardless of whether one shuffles only the sector positions (N46(2 + 3)), the sector and cosector positions (N46(3)), or nonsector positions (N46(1 + 2)), the results are the same: 30–50% of the resulting domains display a cooperative folding transition in the 4–90 °C temperature range (Fig. 10.4F). Taken together, these findings suggest that the capacity to fold and exhibit specific molecular recognition is localized to the subset of positions showing coevolution (the sector/cosector), while the stability of the fold is a more distributed property of the protein structure, even involving weakly correlated and less conserved positions.

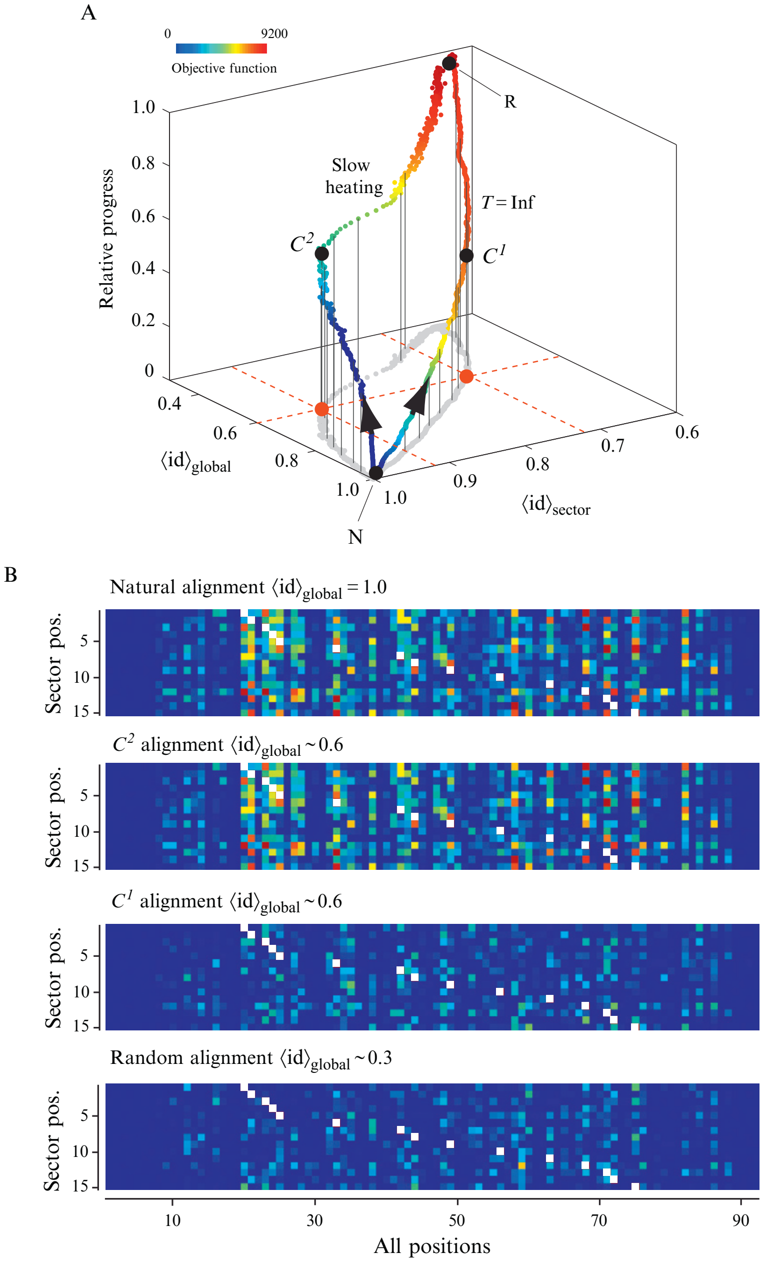
From an evolutionary perspective, this suggests the possibility that the amino acid interactions underlying thermodynamic stability need not be deeply conserved in protein families, but rather can be easily varied. That is, the origin of thermodynamic stability in any particular member of a protein family may be a rapidly changing and perhaps even idiosyncratic feature emerging from small local groups of amino acids with many degenerate possible solutions. The generality and validity of this apparent parsing of fold stability and function should be more deeply examined in larger and more stable proteins and with greater sampling of synthetic designs.

5. FUTURE MONTE CARLO STRATEGIES FOR EXPLORING SEQUENCE SPACE

The experiments described above provide a simple coarse-grained preview of how properties of protein folding and function might be encoded in the hierarchy of positional correlations. But, to examine the pattern of residue couplings with greater resolution and less interpretational bias (i.e., without heuristically defined objective functions and parsing of correlations into discrete blocks such as sector and cosector), we need a way to systematically add or remove correlations along the hierarchy. Here, we describe a strategy to address this question. Implementing this method requires three things: (1) a computational method to design protein sequences that smoothly vary couplings along a hierarchy observed in the SCA matrix, (2) a way to synthesize genes corresponding to a large number of synthetic proteins along this trajectory cheaply and reliably, and (3) high-throughput methods to assess protein function for the libraries of synthetic designs. The methods for assessing protein function are specific for model systems and are not discussed here, but they should be generally possible for any protein in which cell growth rate or fitness can be coupled to protein activity. For example, proteins such as primary metabolic enzymes (Reynolds, McLaughlin, & Ranganathan, 2011; Taylor, Kast, & Hilvert, 2001) or enzymes that mediate antibiotic resistance (Weinreich, Delaney, Depristo, & Hartl, 2006) have obvious advantages in this respect. These experiments are likely to become feasible as advances in technologies for gene synthesis and automated screening for protein stability and function mature (Gerber, Maerkl, & Quake, 2009; Isom, Marguet, Oas, & Hellinga, 2011; Kosuri et al., 2010).

The computational approach is to design sequences as a function of temperature along a Monte Carlo trajectory (i.e., obeying Eq. 10.5), systematically testing for loss (in the case of heating) or gain (in the case of cooling) of protein properties of interest. The cooling trajectory is as described above in Section 3.2—we begin with a vertically shuffled alignment and lower the temperature along a near-equilibrium path to converge on our objective function. The result is an ensemble of sequences that can be characterized at each temperature. Constraints between residues anneal as a function of temperature according to their strength and collective character and thus this design has the property of building the top eigenmodes (defining sectors) first and then slowly annealing on the lower eigenmodes containing weaker and less collective correlations. A dense sampling of sequences along this trajectory would provide a rigorous test of how the statistical structure of correlations is related to various protein properties.

Monte Carlo simulated heating (MCSH) is conceptually similar to MCSA but differs in initial condition and direction of progress (Fig. 10.5A). In this experiment, we begin with a MSA of natural sequences (rather than a vertically shuffled MSA) and conduct a simulation while raising the temperature according to a specified heating schedule until the alignment is completely vertically shuffled. In effect, this is a strategy for computationally introducing mutations in natural sequences constrained by the positional conservation of amino acids and according to a pattern specified by the objective function and the heating protocol. For example, consider the heating trajectories in Fig. 10.5A for an MSA of 240 members of the PDZ family of protein interaction modules. The objective function is the full $\widetilde{C}_{ij}^{(ab)}$ correlation tensor, and we show two trajectories differing by heating protocol starting from the MSA of natural sequences (marked N, Fig. 10.5A). Carrying out a simulation with the temperature set to infinity defines a path in which all positions mutate within their conservation pattern without regard to correlations (see Eq. 10.6), finally approaching the fully randomized, vertically shuffled limit (marked R, Fig. 10.5A). Accordingly, both global and sector identity


A
0
9200
Objective function
1.0
0.8
0.6
0.4
0.2
0
Relative progress
R
Slow heating
T = Inf
C^2
C^1
0.4
0.6
$\langle id \rangle_{global}$
0.8
1.0
N
1.0
0.9
0.8
0.7
0.6
$\langle id \rangle_{sector}$
B
Natural alignment $\langle id \rangle_{global} = 1.0$
Sector pos.
5
10
15
C^2 alignment $\langle id \rangle_{global} \sim 0.6$
Sector pos.
5
10
15
C^1 alignment $\langle id \rangle_{global} \sim 0.6$
Sector pos.
5
10
15
Random alignment $\langle id \rangle_{global} \sim 0.3$
Sector pos.
5
10
15
10
30
50
70
90
All positions

to the natural sequences are lost equally, and sequences at the position marked C^1 show a pattern of correlation for sector positions that is nearly randomized (Fig. 10.5B). In contrast, slow near-equilibrium heating produces a very different trajectory; in this process, global identity is initially lost without much loss in sector identity until a characteristic temperature at which the sector "melts" and the trajectory approaches the same fully vertically shuffled limit. Accordingly, sequences at the position marked C^2 have a pattern of correlations for sector positions that is nearly the same as for natural sequences, despite the same global divergence as C^1 sequences (Fig. 10.5B).

An interesting experiment is to choose one natural sequence within the MSA as a model system, carry out many trials of MCSH, and build versions of this protein at different temperatures along the trajectories. Experimental characterization of the natural sequence is a specific reference for folding, stability, and function in the synthetic "heated" sequences. Study of ensembles of synthetic variants sampled along the two different heating trajectories should provide a clear answer to how these properties of the selected natural sequence differentially diverge as a function of systematically removing the information contained in the SCA correlation tensor.

Both MCSA and MCSH methods provide a means to explore the mapping between statistical correlations and protein structure/function, a mapping that deserves study in several protein systems. Indeed, it will be important to see how the results compare for small versus large domains and for enzymes versus more simple binding proteins.

Figure 10.5 MCSH trajectories for the PDZ domain family. (A) A plot mapping the progress of two different heating trajectories against the average "top-hit" sequence identities of designed sequences calculated for the full-length sequence ($<\text{id}>_{\text{global}}$) or for just sector positions ($<\text{id}>_{\text{sector}}$). In the $T=\text{Inf}$ trajectory, each position is allowed to mutate within its conservation pattern without regard to correlations, and global and sector identity drop together. In the slow heating trajectory, the temperature is gradually increased to "melt out" couplings between positions in an order that depends on the strength and collective nature of the correlations. The value of the objective function along the trajectories is indicated by the color bar, and four points are marked for reference with panel B: N, the natural MSA; R, the fully randomized, vertically shuffled MSA; and C^1 and C^2, two intermediary points that share the same global sequence divergence but differ significantly in sector divergence. (B) Subset of the SCA matrix $\tilde{C}_{ij}$ for 15 sector positions in the PDZ family to illustrate the property of the heating trajectories. Despite identical global sequence divergence, C^1 sequences show a pattern of correlations that nearly approaches the fully randomized case while C^2 sequences show correlations that are nearly the same as for the natural MSA. Experimental analysis of C^2 and C^1 sequences or more generally sequences drawn from both trajectories represent a systematic investigation of how properties of natural proteins are stored in the pattern of correlations.

6. CONCLUSION

Protein design represents one approach for understanding how the pattern of pairwise residue couplings inferred from the statistics of natural protein sequences is related to the encoding of protein structure and function. Further, it may provide insight into decomposability of biochemical properties. In cases such as the S1A serine proteases, the finding of multiple statistically independent sectors offers the exciting possibility of orthogonal control of different biochemical properties of these enzymes. More generally, it may be that a broad study of the Monte Carlo-based design trajectories for proteins will reveal rules for tuning stability and function independently through targeted variation of protein sequences. The methods described here present one approach to distill the general principles, if any, for the design of natural proteins. Practically, such rules might permit the design of improved synthetic proteins that show natural-like properties of high catalytic efficiency, mutational robustness, and adaptability to new functional challenges.

ACKNOWLEDGMENTS

We thank Dr. Walraj Gosal for his contributions on Monte Carlo Simulated Heating for the PDZ domain. We acknowledge support from the NIH (R01EY018720-05, R. R.), The Robert A. Welch Foundation (I-1366, R. R.), and the Green Center for Systems Biology (R. R.).

REFERENCES

Bialek, W., & Ranganathan, R. (2007). *Rediscovering the power of pairwise interactions.* arXiv:0712.4397.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990). Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science*, *247*, 1306–1310.

Dahiyat, B. I., & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science*, *278*, 82–87.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003). A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, *332*, 449–460.

Ferguson, A. D., Amezcua, C. A., Halabi, N. M., Chelliah, Y., Rosen, M. K., Ranganathan, R., et al. (2007). Signal transduction pathway of TonB-dependent transporters. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 513–518.

Gerber, D., Maerkl, S. J., & Quake, S. R. (2009). An in vitro microfluidic approach to generating protein-interaction networks. *Nature Methods*, *6*, 71–74.

Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: Evolutionary units of three-dimensional structure. *Cell*, *138*, 774–786.

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, *282*, 1462–1467.

Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G., & Ranganathan, R. (2003). Allosteric determinants in guanine nucleotide-binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 14445–14450.

Isom, D. G., Marguet, P. R., Oas, T. G., & Hellinga, H. W. (2011). A miniaturized technique for assessing protein thermodynamics and function using fast determination of quantitative cysteine reactivity. *Proteins*, *79*, 1034–1047.

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Kosuri, S., Eroshenko, N., Leproust, E. M., Super, M., Way, J., Li, J. B., et al. (2010). Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature Biotechnology*, *28*, 1295–1299.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*, 1364–1368.

Lee, J., Natarajan, M., Nashine, V. C., Socolich, M., Vo, T., Russ, W. P., et al. (2008). Surface sites for engineering allosteric control in proteins. *Science*, *322*, 438–442.

Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, *286*, 295–299.

McLaughlin, R. N., Jr, Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature*, *491*, 138–142.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. C., & Stevens, R. C. (2001). Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature Structural and Molecular Biology*, *8*, 238–242.

Reidhaar-Olson, J. F., & Sauer, R. T. (1990). Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. *Proteins*, *7*, 306–316.

Reynolds, K. A., McLaughlin, R. N., & Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell*, *147*, 1564–1575.

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., & Ranganathan, R. (2005). Natural-like function in artificial WW domains. *Nature*, *437*, 579–583.

Shulman, A. I., Larson, C., Mangelsdorf, D. J., & Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell*, *116*, 417–429.

Smock, R. G., Rivoire, O., Russ, W. P., Swain, J. F., Leibler, S., Ranganathan, R., et al. (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Molecular Systems Biology*, *6*, 414.

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature*, *437*, 512–518.

Suel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, *10*, 59–69.

Taylor, S. V., Kast, P., & Hilvert, D. (2001). Investigating and engineering enzymes by genetic selection. *Angewandte Chemie (International Ed. in English)*, *40*, 3310–3335.

Weinreich, D. M., Delaney, N. F., Depristo, M. A., & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, *312*, 111–114.