Knowledge-based potential functions in protein design

William P Russ and Rama Ranganathan*

Predicting protein sequences that fold into specific native three-dimensional structures is a problem of great potential complexity. Although the complete solution is ultimately rooted in understanding the physical chemistry underlying the complex interactions between amino acid residues that determine protein stability, recent work shows that empirical information about these first principles is embedded in the statistics of protein sequence and structure databases. This review focuses on the use of 'knowledge-based' potentials derived from these databases in designing proteins. In addition, the data suggest how the study of these empirical potentials might impact our fundamental understanding of the energetic principles of protein structure.

Addresses
Howard Hughes Medical Institute and Department of Pharmacology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA
*e-mail: rama@chop.swmed.edu

Current Opinion in Structural Biology 2002, 12:447–452

0959-440X/02/$ – see front matter

Abbreviations
hGH human growth hormone
hGHbp hGH receptor extracellular domain
SH Src homology

Introduction

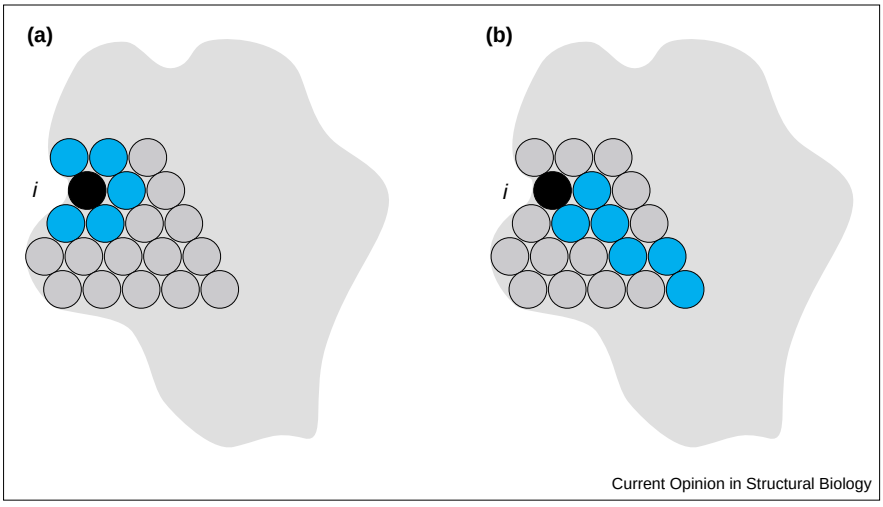
The goal in protein design is to determine the rules for predicting amino acid sequences that will stably fold into a specific target three-dimensional structure. Two features make this problem extraordinarily complex at the present time. First, even small proteins containing on the order of one hundred amino acids can encode an astronomical number of potential sequences (~10^{130}). Numbers of this magnitude clearly preclude exhaustive searching of sequence space with any computational or experimental method; instead, nearly all protein design algorithms implement some techniques for biasing the search towards regions of sequence space with a higher likelihood of identifying members of the target fold family. Second, the scoring (or potential) functions used in assessing the free energy change upon folding are not well defined at a physical chemical level and are often unpredictably imprecise in reporting the experimentally observed energetic properties of proteins. Despite these difficulties at the level of first principles, substantial progress has recently been made in protein design through the application of empirical information from databases of protein sequence and structure. When extracted in the form of statistical quantities suitable for use in computational algorithms, this information is collectively referred to as knowledge-based potentials, which have been demonstrated to help in both reducing the combinatorial complexity of searching sequence space and refining the scoring functions. In this review, we examine the ways in which knowledge-based potentials have been used to influence protein design and suggest that further study of the mechanistic origin of these potentials may lead to a better understanding of the physical chemistry of proteins.

Sources of imprecision and complexity in protein design

The core problem of rational protein design is the construction of a potential function that describes protein stability. The logical basis for this energy function arises from the thermodynamic hypothesis that the native structure of a protein is given by the conformation of amino acids that minimizes the net free energy of the molecule [1]. In design algorithms, sequences that minimize the potential function are expected to have the greatest likelihood of adopting the target structure. However, despite substantial knowledge of the nature of the fundamental forces between atoms and of many high-resolution protein structures, we do not yet have a general form for this energy function. Many examples illustrate the difficulty of deducing energy from a protein structure. The human growth hormone (hGH) buries many residues at the protein–protein interface upon binding to the extracellular domain of its receptor (hGHbp), but most of these contacts show a net free energy change close to zero upon mutagenesis [2]. Instead, only a few so-called 'hot spot' residues seem to account for most of the interaction energy. Importantly, neither the pattern nor the magnitude of energy change at sites are predictable from either the free or bound structures of the hormone and receptor [2,3]. Similarly, the GCN4 leucine zipper makes a set of interhelical salt bridges upon dimerization [4], but mutagenesis shows that these structurally observed interactions actually destabilize the complex [5]. From a protein design point of view, our inability to predict the energetic value of atomic interactions given an actual atomic structure highlights the serious difficulty in estimating these values from hypothetical sequences that must be scored during iterations of the design process.

An additional level of difficulty in specifying the energetics of a protein fold is the potential for nonindependence of interactions between amino acid residues in determining stability. To illustrate this point, Figure 1 shows a schematic representation of a protein with one residue, *i*, labeled. What function gives the total energetic contribution of site *i* to the stability of the protein? If site *i* contributes only through its intrinsic change in free energy relative to the unfolded state ($\Delta G_i^{\text{intrinsic}}$) plus the sum of its pairwise interactions with immediate neighbors, we would write the function as:

Figure 1



Complexity in the interatomic energy function. (a) A common simplification in atomistic energy calculations is to consider the free energy contribution of one site *i* (black) to fold stability to be the intrinsic change at this site plus the sum of pairwise interactions with directly contacting amino acids (blue), as described in Equation 1. This approximation assumes that interactions between more distant residues are negligible (gray). However, the pattern of free energy coupling between amino acid residues does not always follow such a simple rule. (b) As has been measured in several cases, energetic interactions responsible for structure and function may be distributed through the protein in patterns that are not obvious in atomic structures. If so, the potential for many combinatorial networks of inter-residue interactions arises, a possibility that dramatically increases the complexity of the atomistic energy function (Equation 2).

$$\Delta G_i = \Delta G_i^{\text{intrinsic}} + \sum_j \Delta\Delta G_{i,j} \qquad (1)$$

where *j* represents the few other residues directly interacting with *i* (colored in blue) and $\Delta\Delta G_{i,j}$ is the so-called coupling free energy between the pair of residues *i* and *j*. Indeed, many scoring functions in protein design use such a formalism to calculate site-specific energies [6–10,11•,12,13].

However, data from many protein systems suggest that this view is too simple to describe natural proteins; in some cases, amino acids act in cooperative, higher-order units in which residues distant in the atomic structure might nevertheless energetically interact (Figure 1). For example, hot spot residues in the hGHbp [3], antigen recognition sites in antibodies [14] and residues at the active site of serine proteases [15,16] all display thermodynamically coupled interactions with other positions, some of which are located at sites unexpectedly distant in the tertiary structure. Structural studies of these and other proteins [14,17] provide a rationale for explaining such long-range interactions in proteins: the propagated energetic coupling of residues seems to reflect the mechanical coupling of a few amino acid residues, a property necessary in allosteric and signaling proteins for mediating communication at a distance.

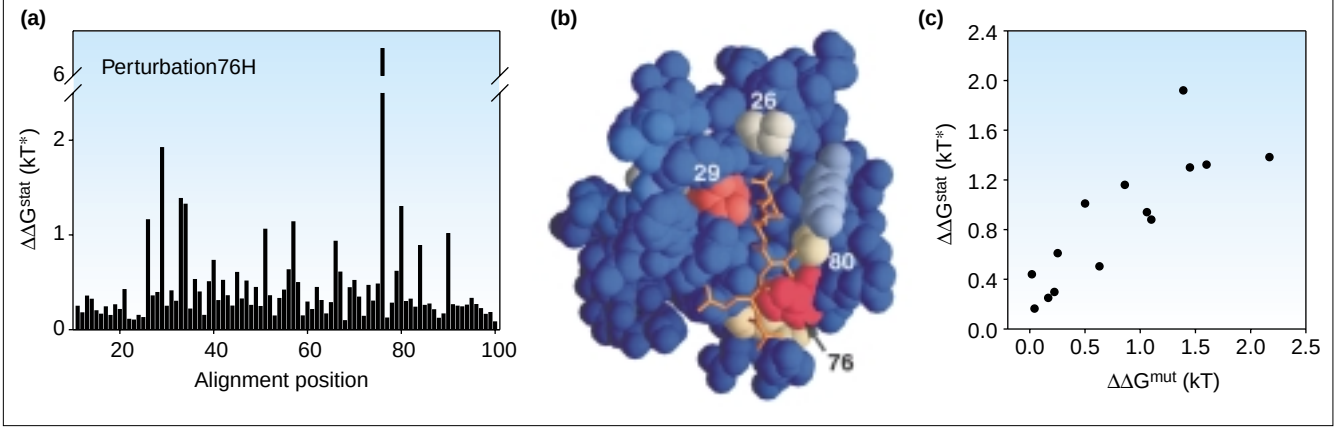
Although these energetic couplings are often interpreted as functional elements in folded proteins, recent work has suggested that these couplings are also determinants of the stability of the folded state [18,19]. These observations, based mainly on dynamics measurements, such as hydrogen exchange rates [20], and on local stability calculations [18], argue that, although folding stability and functional properties seem like distinct manifestations of the free energy, they may be mechanistically inseparable in the protein structure. From the perspective of protein design, such distributed networks of thermodynamic coupling imply that pairwise interactions observed in protein structures cannot necessarily be taken independently. Thus, in the absence of prior knowledge about the pattern or location of thermodynamic coupling, the full potential complexity of the energetics of site *i* (Figure 1) is described by all the possible ways in which site *i* might interact with other sites:

$$\Delta G_i = \Delta G_i^{\text{intrinsic}} + \sum_j \Delta\Delta G_{i,j} + \sum_k \sum_j \Delta^3 G_{i,j,k} + \sum_l \sum_k \sum_j \Delta^4 G_{i,j,k,l} + \ldots \qquad (2)$$

where $\Delta^n G$ represents the *n*-way coupling of residues, for instance, the influence of residue *k* on the coupling between residues *i* and *j*, and so on. The potential complexity of energy parsing in protein structures implied by this description is extraordinary and certainly experimentally intractable. Though several lines of evidence suggest that natural proteins are unlikely to encode much of this potential complexity [21–23,24•,25,26], it serves to illustrate that even precise knowledge of the interaction energy of pairs of atoms in a protein structure taken in isolation may not suffice in describing protein stability.

Given these sources of imprecision and complexity, how can we rationally build proteins and why do design algorithms work? Nearly all scoring functions utilize empirical information learned from protein sequence, structure, and function databases to either improve the calculation of site-specific energies or drastically reduce

Figure 2



A sequence-based method for mapping energetic interactions in proteins. Evolution of a protein fold occurs through random mutagenesis with selection constraints imposed by structure and function. In principle, information about both the importance of individual sites and the degree of interaction between sites may be embedded in large and diverse multiple sequence alignments comprising a protein family [23]. (a) The graph shows the degree of statistical coupling ($\Delta\Delta G^{stat}$) between position 76 in an alignment of 274 PDZ domains and all other positions. $\Delta\Delta G^{stat}$ is an energy-like parameter that measures the degree of co-evolution between two positions. The data show that most positions are only weakly coupled to position 76, a finding consistent with the idea that most positions in proteins are independent of each other [19,21]. However, a few positions show large statistical coupling to position 76. (b) A mapping of the data shown in (a) onto a representative structure of the PDZ family. Position 76 (marked with arrow) is located in the active site of the domain and is involved in mediating interactions with target peptides (shown in stick bonds). The data show that sites evolutionarily coupled to position 76 occur both near and far in the tertiary structure. The color scale ranges from blue (0.33 *kT**) to red (2.33 *kT**). (c) The scatter plot shows the correlation of the statistical energy function derived from sequence analysis with actual thermodynamic coupling measured through double-mutant cycle experiments. The data support the view that the statistical coupling parameter may be a good reporter of the pattern and strength of inter-residue energetic interactions in proteins.

the combinatorial complexity of inter-residue interactions. As might be expected, the accuracy of these knowledge-based potentials depends on the statistical quality (size and unbiased diversity) of the experimental databases. The dramatic growth in databases of both sequences and tertiary structures has allowed much progress in the development of the empirical potentials.

Knowledge-based potentials: information from protein sequences

One powerful source of empirical information about protein stability is contained in the statistics of a multiple sequence alignment of a fold family. Sequence alignments have typically been used to study the evolutionary constraint (or conservation) of sites in a protein family, a parameter given by the distribution of amino acids allowed at each position. By postulating that this constraint reflects the physical chemistry underlying protein structure and function, many studies have attempted to use the pattern of conservation in designing proteins. For example, Lehmann *et al.* [27••] have created hyperstable fungal phytases, proteins that dephosphorylate phytic acid, by choosing the most prevalent residue at each position in a sequence alignment of the family. This 'consensus sequence' approach [28,29] has similarly been used to significantly improve the stability of the SH3 domain [24•], the p53 DNA-binding domain [25] and a GroEL mini-chaperone [26], and to design a WW domain 'prototype' [30]. The successful prediction of sites that control stability in many different proteins strongly supports the basic premise of this approach, that the sequence database usefully encodes information about the global stability of the fold.

What about the problem of combinatorial complexity in interactions between residues? Conservation in a sequence alignment is, by nature, a statistical property of sites taken as if independent of all others. However, some amino acid residues may act in concert and knowledge of such interactions should improve the empirical potential functions that guide protein design. How can we map the location and pattern of thermodynamic coupling in proteins? One sequence-based approach to this problem is to consider that thermodynamic coupling between two sites in a protein should mutually constrain evolution at the two sites if the interaction is important for folding or function. In principle, this mutual constraint should be encoded in the covariance of the amino acid distributions at the two positions in a multiple sequence alignment. Importantly, this approach assumes nothing about the structural proximity or mechanism of the coupling; it simply requires a mutual energetic constraint that forces evolutionary variance at one site to affect the outcome at the second site. On the basis of these principles, Lockless and Ranganathan [23] described a statistical energy function that measures the coupling between pairs of sites in an

alignment and tested this method using the PDZ domain, a small protein interaction module (Figure 2). The statistical energy function predicted both short- and long-range thermodynamic couplings in the domain (Figure 2a,b) that were experimentally verified by mutagenesis (Figure 2c). Using a roughly similar approach, Larson *et al.* [31•] have described an analysis of amino acid covariation in the SH3 domain and again show the effective prediction of groups of substitutions that act cooperatively to stabilize the fold. These studies offer the exciting possibility of systematically mapping thermodynamic coupling in proteins at a scale that was previously inaccessible; it will be interesting to see how this information might be used to devise new functions for protein design that may help better define Nature's rules for building proteins.

Knowledge-based potentials: information from protein structures

Knowledge-based potentials derived from secondary and tertiary structures of proteins comprise key elements of nearly every protein design algorithm. Like sequence-based information, these potentials influence the design process by correcting deficiencies in the interatomic energy function and by constraining the complexity of searching sequence space. Structure-based knowledge potentials include: rotamer libraries derived from known protein structures, which restrict sidechain conformations and provide a backbone-dependent energetic value for each conformation [32,33]; binary patterning [7,34,35,36•,37–39], which imposes bias on the character of amino acids allowed at sites according to knowledge of the hydrophobic or hydrophilic environment; an implicit solvation model that accounts for the hydrophobic effect through calculation of the fractional buried surface area of residues [40] or pairs of residues [41]; secondary structure propensity [8,42–47]; and libraries correlating backbone structures with small segments of protein sequences in the Protein Data Bank [48]. A key aspect of successfully applying knowledge-based potentials has been empirical adjustment of the parameters of the energy function to achieve the best correlation of computed and experimentally observed properties of the designed sequences [10,49]. Iterative adjustment of the energy function through theory and experiment comprises the so-called 'protein design cycle', a method that has now achieved the complete redesign of a globular protein [6,12].

Several groups have reported successful protein design through the application of one or more of these knowledge-based potentials. Sarisky and Mayo [12] have described several redesigned members of the zinc finger ββα fold. The algorithm applied (known as ORBIT) searches for the combination of rotamers on a fixed backbone that minimizes the global free energy of the fold using an elegant strategy for reducing computational time [50]. Binary patterning is included to classify residues as core, surface or boundary; this information is used to adjust the chemical character of each position, a manipulation that helps dramatically reduce the combinatorial complexity of sequences to be searched. Importantly, the computed stabilities of the synthetic sequences correlated with the experimentally determined values, a result suggesting that the scoring function is physically meaningful. Using similar structure-based rotamer libraries and empirical solvation parameters, β turns in two proteins have been replaced [48] and redesigned to enhance stability [51] or to stabilize a domain-swapped dimer [52]. In addition, Isogai *et al.* employed a method that estimates site-specific conformational potentials from tertiary structures to rank amino acid preferences at each site [8,49]; this method was used to build a synthetic member of the globin fold that was monomeric, helical and capable of binding heme. However, the synthetic globin was less ordered than natural globins and was unable to bind oxygen.

One serious limitation in all of these studies is the need to fix the geometry of the backbone in order to reduce computational time searching through combinations of rotamers. This approximation does not allow mainchain adjustments to accommodate sequence variation, although such rearrangements have been experimentally observed in response to mutations of the protein core [22,53]. Two approaches have resulted in the design of proteins that allow alternative backbone conformations, either through explicitly modeled conformations [7,48,52] or by simply not constraining the backbone geometry [35,36•,37,54]. In the former approach, the design algorithms also incorporated information to reduce the computational complexity, such as residue patterning in coiled-coil motifs and sidechain rotamer libraries. The resulting proteins showed excellent agreement between designed and experimental structures [7]. The second approach is remarkable for the simplicity of its scoring potential. In these studies, the binary pattern of hydrophobic and hydrophilic residues was the predominant constraint in the design of combinatorial sequence libraries [35,36•,37,54]. For instance, Kamtekar *et al.* [35] described the design of a combinatorial library of α-helical proteins by constraining only binary patterning; many of these proteins exhibit cooperative unfolding transitions [55•], consistent with the properties of native structures.

A major future goal in protein design is the rational engineering of functional properties such as binding specificity, allosteric regulation and catalysis. Towards this goal, two groups have isolated new functional proteins from combinatorial sequence libraries that incorporate knowledge-based design parameters [36•,37]. The two studies employ similar strategies. Both groups impose a binary pattern on the libraries corresponding to the pattern of hydrophobic and polar residues in the parental enzymes, as well as partial randomization and recombination, whereby portions of the proteins are independently assembled. Both studies conclude that functional proteins are rare in the libraries, a result that highlights the value of knowledge-based design, but also suggests that additional

constraints are necessary. In addition, Voigt *et al.* [56••] have used protein design to test an algorithm for identifying sites of productive recombination events and are able to build β-lactamase variants that support their hypotheses. One study has now reported the construction of 'protozymes', designed proteins with weak but quantifiable catalytic power in mediating hydrolysis of *p*-nitrophenyl acetate [11•]. The key feature of this work is that the active site was designed *de novo* into the surface of *Escherichia coli* thioredoxin, a protein with no intrinsic hydrolytic activity. These studies suggest the exciting possibility that the fundamental principles of catalysis might now be studied not only through the analysis of natural enzymes, but through understanding the engineering principles of building them. Further improvement of empirical potentials for predicting the stability of protein structures may allow rapid progress in this area in the near future.

Conclusions: towards the physical basis of knowledge-based potentials

Key problems of protein design are imprecision in our estimation of the net value of amino acid interactions in folded structures and poor understanding of the complex ways in which energy can be stored in high-order interactions of residues. The knowledge-based approach provides a short cut to rational protein design by imposing mechanistically unclear but predictive site-specific potentials learned from databases of protein structures and sequences. The success of this approach to date provides significant motivation for future work in at least two areas. First, to define the minimal set of rules for building a protein, a result that may help bound the extraordinary potential complexity of the protein folding problem. Second, to determine the mechanistic origin of knowledge-based potentials, which should help further refine and simplify the potential functions used in design algorithms. Perhaps more importantly, however, a detailed understanding of knowledge-based potentials should lead to a better fundamental understanding of how proteins work.

Acknowledgements

We thank members of the Ranganathan laboratory for critical reading of the manuscript. This work was partially supported by a grant from the Robert A Welch Foundation to RR, who is also a recipient of the Burroughs-Wellcome Fund New Investigator Award in the Basic Pharmacological Sciences and the Mallinckrodt Scholar Award. WPR is a Research Associate and RR is an Assistant Investigator of the Howard Hughes Medical Institute

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Anfinsen CB: Principles that govern the folding of protein chains. *Science* 1973, 181:223-230.

2. Clackson T, Wells JA: A hot spot of binding energy in a hormone-receptor interface. *Science* 1995, 267:383-386.

3. Atwell S, Ultsch M, De Vos AM, Wells JA: Structural plasticity in a remodeled protein-protein interface. *Science* 1997, 278:1125-1128.

4. O'Shea EK, Klemm JD, Kim PS, Alber T: X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* 1991, 254:539-544.

5. Lumb KJ, Kim PS: Measurement of interhelical electrostatic interactions in the GCN4 leucine zipper. *Science* 1995, 268:436-439.

6. Dahiyat BI, Mayo SL: ***De novo*** protein design: fully automated sequence selection. *Science* 1997, 278:82-87.

7. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS: High-resolution protein design with backbone freedom. *Science* 1998, 282:1462-1467.

8. Isogai Y, Ota M, Fujisawa T, Izuno H, Mukai M, Nakamura H, Iizuka T, Nishikawa K: Design and synthesis of a globin fold. *Biochemistry* 1999, 38:7431-7443.

9. Koehl P, Levitt M: Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci USA* 2002, 99:1280-1285.

10. Street AG, Datta D, Gordon DB, Mayo SL: Designing protein beta-sheet surfaces by Z-score optimization. *Phys Rev Lett* 2000, 84:5010-5013.

11. • Bolon DN, Mayo SL: Enzyme-like proteins by computational design. *Proc Natl Acad Sci USA* 2001, 98:14274-14279.
The authors describe the design of 'protozymes', enzyme-like proteins that show moderate but clearly measurable catalysis of hydrolysis of a chosen model substrate, *p*-nitrophenyl acetate. The design method uses information from structure databases to constrain rotamer choices and to impose binary patterning rules, and implements an iterative design strategy to identify regions for active site design and refinement of active site energetics. Importantly, the study demonstrates that catalytic power can be transferred to a noncatalytic protein fold.

12. Sarisky CA, Mayo SL: The beta-beta-alpha fold: explorations in sequence space. *J Mol Biol* 2001, 307:1411-1418.

13. Zou J, Saven JG: Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J Mol Biol* 2000, 296:281-294.

14. Patten PA, Gray NS, Yang PL, Marks CB, Wedemayer GJ, Boniface JJ, Stevens RC, Schultz PG: The immunological evolution of catalysis. *Science* 1996, 271:1086-1091.

15. Hedstrom L, Szilagyi L, Rutter WJ: Converting trypsin to chymotrypsin: the role of surface loops. *Science* 1992, 255:1249-1253.

16. Hedstrom L: Trypsin: a case study in the structural determinants of enzyme specificity. *Biol Chem* 1996, 377:465-470.

17. Perona JJ, Hedstrom L, Rutter WJ, Fletterick RJ: Structural origins of substrate discrimination in trypsin and chymotrypsin. *Biochemistry* 1995, 34:1489-1499.

18. Freire E: The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc Natl Acad Sci USA* 1999, 96:10118-10122.

19. Luque I, Leavitt SA, Freire E: The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu Rev Biophys Biomol Struct* 2002, 31:235-256.

20. Williams DC Jr, Benjamin DC, Poljak RJ, Rule GS: Global changes in amide hydrogen exchange rates for a protein antigen in complex with three different antibodies. *J Mol Biol* 1996, 257:866-876.

21. Skinner MM, Terwilliger TC: Potential use of additivity of mutational effects in simplifying protein engineering. *Proc Natl Acad Sci USA* 1996, 93:10753-10757.

22. Baldwin E, Xu J, Hajiseyedjavadi O, Baase WA, Matthews BW: Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. *J Mol Biol* 1996, 259:542-559.

23. Lockless SW, Ranganathan R: Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999, 286:295-299.

24. • Rath A, Davidson AR: The design of a hyperstable mutant of the Abp1p SH3 domain by sequence alignment analysis. *Protein Sci* 2000, 9:2457-2469.
Conservation-based approaches [24•,25,26] employ the hypothesis that evolution has typically selected the most stabilizing amino acid at each position in the sequence and that substitution of atypical residues with the most frequently occurring counterpart at each site in the alignment should result in a stabilized protein. Although this approach dramatically reduces the search space for stabilizing mutations, its predictive power is relatively low – in each case, less than 50% of the substitutions resulted in significant stabilization of the protein.

25. Nikolova PV, Henckel J, Lane DP, Fersht AR: Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc Natl Acad Sci USA* 1998, 95:14675-14680.

26. Wang Q, Buckle AM, Foster NW, Johnson CM, Fersht AR: Design of highly stable functional GroEL minichaperones. *Protein Sci* 1999, 8:2186-2193.

27. •• Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon AP: From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng* 2000, 13:49-57.

In this paper, a novel phytase enzyme is created using sequence conservation as the only design parameter, resulting in an enzyme that is substantially stabilized and exhibits a higher functional temperature optimum than the proteins comprising the alignment. Substitution at sites was made to the most prevalent residue at each site in a multiple sequence alignment, the empirical rule of the consensus sequence approach. It is interesting to consider that imposing the consensus pattern in selecting designed sequences may be a form of cryptically introducing high-order coupling between amino acid residues. That is, the probability of the most likely residue at one site may often be coupled to the most likely residue at another site.

28. Lehmann M, Pasamontes L, Lassen SF, Wyss M: The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta* 2000, 1543:408-415.

29. Lehmann M, Wyss M: Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr Opin Biotechnol* 2001, 12:371-375.

30. Macias MJ, Gervais V, Civera C, Oschkinat H: Structural analysis of WW domains and design of a WW prototype. *Nat Struct Biol* 2000, 7:375-379.

31. • Larson SM, Di Nardo AA, Davidson AR: 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* 2000, 303:433-446.

In this work, the authors employed information about evolutionary covariance between positions in a multiple sequence alignment to design energetically interacting mutations in the SH3 domain hydrophobic core. The authors demonstrate that this approach works and additionally show that the covariance seems to have predictive power for the physical interaction of amino acids; 80% of significantly covarying residues are close in tertiary structure. The remaining 20% may also be of interest, because residue covariation may also result from long-range energetic interaction.

32. Janin J, Wodak S, Levitt M, Maigret B: Conformation of amino acid side-chains in proteins. *J Mol Biol* 1978, 125:357-386.

33. Ponder JW, Richards FM: Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 1987, 193:775-791.

34. Marshall SA, Mayo SL: Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 2001, 305:619-631.

35. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH: Protein design by binary patterning of polar and nonpolar amino acids. *Science* 1993, 262:1680-1685.

36. • Taylor SV, Walter KU, Kast P, Hilvert D: Searching sequence space for protein catalysts. *Proc Natl Acad Sci USA* 2001, 98:10596-10601.

A design strategy employing sequence conservation, binary patterning and a reduced amino acid alphabet was coupled with genetic selection to isolate catalytically active chorismate mutase enzymes from a randomized library. The rare occurrence of functional proteins emphasizes the importance of using knowledge to reduce the complexity of the library; it would have been virtually impossible to isolate sequences encoding functional proteins if all positions were allowed to vary without constraint.

37. Silverman JA, Balakrishnan R, Harbury PB: Reverse engineering the (beta/alpha)8 barrel fold. *Proc Natl Acad Sci USA* 2001, 98:3092-3097.

38. Li H, Cocco MJ, Steitz TA, Engelman DM: Conversion of phospholamban into a soluble pentameric helical bundle. *Biochemistry* 2001, 40:6636-6645.

39. Frank S, Kammerer RA, Hellstern S, Pegoraro S, Stetefeld J, Lustig A, Moroder L, Engel J: Toward a high-resolution structure of phospholamban: design of soluble transmembrane domain mutants. *Biochemistry* 2000, 39:6825-6831.

40. Eisenberg D, McLachlan AD: Solvation energy in protein folding and binding. *Nature* 1986, 319:199-203.

41. Street AG, Mayo SL: Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* 1998, 3:253-258.

42. Minor DL Jr, Kim PS: Measurement of the beta-sheet-forming propensities of amino acids. *Nature* 1994, 367:660-663.

43. Smith CK, Withka JM, Regan L: A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* 1994, 33:5510-5517.

44. Chakrabartty A, Baldwin RL: Stability of alpha-helices. *Adv Protein Chem* 1995, 46:141-176.

45. Dalal S, Balasubramanian S, Regan L: Protein alchemy: changing beta-sheet into alpha-helix. *Nat Struct Biol* 1997, 4:548-552.

46. Dalal S, Regan L: Understanding the sequence determinants of conformational switching using protein design. *Protein Sci* 2000, 9:1651-1659.

47. Bishop B, Koay DC, Sartorelli AC, Regan L: Reengineering granulocyte colony-stimulating factor for enhanced stability. *J Biol Chem* 2001, 276:33465-33470.

48. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D: Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J Mol Biol* 2002, 315:471-477.

49. Ota M, Nishikawa K: Assessment of pseudo-energy potentials by the best-five test: a new use of the three-dimensional profiles of proteins. *Protein Eng* 1997, 10:339-351.

50. Dahiyat BI, Mayo SL: Protein design automation. *Protein Sci* 1996, 5:895-903.

51. Nauli S, Kuhlman B, Baker D: Computer-based redesign of a protein folding pathway. *Nat Struct Biol* 2001, 8:602-605.

52. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D: Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc Natl Acad Sci USA* 2001, 98:10687-10691.

53. Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW: The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* 1993, 262:1715-1718.

54. West MW, Wang W, Patterson J, Mancias JD, Beasley JR, Hecht MH: ***De novo*** amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci USA* 1999, 96:11211-11216.

55. • Roy S, Hecht MH: Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry* 2000, 39:4603-4607.

The authors show that binary patterning can specify cooperatively denaturing small folds. This result suggests that the pattern of hydrophobicity and hydrophilicity encodes substantial information about protein topology. It is interesting to note that binary patterning is a way of specifying coupling between positions, because the chemical character of a residue at one site imposes a bias in the character of residues at one or more other sites.

56. •• Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH: Protein building blocks preserved by recombination. *Nat Struct Biol* 2002, 9:553-558.

An interesting paper that describes a theory for predicting sites of recombination in proteins that are likely to be structurally and functionally tolerated. Experiments verify that the method in fact provides a novel method for the combinatorial synthesis of functionally active synthetic proteins. Interestingly, the predictions do not always coincide with domain boundaries, possibly suggesting a new definition of what is a minimal building block for natural proteins.