ELSEVIER

Knowledge-based potentials in protein design

Alan M Poole and Rama Ranganathan

Knowledge-based potentials are statistical parameters derived from databases of known protein properties that empirically capture aspects of the physical chemistry of protein structure and function. These potentials play a key role in protein design by improving the accuracy of physics-based models of interatomic interactions and enhancing the computational efficiency of the design process by limiting the complexity of searching sequence space. Recently, knowledge-based potentials (in isolation or in combination with physics-based potentials) have been applied to the modification of existing protein function, the redesign of natural protein folds and the complete design of a non-natural protein fold. In addition, knowledge-based potentials appear to be providing important information about the global topology of amino acid interactions in natural proteins. A detailed study of the methods and products of these protein design efforts promises to greatly expand our understanding of proteins and the evolutionary process that created them.

Addresses
Howard Hughes Medical Institute, Department of Pharmacology and the Green Comprehensive Center Division for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA

Corresponding author: Ranganathan, Rama (rama.ranganathan@utsouthwestern.edu)

Current Opinion in Structural Biology 2006, **16**:508–513

This review comes from a themed issue on
Engineering and design
Edited by William F DeGrado and Derek N Woolfson

Available online 14th July 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.06.013

Introduction

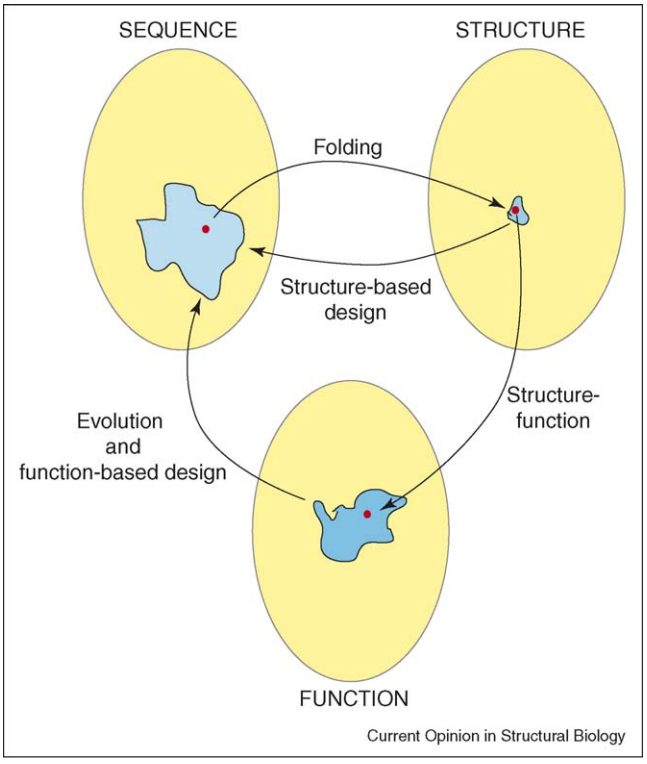
Efforts in protein design are motivated by two overlapping goals: developing novel molecular reagents that expand upon and improve the function of natural proteins; and testing our fundamental understanding of how proteins fold and function. Both aims, however, share the problem of contending with the dramatic complexity of searching sequence and conformation space. For even small proteins (say, comprising 100 amino acids), 20^{100} possible sequences exist and each sequence may have an enormous number of conformational states. As has been discussed extensively [1], the vastness of these spaces precludes exhaustive computational or experimental searching. In addition, a computational protein design algorithm must also be able to score sequences for the one(s) most likely to adopt the desired structure and/or function; in essence, this process involves finding the sequences that lie at the global minimum of an energy function that models the physical interactions between atoms. In an ideal case, this energy function would include just the appropriate combination of fundamental physical forces and computation would be efficient enough to identify the global minima in a reasonable amount of time. At this time, such a target energy function for protein design does not exist; instead, investigators incorporate information from databases of protein structure and function into knowledge-based potentials (KBPs) that serve to both increase the accuracy of scoring functions and restrict the conformational search problem. In this review, we begin with several exciting new advances in protein design, grouped by the primary sources of the KBPs utilized. In addition, we discuss new applications of KBPs to define the quantity of information encoded in protein sequences by mapping the statistically independent components of proteins. This work suggests that KBPs may help define the basic evolutionary design rules of proteins.

To facilitate discussion, Figure 1 shows a schematic representation of relationships between protein sequence, structure and function. The three 'worlds' represent our databases of acquired knowledge about proteins, and mappings between them represent key topical problems of protein analysis and design. A particular sequence (red dot in the sequence world) adopts a native state ensemble in the structure world (the protein folding problem), which then displays a particular biochemical activity (the structure-function problem). Knowledge-based protein design can be abstractly defined as the attempt to use information from the distribution of natural proteins in any of these three worlds to identify new, non-natural sequences that encode desired target regions in the structure and function worlds. Evolution is represented in this schematic as the natural process of function-based design through random sequence variation and selection based on functional fitness. In this view, the sequence space corresponding to a natural protein family is intimately linked to the nature of the evolutionary fitness function. Thus, an account of the selection constraints that shape the sequence space distribution of protein families may help in defining general principles, if any, of the evolutionary design of proteins.

Knowledge-based potentials derived from structural databases

The ever-expanding Protein Data Bank (PDB) is a valuable database of structural information about proteins and comprises the basis of many KBPs. Potentials derived

Figure 1



Schematic representation of the protein sequence-structure-function problem. Each 'world' (see labels) abstractly represents the space of possible sequences, structures and biochemical functions. Thus, one natural sequence (red dot in the sequence world) corresponds to one native state ensemble and in turn corresponds to one region of the function world. Mappings between the worlds represent various core problems in protein analysis and design, all of which ultimately depend on understanding the global pattern of physical forces between atoms specified by the evolutionary fitness function. A key concept is that KBPs represent statistical parameters extracted from analyses of the distribution of natural proteins in any of these worlds; these potentials can provide key information that contributes to efficiently selecting well-folded and possibly functional non-natural sequences.

from structural libraries include, but are not limited to: solvation potentials that bias hydrophobic and hydrophilic residues to the core and surface of a protein, respectively [2–4]; binary patterning [5,6]; rotamer libraries that constrain amino acid sidechain orientations to those commonly observed in natural proteins [7]; libraries of short residue fragments observed in the PDB [8]; hydrogen-bonding potentials [9]; and electrostatic potentials [10]. These various potentials are often combined in various ways with and without approximate physical potentials to produce hybrid energy functions used to score sequences during protein design. As an example, RosettaDesign, a highly successful algorithm for protein design, has an energy function consisting of a 6,12 Lennard–Jones potential, an implicit solvation model, a hydrogen-bonding potential, backbone-dependent rotamer probabilities, ϕ, ψ space-dependent amino acid probabilities, and an electrostatics term [11].

The power of these structure-derived potentials is evident in the remarkable advances in protein design efforts in recent years. In an approach that is inspiring for its simplicity, Hecht *et al.* [12] have designed both α-helix bundles and β-sheet proteins using no information other than the binary patterning of polar and non-polar residues commonly observed in α helices and β strands. Recently, they discovered that these four-helix bundles can possess intrinsic esterase activity and suggested that the activity of these minimally designed sequences serve as a benchmark for more sophisticated design attempts [13]. Matching polar and non-polar residues to compatible environments also anchored the principles that enabled the design of water-soluble analogs of phospholamban [14] and the KcsA potassium channel [15]. Although the potentials used to select amino acid identities were more complicated than simple binary patterning, the selection of positions to mutate was determined based on solvent exposure, and the swapping of hydrophobic surface residues for hydrophilic ones was a key aspect in both of the design processes. Coiled-coil proteins have also been amenable to design using relatively simple patterning of hydrophobic, charged and polar residues [16]. Recent work has used these simple design elements to identify coiled-coil interaction determinants and, through both positive and negative design of a small number of residues, has engineered specificity for homodimeric and heterodimeric interfaces [17]. This work in several different systems shows how remarkably simple design rules can, in some cases, yield successful protein designs.

Several groups have combined structure-based potentials, physics-based potentials, and efficient sequence and conformation searching to redesign portions of proteins and recover novel sequences with desired functions. Recent achievements include engineering triose phosphate isomerase activity into a non-enzymatic protein [18•], and redesigning the ligand-binding pocket of periplasmic binding proteins to bind Zn^{2+} [19] and several different small molecules [20,21]. In each of these examples, a limited set of residues near the binding region were allowed to vary in the design process, resulting in 5–22 mutations. A methodologically similar effort by Shifman and Mayo [22,23] yielded a calmodulin variant (13 mutations) with similar affinity for its target, but significantly increased specificity when considering other natural targets. Whereas the above efforts used deterministic dead-end elimination algorithms to identify the single best sequences, Kono and Saven [24] developed a knowledge-based statistical method that estimates the site-specific amino acid probabilities for each position in a protein. By choosing the most probable amino acid for most positions in the sequence (some amino acids considered essential for function were not allowed to vary), a redox-active

minimal rubredoxin mimic [25] and a monomeric helical dinuclear metalloprotein [26] were constructed and functionally verified.

The complete design of protein folds has also been successfully demonstrated [11,27–29]. These results convincingly show that stable proteins with well-packed cores can be designed computationally. In fact, preliminary observations suggest that some of these designed proteins may actually be significantly more stable than their natural counterparts [11]. Perhaps the most exciting protein design using structure-based potentials is the report from Kuhlman *et al.* [8] describing a novel protein, Top7. The structure of Top7 is highly similar (~1.2 Å) to its design model and represents a fold topology not found in natural proteins. Top 7 was designed from models built with three- and nine-residue fragments found in the PDB, employing iterative cycling between sequence optimization (evaluating rotamer selections with a combination of physics- and structure-derived potentials) and backbone conformation adjustment. The resulting protein is extremely stable ($T_m > 98$ °C), which is consistent with the target design principle of optimal atomic packing.

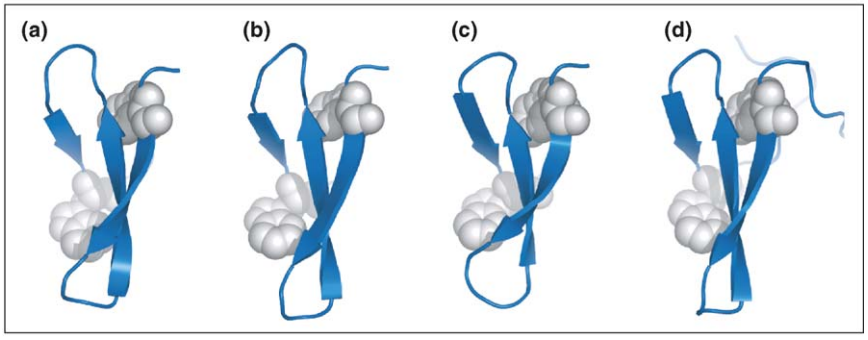
Knowledge-based potentials derived from sequence databases

When sufficiently large and diverse, multiple sequence alignments (MSAs) can provide a measure of the sequence constraints for protein folding and function. Basic principles of molecular evolution suggest that, if the identity of an amino acid at a particular position is important for fitness, it should be conserved in related proteins. This hypothesis is the basis of the 'consensus sequence' approach to modifying protein function, whereby a particular sequence is altered by changing one or more amino acids in a particular protein to the most commonly encountered amino acid in an MSA of the protein family. Two recent examples of this approach are a β-lactamase variant containing eight mutations whose melting temperature was 9 °C higher than that of the natural molecule [30] and a triple mutant of an IgG1 C_H3 domain that was stabilized by 10 °C [31]. Evolutionary relationships within a protein family have also been used to infer ancestral protein sequences and functions — a fascinating application of KBPs to understand the paths of evolution. Ancestral gene reconstruction for fluorescent proteins [32], serine proteases [33], steroid receptors [34], bacterial elongation factors [35] and archosaur rhodopsin [36] not only has been very successful at generating functional protein sequences, but also has led these authors to suggest that ancient proteins are less specialized [32–34].

How much of the total information content of protein sequences can be extracted from analyses of MSAs? Classical studies show that the amino acid sequence contains all the information necessary for protein folding and function [37], but, in principle, this information might be held in complex, high-order statistical interactions between amino acid positions that are not necessarily evident in simple models of sequence conservation that treat sites as if they were statistically independent of one another. To develop a more physically consistent model of sequence conservation that takes into account the cooperative interactions of residues, several groups have now developed algorithms for capturing the correlated evolution of residues from MSAs [38–43]. One such method, statistical coupling analysis (SCA), provides a global analysis of conserved evolutionary interactions between pairs of sequence positions in large and diverse MSAs, and shows strong consistency with experimental data [38,44–46]. Recently, Socolich *et al.* [47••] demonstrated that the statistical co-evolution information extracted by the SCA is sufficient to specify the fold of a small protein — the WW domain (Figure 2). In a companion study, Russ *et al.* [48••] showed that the SCA-designed artificial proteins exhibited the same range of functional properties (ligand affinity and specificity) as natural WW domains. This approach for creating artificial proteins is interesting in that no structural or physiochemical information was used in the design, and that only a small fraction of the possible pairwise inter-residue constraints were incorporated. More importantly, this work argues that the number of constraints necessary to define a protein's fold and function may be far less than theoretically possible.

Other sequence-based designs support the notion that the constraints required to specify protein structure and function are relatively sparse. Gene-shuffling experiments start with several different sequences and use recombination at the level of predefined blocks or single amino acids to create libraries of chimeric sequences. One recent experiment used three cytochrome P450 sequences (~65% amino acid identity to each other) and recombined eight sequence blocks to successfully create a library containing significant numbers of folded and functional chimeric P450 proteins [49]. This fragment-based method relies on a structure-based computational analysis to identify the optimal recombination sites for sequence variation. Genetic shuffling at the level of single amino acids has also been shown to generate functional proteins with significant levels of sequence diversity [50]. An interesting example of the combination of these methods is a project to develop a glyphosate tolerance gene [51•]. Castle *et al.* began with three glyphosate N-acetyltransferase (GAT) genes and conducted iterative rounds of fragment-based or site-independent recombination followed by functional selection, whereby the most improved variants were used as the parents of the next recombination reaction. After eleven total iterations and the incorporation of diversity from four other sequences, enzyme efficiency was improved ~10,000-fold and high levels of glyphosate resistance were conferred on transformed plants. That these gene-shuffling techniques generate libraries with many folded and functional

Figure 2



Structural comparison of natural and designed WW proteins. **(a–c)** Cartoon representations of three natural WW domains from dystrophin (PDB code 1eg3), YAP65 (PDB code 1k9r) and Nedd4 (PDB code 1i5h). **(d)** Solution NMR structure of an artificial WW domain, CC45 (PDB code 1ymz), designed using SCA-based rules of sequence co-evolution extracted from an MSA of many WW domains [47••]. Quantitative comparisons of the root mean squared deviation of backbone atoms show that the CC45 structure is about as different from natural WW domains as natural domains are from each other.

proteins provides further support for the notion that natural proteins encode a great deal of near energetic independence between residues.

Knowledge-based potentials derived from functional databases

Databases that correlate functional variation with sequence and/or structure variations can significantly aid design efforts by targeting sequence variation to specific determinants of protein function. For example, Yang *et al.* [52] introduced a Ca^{2+}-binding site into a non-Ca^{2+}-binding protein. The key aspect of this study is the computational survey of naturally occurring Ca^{2+}-binding motifs in the context of their target protein to identify a Ca^{2+}-binding site that is structurally and chemically compatible with both Ca^{2+} binding and the surrounding protein environment. Another study described the redesign of a promiscuous sesquiterpene synthase into several novel enzymes with greatly improved specificities [53••]. This experiment involved creating and characterizing several enzyme libraries, each containing a single degenerate position. The product profiles from the variant enzymes were then recombined computationally to predict a set of mutations that would bias the product distribution toward the desired distributions. Construction and testing of the novel proteins revealed that the predicted mutations did indeed alter product specificity as desired. Again, these studies highlight the energetic simplicity of proteins; tuning function appears to be possible by varying a few positions in a combinatorial fashion to yield phenotypic diversity.

What can artificial proteins tell us about natural proteins?

In addition to providing practical avenues for the design of novel reagents, a major goal of protein design is to understand the basic principles of the folding and function of natural proteins. Indeed, how well do designed proteins really recapitulate the design of natural proteins? As reviewed above, several studies have now described the design of proteins that exhibit biochemical activities similar to those of natural proteins, and there is considerable promise for the total computational design of even complex functional proteins. However, we suggest that a proper validation of our understanding of the design principles of natural proteins must go well beyond the creation of artificial proteins that recapitulate structural stability and biochemical activity *in vitro*. For example, proteins must function within complex physiological environments in which negative and positive selection for activities may exist [54], as well as other as yet unclear constraints. In addition, a large body of work now argues that natural proteins are selected for robustness — meaning that they tolerate mutation at many sites without dramatic changes in function. Nevertheless, natural proteins also somehow maintain the capacity for rapid adaptive change through mutation of a few critical sites. This type of functional response to mutation implies a heterogeneous architecture of natural proteins in which many residues contribute independently or not at all, and a few emerge as cooperative determinants of structure and function. The success of current protein design methods based largely on optimizing the packing of atoms suggests that these proposed natural design properties are not necessary conditions for producing well-folded and perhaps even functional artificial proteins. However, it may be that they are necessary for designing stable and functional proteins that can also evolutionarily compete with their natural counterparts. It will be interesting to carry out a comparative analysis of the functional fitness of artificial proteins built through different design strategies. Thanks to the significant recent advances in diverse protein design methods, the reagents to drive these exciting experiments are now becoming available.

Acknowledgements

We thank members of the Ranganathan laboratory for critical reading of the manuscript and Bill Russ for many productive discussions. RR acknowledges support from the Mallinckrodt Scholar award, the Keck Future Initiatives award and the Robert A Welch Foundation, and is an investigator of the Howard Hughes Medical Institute. AMP was supported in part by the Perot Family Foundation and National Institutes of Health grant T32-GM008014 to the UT Southwestern Medical Scientist Training Program.

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. Russ WP, Ranganathan R: **Knowledge-based potential functions in protein design**. *Curr Opin Struct Biol* 2002, **12**:447-452.

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

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

4. Zhang N, Zeng C, Wingreen NS: **Fast accurate evaluation of protein solvent exposure**. *Proteins* 2004, **57**:565-576.

5. 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.

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

7. Dunbrack RL Jr: **Rotamer libraries in the 21st century**. *Curr Opin Struct Biol* 2002, **12**:431-440.

8. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy**. *Science* 2003, **302**:1364-1368.

9. Kortemme T, Morozov AV, Baker D: **An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes**. *J Mol Biol* 2003, **326**:1239-1259.

10. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins**. *Proteins* 1999, **34**:82-95.

11. Dantas G, Kuhlman B, Callender D, Wong M, Baker D: **A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins**. *J Mol Biol* 2003, **332**:449-460.

12. Hecht MH, Das A, Go A, Bradley LH, Wei Y: ***De novo* proteins from designed combinatorial libraries**. *Protein Sci* 2004, **13**:1711-1723.

13. Wei Y, Hecht MH: **Enzyme-like proteins from an unselected library of designed amino acid sequences**. *Protein Eng Des Sel* 2004, **17**:67-75.

14. Slovic AM, Summa CM, Lear JD, DeGrado WF: **Computational design of a water-soluble analog of phospholamban**. *Protein Sci* 2003, **12**:337-348.

15. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF: **Computational design of water-soluble analogues of the potassium channel KcsA**. *Proc Natl Acad Sci USA* 2004, **101**:1828-1833.

16. Woolfson DN: **The design of coiled-coil structures and assemblies**. *Adv Protein Chem* 2005, **70**:79-112.

17. Havranek JJ, Harbury PB: **Automated design of specificity in molecular recognition**. *Nat Struct Biol* 2003, **10**:45-52.

18. • Dwyer MA, Looger LL, Hellinga HW: **Computational design of a biologically active enzyme**. *Science* 2004, **304**:1967-1971.

This work is noteworthy because it demonstrates that significant alteration or acquisition of protein function can be accomplished through the redesign of a relatively small proportion of a protein.

19. Dwyer MA, Looger LL, Hellinga HW: **Computational design of a Zn^{2+} receptor that controls bacterial gene expression**. *Proc Natl Acad Sci USA* 2003, **100**:11255-11260.

20. Allert M, Rizk SS, Looger LL, Hellinga HW: **Computational design of receptors for an organophosphate surrogate of the nerve agent soman**. *Proc Natl Acad Sci USA* 2004, **101**:7907-7912.

21. Looger LL, Dwyer MA, Smith JJ, Hellinga HW: **Computational design of receptor and sensor proteins with novel functions**. *Nature* 2003, **423**:185-190.

22. Shifman JM, Mayo SL: **Exploring the origins of binding specificity through the computational redesign of calmodulin**. *Proc Natl Acad Sci USA* 2003, **100**:13274-13279.

23. Shifman JM, Mayo SL: **Modulating calmodulin binding specificity through computational protein design**. *J Mol Biol* 2002, **323**:417-423.

24. Kono H, Saven JG: **Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure**. *J Mol Biol* 2001, **306**:607-628.

25. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG, Degrado WF: ***De novo* design of a redox-active minimal rubredoxin mimic**. *J Am Chem Soc* 2005, **127**:5804-5805.

26. Calhoun JR, Kono H, Lahr S, Wang W, DeGrado WF, Saven JG: **Computational design and characterization of a monomeric helical dinuclear metalloprotein**. *J Mol Biol* 2003, **334**:1101-1115.

27. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL *et al.*: ***De novo* backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure**. *J Mol Biol* 2003, **325**:163-174.

28. Kraemer-Pecore CM, Lecomte JT, Desjarlais JR: **A *de novo* redesign of the WW domain**. *Protein Sci* 2003, **12**:2194-2205.

29. Isogai Y, Ito Y, Ikeya T, Shiro Y, Ota M: **Design of lambda Cro fold: solution structure of a monomeric variant of the *de novo* protein**. *J Mol Biol* 2005, **354**:801-814.

30. Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, Schellenberger V: **Construction of stabilized proteins by combinatorial consensus mutagenesis**. *Protein Eng Des Sel* 2004, **17**:787-793.

31. Demarest SJ, Rogers J, Hansen G: **Optimization of the antibody C(H)3 domain by residue frequency analysis of IgG sequences**. *J Mol Biol* 2004, **335**:41-48.

32. Ugalde JA, Chang BS, Matz MV: **Evolution of coral pigments recreated**. *Science* 2004, **305**:1433.

33. Wouters MA, Liu K, Riek P, Husain A: **A despecialization step underlying evolution of a family of serine proteases**. *Mol Cell* 2003, **12**:343-354.

34. Thornton JW, Need E, Crews D: **Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling**. *Science* 2003, **301**:1714-1717.

35. Gaucher EA, Thomson JM, Burgan MF, Benner SA: **Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins**. *Nature* 2003, **425**:285-288.

36. Chang BS, Jonsson K, Kazmi MA, Donoghue MJ, Sakmar TP: **Recreating a functional ancestral archosaur visual pigment**. *Mol Biol Evol* 2002, **19**:1483-1489.

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

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

39. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles**. *J Mol Biol* 2001, **312**:289-307.

40. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families**. *J Mol Biol* 1996, **257**:342-358.

41. 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.

42. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis**. *Mol Biol Evol* 2000, **17**:164-178.

43. Dima RI, Thirumalai D: **Determination of network of residues that regulate allostery in protein families using sequence analysis**. *Protein Sci* 2006, **15**:258-268.

44. Suel GM, Lockless SW, Wall MA, Ranganathan R: **Evolutionarily conserved networks of residues mediate allosteric communication in proteins**. *Nat Struct Biol* 2003, **10**:59-69.

45. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R: **Structural determinants of allosteric ligand activation in RXR heterodimers**. *Cell* 2004, **116**:417-429.

46. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R: **Allosteric determinants in guanine nucleotide-binding proteins**. *Proc Natl Acad Sci USA* 2003, **100**:14445-14450.

47. •• Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R: **Evolutionary information for specifying a protein fold**. *Nature* 2005, **437**:512-518.
A small set of sequence constraints derived from the conservation of residues and the co-evolution of pairs of residues in an MSA of WW domains were found to be sufficient to create a library of folded proteins. These proteins were shown to adopt the same fold and possess the same range of thermostability as natural WW domains.

48. •• Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R: **Natural-like function in artificial WW domains**. *Nature* 2005, **437**:579-583.
In a companion paper to that of Socolich *et al.* [45], the functional properties (peptide-binding affinity and specificity) of artificial WW domains were found to be consistent with those of natural WW domains. This finding illustrates that a very small set of constraints are sufficient to encode function in a protein.

49. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH: **Structure-guided recombination creates an artificial family of cytochromes P450**. *PLoS Biol* 2006, **4**:e112.

50. Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J: **Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently**. *Nat Biotechnol* 2002, **20**:1251-1255.

51. • Castle LA, Siehl DL, Gorton R, Patten PA, Chen YH, Bertain S, Cho HJ, Duck N, Wong J, Liu D *et al.*: **Discovery and directed evolution of a glyphosate tolerance gene**. *Science* 2004, **304**:1151-1154.
In a very successful venture, this group was able to blend the iterative incorporation of sequence diversity and functional selection to find a genetic variant of GAT with enzymatic efficiency four log orders higher than that of any of the parent sequences. The new enzyme was also effective *in vivo* and rescued plants from lethal doses of herbicide.

52. Yang W, Wilkins AL, Ye Y, Liu ZR, Li SY, Urbauer JL, Hellinga HW, Kearney A, van der Merwe PA, Yang JJ: **Design of a calcium-binding protein with desired structure in a cell adhesion molecule**. *J Am Chem Soc* 2005, **127**:2085-2093.

53. •• Yoshikuni Y, Ferrin TE, Keasling JD: **Designed divergent evolution of enzyme function**. *Nature* 2006, **440**:1078-1082.
Libraries containing degenerate codons at a single amino acid position were constructed and characterized based on their enzymatic products. The authors then demonstrated that the product profiles could be rationally combined to predict enzyme variants with specific activity profiles. This construction of a specialized functional database illustrates how functional information can be cleverly incorporated into protein design.

54. Zarrinpar A, Park SH, Lim WA: **Optimization of specificity in a cellular protein interaction network by negative selection**. *Nature* 2003, **426**:676-680.