ELSEVIER

Simulating protein evolution in sequence and structure space

Yu Xia[1] and Michael Levitt[2]

Naturally occurring proteins comprise a special subset of all plausible sequences and structures selected through evolution. Simulating protein evolution with simplified and all-atom models has shed light on the evolutionary dynamics of protein populations, the nature of evolved sequences and structures, and the extent to which today's proteins are shaped by selection pressures on folding, structure and function. Extensive mapping of the native structure, stability and folding rate in sequence space using lattice proteins has revealed organizational principles of the sequence/structure map important for evolutionary dynamics. Evolutionary simulations with lattice proteins have highlighted the importance of fitness landscapes, evolutionary mechanisms, population dynamics and sequence space entropy in shaping the generic properties of proteins. Finally, evolutionary-like simulations with all-atom models, in particular computational protein design, have helped identify the dominant selection pressures on naturally occurring protein sequences and structures.

Addresses
[1]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA
e-mail: yuxia@csb.yale.edu
[2]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA
e-mail: michael.levitt@stanford.edu

Current Opinion in Structural Biology 2004, **14**:202–207

This review comes from a themed issue on
Theory and simulation
Edited by Joel Janin and Thomas Simonson

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2004.03.001

Introduction

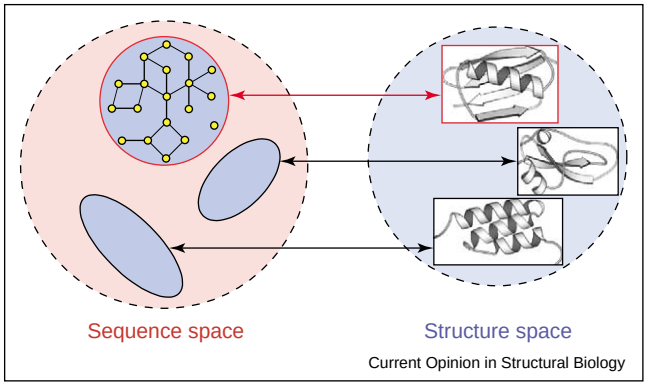
A protein is an evolved molecular machine capable of self-assembly and reliable functioning in a fluctuating environment. Understanding how these remarkable properties arise as the result of evolution is central to the development of a protein folding theory [1], and can lead to better strategies for protein design and structure prediction. One way to understand protein evolution is to directly simulate the evolutionary dynamics of a protein population, by combining general theories of molecular evolution with physical models of protein folding. Such simulation studies, when combined with experiments and sequence/structure database analyses, can help delineate major evolutionary factors responsible for shaping today's proteins. The following sample of questions can be addressed by evolutionary simulation: how does the large-scale organization of protein sequence and structure space affect protein evolution; can dynamic simulations of protein evolution explain the generic properties of proteins, such as mutational robustness and marginal stability; which one is a more dominant selection pressure, stability or folding rate; are popular protein folds selected for favorable innate properties or is it largely a stochastic process?

In this review, we survey recent computational studies of protein evolution using lattice and all-atom models. Several related topics are beyond the scope of this article and have been reviewed elsewhere; these include computational studies of RNA evolution [2] and the theory of *in vitro* directed evolution [3,4].

Simplified protein models and protein evolution

In the first part of this review, we survey recent evolutionary studies using simplified protein models. Simplified models of proteins, such as spin glass models [5], and two- and three-dimensional lattice models [6,7], have led to many insights into protein folding mechanisms. When applied to evolutionary studies, simplified models have a number of advantages. First, there are only a few tunable parameters that define the fundamental physics of the system. Second, the fitness landscape can be defined in a precise way. Different fitness criteria have been used to address different evolutionary questions. For example, fitness of a protein sequence may depend on whether it folds to a target structure (i.e. the structure is a unique global energy minimum for the sequence), native structure stability, folding rate, ability to bind a model ligand, or a combination of the above. The fitness value can be either continuous or binary. In the latter case, all viable proteins that meet the fitness criterion are considered to be equally fit, and mutations are either neutral or lethal. Third, the mapping between sequence and structure space can be carried out extensively as a result of the reduced model complexity, and few approximations are required to set up the protocols for evolutionary simulation. These properties make simplified models ideal for studying how the generic properties of proteins are shaped by evolution. Of course, due to the nature of simplified models, the interpretation of results requires considerable care to exclude model artifacts. A comprehensive review on applying simple exact models to problems in protein evolution is given in [8••]; here, we focus on recent developments.

Figure 1



Cartoon of the mapping between protein sequence space and structure space. Each point in sequence space represents a protein sequence. The mapping is asymmetric, in that many protein sequences share the same fold. Different protein folds correspond to different regions of protein sequence space with varying shapes. A line connects two sequences that share the same native fold and differ by a single mutation. The resulting connected graph in protein sequence space is called a neutral network. In this cartoon, two neutral networks share the same fold. The large network has eighteen sequences and the small network has one sequence. Sequences that share the same protein fold differ in their stability, folding rate and evolutionary preference. The evolutionary dynamics of protein sequences and structures can be regarded as the dynamic properties of this mapping.

Mapping protein sequence/structure relationships with simplified models

The dynamic properties of protein evolution are defined, to a large degree, by the static properties of protein sequence/structure mapping (Figure 1). For two-dimensional square lattice models with HP (hydrophobic/polar) sequences, it is possible to construct a complete mapping from sequence space to structure space. In addition, various thermodynamic and kinetic properties of protein folding can be computed for all sequences that fold to the same structure. A number of observations emerge from these studies.

First, sequence space is plastic, in that many different sequences can fold to the same structure. Many of these sequences form a network interconnected by single point mutations, termed a neutral network [9]. Within each neutral network, sequences are organized around a special prototype sequence with the largest number of viable neighbors in sequence space [10]. As one walks away from the prototype sequence to the fringe of the neutral network, the average protein stability gradually decreases. This funnel-like organization of protein stability has been called a super-funnel [10]. Similar funnel-like distributions have also been observed for the protein folding rate. Though the effects of point mutations on folding rate are often subtle and difficult to predict for the HP model, the large-scale organization of folding rates in sequence space follows a simple funnel-like distribution [11•].

Second, neutral networks of different structures are well separated in sequence space, with rather sparse connections ('bridges') between them [12,13]. A discussion of these rare but important bridges connecting naturally occurring protein folds can be found in [14]. This lattice protein result is in contrast to RNA, for which neutral networks of any two structures can be connected by just a few point mutations [2].

Finally, the number of sequences that fold into a structure is called the designability of the structure. The distribution of designability among structures is highly skewed. The few highly designable structures have protein-like local structures [15] and symmetries [16], and are, on average, more stable and faster folding [17] than other structures. A highly designable structure is compatible with a larger volume of sequence space and has fewer competing structural neighbors [18]. These results are robust when different energy functions are used, as long as they capture the dominant hydrophobic interactions in proteins [19,20]. Based on these results with lattice models, it is postulated that naturally occurring proteins are highly designable [15,21].

More complicated models have also been used to characterize protein sequence/structure relationships. The evolutionary landscape of all proteins with binding pockets, called functional proteins, has been studied with two-dimensional lattice models [22,23]. The study was later extended to three-dimensional tetrahedral lattice models [24], for which exhaustive sampling of structure space is also possible [25]. A number of studies have explored evolutionary landscapes in sequence space with 20 amino acid alphabets [26–31]. In this case, exhaustive characterization of sequence space is no longer possible. Instead, extensive sampling is required.

Simulating protein evolutionary dynamics with simplified models

Depending on the question to be addressed, evolutionary processes can be simulated in different ways. Stochastic optimization methods are often used to approximate the evolutionary process, but they can only model the end results of strong evolutionary selection. Evolutionary dynamics is usually simulated as an adaptive walk on a fitness landscape or a random walk when the fitness landscape becomes flat. With large population size and high mutation rate, however, population effects become important and evolutionary dynamics can be best simulated as a population of slightly different sequences undergoing replication with mutation and/or recombination, and selection. Here, evolutionary selection does not act on a single sequence, but rather on the population of sequence variants interconnected by mutations, termed

quasi-species [32]. Pronounced population dynamics effects have been observed for the simulated evolution of RNA secondary structures and digital organisms [33,34]. Here, we survey recent simulation studies with lattice proteins that address the evolutionary origins of many generic properties of proteins.

Unlike random polymers, a protein folds quickly to the native structure, which corresponds to a pronounced global energy minimum. Stochastic and adaptive optimization with lattice protein sequences have demonstrated that these protein-like properties can arise as a result of evolutionary selection for fast folding and native state stability [7,35]. Even when folding is under kinetic control and the native state is initially not the energetic ground state, the protein sequence could evolve so that the native state is most often the global energy minimum [36].

Proteins are robust to mutation. This property can be explained by population dynamics of evolution. Proteins evolved by population dynamics are more robust to mutation than those evolved by random walk within the viable region [37•]. For an evolving sequence population on a neutral network, certain sequences are evolutionarily preferred at steady state over others, even when all viable sequences share the same fitness. Evolutionary preference for a sequence correlates largely with the number of its viable neighbors [10,11•,38], giving rise to mutational robustness. These results remain unchanged when explicit selection for protein function is considered [39].

Proteins are marginally stable. This can be at least partly understood in terms of sequence space entropy. Due to the high dimensionality of sequence space, most sequences in an evolving population on a neutral network are located near the boundary of the viable region and are thus marginally stable, despite the evolutionary preference for the prototype sequence located at the center [10,38]. As a result of this sequence space entropy effect, functionality consistent with marginal stability tends to be selected by evolution [40].

Are highly designable protein structures favored by evolution? By allowing both sequences and structures to evolve during simulation, it was found that the relative frequency of highly designable structures in the population increases as a result of population dynamics of evolution [41]. During evolutionary simulation, the protein population explores structure space only for a short amount of time. As the average fitness of the population increases, sequences are increasingly confined to isolated regions of sequence space and the population quickly hones in on a single native structure that is preserved for the rest of the simulation [42]. These results indicate that highly designable structures are favored by evolution as a result of population dynamics; at the same time, protein structure is far less mutable than sequence during evolution [42].

Finally, several studies have considered the effects of recombination events. When sequences evolve while maintaining the native structure, recombination events can drastically increase the evolutionary preference for the prototype sequence [38]. In an evolving population of different protein structures, recombination allows rapid and effective exploration of structure space [13].

Evolutionary simulations can also be used to quantitatively test evolutionary hypotheses. For example, by simulating neutral evolution with a fixed structure, it was found that the requirement of maintaining a constant structure leads to an unexpected non-Poissonian substitution process [43,44].

All-atom protein models and protein evolution

The same procedures for studying protein evolution with simplified models can, in principle, be applied to all-atom models as well. In practice, however, because of the complexity of all-atom models, major assumptions are needed to make these procedures feasible. For example, it is no longer possible to exhaustively sample sequence and structure space. Protein thermodynamic properties can only be calculated approximately. In computational protein design, evolutionary-like optimization protocols are often used, whereby the stability of a protein is optimized by successive mutational operations, at the same time keeping the protein structure fixed. The aim of these procedures is not to faithfully reproduce the evolutionary trajectory of naturally occurring proteins, but rather to capture the end result of an evolutionary process with dominant selection pressure for optimal stability. Recent computational design of a new fold and subsequent experimental validation indicate that, despite many approximations, current design methods are accurate enough for direct comparisons with experiments [45••]. Methods for computational protein design are reviewed in [46]; here, we focus on recent insights from evolutionary studies with all-atom models, in particular, what makes naturally occurring protein sequences and structures so special among all plausible sequences and structures.

Evolutionary selection of naturally occurring protein sequences

What role does selection for native state stability play in the evolution of protein sequences? Sequences computationally optimized for native state stability converge to a region of sequence space close to the wild-type sequence and the conservation patterns in the designed sequences are similar to those in naturally occurring proteins [47–49,50••,51,52]. These results suggest that selection for

stability plays a major role in the evolution of naturally occurring protein sequences.

Are the folding rates of small proteins extensively optimized by natural selection? This question can be addressed using computational protein design. Residues that play a consistent role in the transition states of SH3 domain homologs tend to be optimized for native stability and vice versa [53]. Furthermore, for many small proteins with diverse topologies, sequences computationally optimized for thermodynamic stability, when synthesized, were found to often fold as fast as, and in many cases even faster than, the wild-type sequences [46,54]. Taken together, these results suggest that fast folding is a by-product of natural selection for thermodynamic stability, at least for the many small proteins studied so far.

Evolutionary selection of naturally occurring protein structures

Are naturally occurring protein folds highly designable? Because the designability of a structure is directly correlated with the density of structures when they are mapped to sequence space [18], a complete demonstration of the designability principle requires extensive sampling of structure space, a formidable task even for proteins of modest size. Several attempts have been made to approximate designability by an easily computable geometrical measure of the structure [15,55–59]. The results are encouraging, but currently there is no consensus on how good these approximations are. Recently, Tang and colleagues [60,61•] extensively sampled structure space for small protein fragments and four-helix bundles, and found that many of the highly designable structures generated are indeed similar to some naturally occurring proteins.

Current all-atom computational protein design methods can be used to address the designability problem. By using all-atom protein design, it was found that the size of sequence space compatible with a fold correlates well with the sequence diversity observed in nature [50••,51]. This indicates that naturally occurring protein folds differ in their designability and that the volume of sequence space compatible with a fold has been greatly exploited by evolution to generate diverse sequences.

Despite the promising progress of the above studies, much work is needed to assess the full extent to which the designability principle can be applied to naturally occurring protein folds.

Recently, several stochastic evolutionary models have been proposed to account for the fact that the genomic occurrence of protein folds follows a power-law distribution [62–67]. These models assume that genome evolution involves duplication of existing genes, introduction of new genes and deletion of existing genes. It should be emphasized that these models are not incompatible with the hypothesis that popular protein folds are selected for favorable innate properties, such as high designability or, in general, the ability to accommodate new functions. Indeed, introducing a fitness function for folds into the model does not change the overall power-law picture, but does affect the exact occurrence of an individual fold [64]. The occurrence of individual folds is likely to be mainly the result of selection for function [68], even though the global distribution of the genomic occurrence of folds can be simulated by an evolutionary process that is largely stochastic.

Conclusions

By directly simulating protein evolution with simplified and all-atom models, much has been learned about the evolutionary mechanisms behind the selection of modern protein sequences and structures. Despite recent progress, significant challenges remain. First, better forcefields and more efficient ways of sampling sequence and structure space are always needed. Second, more computational and experimental work is required to fully resolve the current debate on many aspects of protein evolution. Third, the scope of current simulations can be expanded to address new evolutionary problems. For example, most evolutionary simulations deal with a protein population in isolation. In reality, a protein performs its function by interacting with other proteins, nucleic acids and substrates. The current simulation protocol can be expanded to include these interactions. Most importantly, the continuing combination of evolutionary simulations, bioinformatics and experiments is crucial to solving the many fundamental problems in protein evolution.

Update

A recent review on the designability principle can be found in [69]. Simulated evolution with model proteins was performed in two recent studies. The first study [70] attempted to reconcile the two views of protein fold evolution (the designability principle and the divergent model of fold evolution). Highly designable folds were found to be evolutionarily favored in a divergent model of protein evolution previously shown to reproduce the power-law distribution of genomic fold occurrence. Furthermore, eukaryotic-only folds were found to be, on average, more designable than prokaryotic-only folds, according to an approximate measure of designability. The second study [71] investigated the relationship between the selection pressure on stability and the ability of a model protein to evolve ligand-binding function. It was found that proteins evolve function more efficiently when the selection pressure on stability is low, and that it is easier to enhance stability while maintaining high function than to enhance function while maintaining high stability.

Acknowledgements

We thank Tanya Raschke for helpful comments. YX thanks Mark Gerstein for support.

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. Onuchic JN, Wolynes PG: **Theory of protein folding**. *Curr Opin Struct Biol* 2004, **14**:70-75.

2. Fontana W: **Modelling 'evo-devo' with RNA**. *Bioessays* 2002, **24**:1164-1177.

3. Voigt CA, Kauffman S, Wang ZG: **Rational evolutionary design: the theory of *in vitro* protein evolution**. *Adv Protein Chem* 2000, **55**:79-160.

4. Arnold FH: **Combinatorial and computational challenges for biocatalyst design**. *Nature* 2001, **409**:253-257.

5. Bryngelson JD, Wolynes PG: **Spin glasses and the statistical mechanics of protein folding**. *Proc Natl Acad Sci USA* 1987, **84**:7524-7528.

6. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding–a perspective from simple exact models**. *Protein Sci* 1995, **4**:561-602.

7. Shakhnovich EI: **Theoretical studies of protein-folding thermodynamics and kinetics**. *Curr Opin Struct Biol* 1997, **7**:29-40.

8. •• Chan HS, Bornberg-Bauer E: **Perspectives on protein evolution from simple exact models**. *Appl Bioinformatics* 2002, **1**:121-144.
An in-depth and comprehensive review of the use of simple exact models to study protein evolution.

9. Bornberg-Bauer E: **How are model protein structures distributed in sequence space?** *Biophys J* 1997, **73**:2393-2403.

10. Bornberg-Bauer E, Chan HS: **Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space**. *Proc Natl Acad Sci USA* 1999, **96**:10689-10694.

11. • Xia Y, Levitt M: **Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution**. *Proteins* 2004, **55**:107-114.
Using a two-dimensional HP-like model, the authors calculated the stability and folding rate of all sequences that fold to the same 24-mer structure. It was found that the distributions of stability and folding rate in sequence space are different, but both distributions are funnel like. The existence of stability and folding rate funnels in sequence space limits the range of possible dynamic behavior of protein evolution.

12. Bornberg-Bauer E: **Randomness, structural uniqueness, modularity, and neutral evolution in sequence space of model proteins**. *Z Phys Chem* 2002, **216**:139-154.

13. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS: **Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes**. *Proc Natl Acad Sci USA* 2002, **99**:809-814.

14. Grishin NV: **Fold change in evolution of protein structures**. *J Struct Biol* 2001, **134**:167-185.

15. Li H, Helling R, Tang C, Wingreen N: **Emergence of preferred structures in a simple model of protein folding**. *Science* 1996, **273**:666-669.

16. Wang T, Miller J, Wingreen NS, Tang C, Dill KA: **Symmetry and designability for lattice protein models**. *J Chem Phys* 2000, **113**:8329-8336.

17. Melin R, Li H, Wingreen N, Tang C: **Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study**. *J Chem Phys* 1999, **110**:1252-1262.

18. Li H, Tang C, Wingreen NS: **Are protein folds atypical?** *Proc Natl Acad Sci USA* 1998, **95**:4987-4990.

19. Buchler NEG, Goldstein RA: **Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus**. *J Chem Phys* 2000, **112**:2533-2547.

20. Li H, Tang C, Wingreen NS: **Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix**. *Proteins* 2002, **49**:403-412.

21. Helling R, Li H, Melin R, Miller J, Wingreen N, Zeng C, Tang C: **The designability of protein structures**. *J Mol Graph Model* 2001, **19**:157-167.

22. Hirst JD: **The evolutionary landscape of functional model proteins**. *Protein Eng* 1999, **12**:721-726.

23. Blackburne BP, Hirst JD: **Evolution of functional model proteins**. *J Chem Phys* 2001, **115**:1935-1942.

24. Blackburne BP, Hirst JD: **Three-dimensional functional model proteins: structure, function and evolution**. *J Chem Phys* 2003, **119**:3453-3460.

25. Hinds DA, Levitt M: **From structure to sequence and back again**. *J Mol Biol* 1996, **258**:201-209.

26. Bastolla U, Roman HE, Vendruscolo M: **Neutral evolution of model proteins: diffusion in sequence space and overdispersion**. *J Theor Biol* 1999, **200**:49-64.

27. Bastolla U, Vendruscolo M, Roman HE: **Structurally constrained protein evolution: results from a lattice simulation**. *Eur Phys J B* 2000, **15**:385-397.

28. Tiana G, Broglia RA, Shakhnovich EI: **Hiking in the energy landscape in sequence space: a bumpy road to good folders**. *Proteins* 2000, **39**:244-251.

29. Tiana G, Broglia RA, Shakhnovich EI: **Energy profile of the space of model protein sequences**. *J Biol Phys* 2001, **27**:147-159.

30. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF: **Exploring protein sequence space using knowledge-based potentials**. *J Theor Biol* 2001, **212**:35-46.

31. Aita T, Ota M, Husimi Y: **An *in silico* exploration of the neutral network in protein sequence space**. *J Theor Biol* 2003, **221**:599-613.

32. Eigen M: **Self-organization of matter and the evolution of biological macromolecules**. *Naturwissenschaften* 1971, **58**:465-523.

33. Fontana W, Schuster P: **Continuity in evolution: on the nature of transitions**. *Science* 1998, **280**:1451-1455.

34. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C: **Evolution of digital organisms at high mutation rates leads to survival of the flattest**. *Nature* 2001, **412**:331-333.

35. Mirny LA, Abkevich VI, Shakhnovich EI: **How evolution makes proteins fold quickly**. *Proc Natl Acad Sci USA* 1998, **95**:4976-4981.

36. Govindarajan S, Goldstein RA: **On the thermodynamic hypothesis of protein folding**. *Proc Natl Acad Sci USA* 1998, **95**:5545-5549.

37. • Taverna DM, Goldstein RA: **Why are proteins so robust to site mutations?** *J Mol Biol* 2002, **315**:479-484.
The authors studied two models of protein sequence evolution. In the first model, a viable sequence diffuses randomly over the range of allowed sequences. In the second model, a sequence population evolves over the range of allowed sequences via mutation, selection and reproduction. It was found that evolved proteins in the second model are more robust to site mutations than those in the first model. This result suggests that mutational robustness is a population dynamics effect.

38. Xia Y, Levitt M: **Roles of mutation and recombination in the evolution of protein thermodynamics**. *Proc Natl Acad Sci USA* 2002, **99**:10382-10387.

39. Sasaki TN, Sasai M: **Correlation between the conformation space and the sequence space of peptide chain**. *J Biol Phys* 2002, **28**:483-492.

40. Taverna DM, Goldstein RA: **Why are proteins marginally stable?** *Proteins* 2002, **46**:105-109.

41. Taverna DM, Goldstein RA: **The distribution of structures in evolving protein populations**. *Biopolymers* 2000, **53**:1-8.

42. Williams PD, Pollock DD, Goldstein RA: **Evolution of functionality in lattice proteins**. *J Mol Graph Model* 2001, **19**:150-156.

43. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Lack of self-averaging in neutral evolution of proteins**. *Phys Rev Lett* 2002, **89**:208101.

44. Bastolla U, Porto M, Eduardo Roman MH, Vendruscolo MH: **Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution**. *J Mol Evol* 2003, **56**:243-254.

45. •• 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.
Computational protein design was used to create a new protein sequence with a novel topology. The subsequent X-ray structure of the protein is very close to the design model, with a root mean square difference of 1.2 Å for 93 residues.

46. Kuhlman B, Baker D: **Exploring folding free energy landscapes using computational protein design**. *Curr Opin Struct Biol* 2004, **14**:89-95.

47. Koehl P, Levitt M: ***De novo* protein design. II. Plasticity in sequence space**. *J Mol Biol* 1999, **293**:1183-1193.

48. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structures**. *Proc Natl Acad Sci USA* 2000, **97**:10383-10388.

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

50. •• Koehl P, Levitt M: **Protein topology and stability define the space of allowed sequences**. *Proc Natl Acad Sci USA* 2002, **99**:1280-1285.
Using all-atom computational protein design and a physical energy function, the authors estimated the size of sequence space compatible with a fold using multiple alignments of the designed sequences. It was found that the volume of sequence space compatible with a fold is similar in size to that observed in nature. This is a promising method for identifying highly designable folds.

51. Larson SM, England JL, Desjarlais JR, Pande VS: **Thoroughly sampling sequence space: large-scale protein design of structural ensembles**. *Protein Sci* 2002, **11**:2804-2813.

52. Jaramillo A, Wernisch L, Hery S, Wodak SJ: **Folding free energy function selects native-like protein sequences in the core but not on the surface**. *Proc Natl Acad Sci USA* 2002, **99**:13554-13559.

53. Larson SM, Pande VS: **Sequence optimization for native state stability determines the evolution and folding kinetics of a small protein**. *J Mol Biol* 2003, **332**:275-286.

54. Gillespie B, Vu DM, Shah PS, Marshall SA, Dyer RB, Mayo SL, Plaxco KW: **NMR and temperature-jump measurements of *de novo* designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics**. *J Mol Biol* 2003, **330**:813-819.

55. Kussell EL, Shakhnovich EI: **Analytical approach to the protein design problem**. *Phys Rev Lett* 1999, **83**:4437-4440.

56. Emberly EG, Miller J, Zeng C, Wingreen NS, Tang C: **Identifying proteins of high designability via surface-exposure patterns**. *Proteins* 2002, **47**:295-304.

57. England JL, Shakhnovich EI: **Structural determinant of protein designability**. *Phys Rev Lett* 2003, **90**:218101.

58. England JL, Shakhnovich BE, Shakhnovich EI: **Natural selection of more designable folds: a mechanism for thermophilic adaptation**. *Proc Natl Acad Sci USA* 2003, **100**:8727-8731.

59. Yahyanejad M, Kardar M, Tang C: **Structure space of model proteins: a principal component analysis**. *J Chem Phys* 2003, **118**:4277-4284.

60. Miller J, Zeng C, Wingreen NS, Tang C: **Emergence of highly designable protein-backbone conformations in an off-lattice model**. *Proteins* 2002, **47**:506-512.

61. • Emberly EG, Wingreen NS, Tang C: **Designability of alpha-helical proteins**. *Proc Natl Acad Sci USA* 2002, **99**:11163-11168.
The authors extensively sampled structure space for all compact four-helix bundles connected by short turns. Highly designable structures were identified from this ensemble, most of which resemble known four-helix-bundle folds. The few novel ones can serve as targets for protein design.

62. Huynen MA, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes**. *Mol Biol Evol* 1998, **15**:583-589.

63. Yanai I, Camacho CJ, DeLisi C: **Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification**. *Phys Rev Lett* 2000, **85**:2641-2644.

64. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model**. *J Mol Biol* 2001, **313**:673-681.

65. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution**. *Nature* 2002, **420**:218-223.

66. Dokholyan NV, Shakhnovich B, Shakhnovich EI: **Expanding protein universe and its origin from the biological Big Bang**. *Proc Natl Acad Sci USA* 2002, **99**:14132-14136.

67. Deeds EJ, Dokholyan NV, Shakhnovich EI: **Protein evolution within a structural space**. *Biophys J* 2003, **85**:2962-2972.

68. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire**. *Science* 2003, **300**:1701-1703.

69. Wingreen NS, Li H, Tang C: **Designability and thermal stability of protein structures**. *Polymer* 2004, **45**:699-705.

70. Tiana G, Shakhnovich BE, Dokholyan NV, Shakhnovich EI: **Imprint of evolution on protein structures**. *Proc Natl Acad Sci USA* 2004, **101**:2846-2851.

71. Bloom JD, Wilke CO, Arnold FH, Adami C: **Stability and the evolvability of function in a model protein**. *Biophys J* 2004, in press.