REVIEW

Perspectives on protein evolution from simple exact models

Hue Sun Chan[1] and Erich Bornberg-Bauer[2]

[1]Protein Engineering Network of Centres of Excellence, Department of Biochemistry, and Department of Medical Genetics and Microbiology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; [2]School of Biological Sciences, The University of Manchester, Manchester, UK

Abstract: Understanding the evolution of biopolymers is important to rationalise the directed and undirected design of functional molecules. Large scale experiments or detailed computational studies are often impractical. Therefore, simple model systems, such as RNA secondary structure and lattice proteins have been adapted to study general statistical and topological features of genotype (sequence) to phenotype (structure) maps. We review findings from such models that address aspects of thermodynamic and mutational robustness, neutral evolution and recombination of proteins. We compare various modelling approaches, and discuss their generality, parameter dependency and experimental verifications of their predictions. The most striking observation is the universal emergence of neutral nets – sets of phenotypically identical genotypes that are interconnected by series of point mutations. However, fast adaptation by point mutations appears to be problematic for proteins. This may explain why proteins appear to be more specific while RNA is rather versatile. This could even be the reason why RNA had to evolve before proteins. Similar principles of biological organisation are reflected in sequence and structure databases of real proteins. Insights gained from modelling are useful for designing more efficient database organisation and search strategies.

Keywords: simple models, neutral evolution, mutational stability, fitness landscapes, sequence databases

Abbreviations:

SEM	simple exact model
REM	random energy model
MJ	Miyazawa-Jernigan potential

Motivation

The study of evolution by computational biophysical models is roughly 15 years old. The scientific origin of this approach is a confluence of two relatively recent developments: (1) extensive applications of physical simple exact models (SEMs) of biopolymers, and (2) the utilisation of landscape concepts borrowed from other areas of theoretical biology and biophysical chemistry. From a vantage point of computation technology, however, these developments could only bear fruit in evolutionary research through the advent of powerful and inexpensive computers in the late 1980s. Many concepts emerging from this approach suggest new questions to be posed, and more efficient ways to exploit bioinformatics data. Examples include the clustering of protein families and the rationalisation of the marginal thermodynamic stabilities of many natural proteins.

Evolutionary optimisation acts on populations and therefore their simulation requires the computation of large ensembles of sequences and structures. The purpose of this paper is to introduce the reader to the rationale of the SEM/landscape approach to evolution, and to discuss results that bear on bioinformatics research, particularly those relevant to search strategies for sequence and structure databases. After a brief survey of general principles of protein evolution and a formal description of simple exact models, our main focus will be on the sequence-structure relations (mappings between sequence and shape or conformational spaces) of proteins under a variety of evolutionary fitness criteria. Examples of explicit simulations of population dynamics will also be presented.

Correspondence: Erich Bornberg-Bauer, School of Biological Sciences, The University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester M13 9PT, UK; tel +44 161 275 7396; fax +44 161 275 5082; email ebb@bioinf.man.ac.uk
Hue Sun Chan, Department of Biochemistry, University of Toronto Faculty of Medicine, 1 King's College Circle, MSB 5207, Toronto, Ontario, M5S 1A8, Canada; tel +1 416 978 2697; fax +1 416 978 8548; email chan@arrhenius.med.utoronto.ca

Simple exact models and related theoretical constructs

SEMs of proteins are useful for investigating the relationship between sequences and structures from basic physical principles, as Dill et al (1995, p 561–2) have stated:

> … simple exact models … can address questions of general principle. Such questions are often difficult to address by other means, through experiments, atomic simulation, Monte Carlo partial sampling, or approximate theoretical models. 'Simple' models have few arbitrary parameters. 'Exact' models have partition functions from which physical properties can be computed without further assumption or approximations.

SEMs are simple representations of biopolymers. For most lattice protein models, each residue is simply a bead on a regular lattice, inter-residue virtual bonds are of equal length and bond angles can take only a few discrete values. An obvious advantage of SEMs is their computability. Their conformational search spaces are significantly reduced vis-à-vis that of chains configured in continuous space. Lattice discretisation enables the use of integers and binary contact maps for efficient computations. Therefore, SEMs can be used to model sufficiently large collections of sequences and structures necessary for addressing issues in evolution. Despite their extreme simplification, however, conformational search problems in several SEMs such as the HP model have been proven to be NP-complete (Patterson and Przytycka 1995; Crescenzi et al 1998). Therefore, in evolutionary applications, their usage is limited to exhaustive enumerations of sequence and conformational spaces of short chains on low coordination lattices, non-comprehensive sampling of sequence space or approximate folding algorithms. This is in contrast to RNA, for which the secondary structure model can be exactly computed within polynomial time by taking advantage of a recursive algorithm (Nussinov and Jacobson 1980; Zuker and Stiegler 1981). Only salient features of SEMs are briefly summarised below, as more detailed reviews of SEM properties are available elsewhere (Chan and Dill 1993; Bryngelson et al 1995; Dill et al 1995; Karplus and Šali 1995; Shakhnovich 1996; Thirumalai and Woodson 1996; Dill and Chan 1997; Pande et al 1997; Chan et al 2002).

SEMs with small alphabets

The letters of an alphabet are the different residue types in a model. Having only two letters, the HP (hydrophobic-polar) model (Lau and Dill 1989; Dill et al 1995) may be viewed as the simplest heteropolymer model (see Figure 1a). Physically, the HP model assumes that hydrophobic forces account for the compact shapes (Chan and Dill 1989)

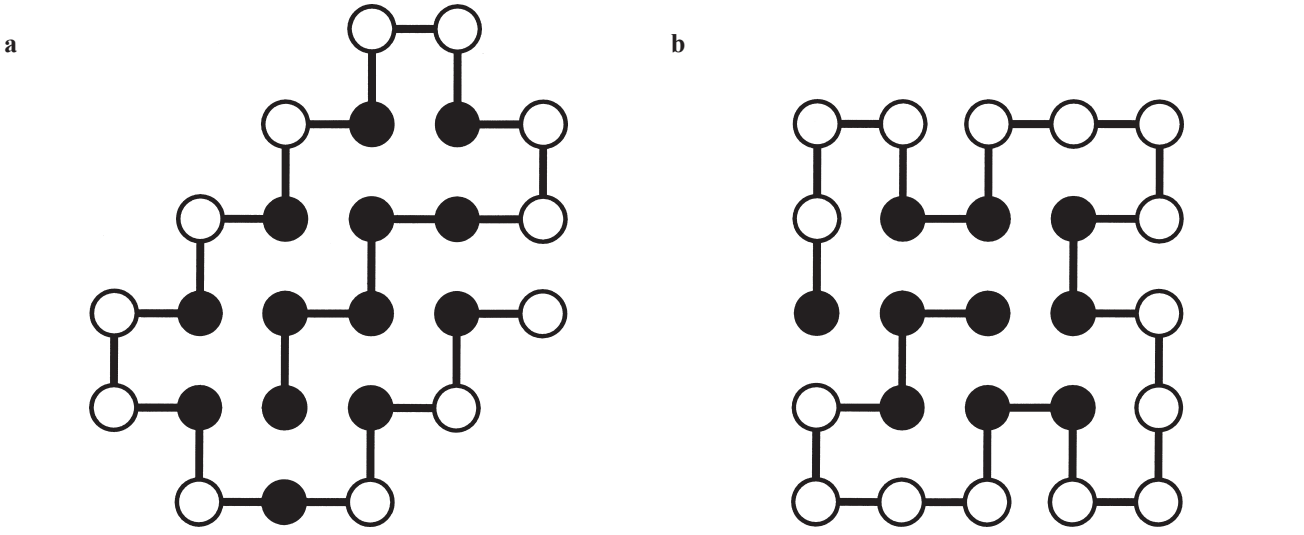

Figure 1 Simple exact models of proteins. **(a)** The most frequently encoded structure in the standard two-dimensional HP model of chain length $n = 25$ is shown by one of the 326 sequences that have it as the unique ground-state conformation (adapted from Irbäck and Troein 2002). Hydrophobic (H) and polar (P) monomers are depicted as filled and open circles respectively. For this HP sequence, the maximum number of favourable hydrophobic-hydrophobic contacts it can form is 13; and this is achievable only if the sequence adopts the structure shown. All 5 768 299 665 square-lattice conformations an $n = 25$ chain can possibly adopt were taken into account in this determination (Irbäck and Troein 2002). **(b)** The most frequently encoded structure in the modified HP model ($n = 25$) of Xia and Levitt (2002) that restricts conformational variation to the 1081 maximally compact conformations. The model potential function is the same as that of Li et al (1996). The sequence shown is one of the 67 615 encoding sequences in the model (adapted from Xia and Levitt 2002). It should be noted, however, that the structure here is not the lowest-energy one for the given sequence if conformations other than 5 × 5 squares can be adopted, as in (a). Conformations searches confined to 5 × 5 squares have been employed extensively in SEM study of evolution using a variety of interaction schemes (see Buchler and Goldstein 1999; Taverna and Goldstein 2002b).

of globular proteins and determine in large measure their backbone arrangements. In many applications, HP chains are configured on two-dimensional square or three-dimensional simple cubic lattices. There is only one type of stabilising interaction in the HP model, namely when two hydrophobic residues (monomers) are spatial nearest neighbours on the lattice (ie they form a contact) but are not nearest neighbours along the chain sequence. This favourable interaction is assigned a negative contact energy. Hydrophobic-polar and polar-polar contacts as well as contacts between either residue type with the solvent are taken to be neutral, ie have zero contact energy. Most randomly generated amino acid sequences do not behave like natural proteins, because the latter are products of natural selection. Likewise, most randomly generated sequences of H and P residues in the HP model do not fold to a single (unique) lowest-energy (ground-state) conformation. For example, only 2.42% (6349) of all (262 144) possible HP sequences of chain length $n = 18$ have a unique ground-state conformation on a square lattice (Chan and Dill 1996). Other sequences have *gemisch* ground states with multiple lowest-energy conformations (Dill et al 1995). The percentage of square-lattice HP sequences that are unique varies within a narrow range between 2.11% and 2.60% for chain lengths $n = 12$ through $n = 25$ (Chan and Dill 1994; Irbäck and Troein 2002).

A variation of the HP model has been applied by Tang and coworkers (Li et al 1996). They used a two-letter HP alphabet that has additional attracting interactions between the H and P residues, similar to the energies in one of the database-derived Miyazawa-Jernigan (MJ) contact potentials (Miyazawa and Jernigan 1985). By restricting conformational searches to the 103 346 maximally compact conformations not related by rotations and inversions on the simple cubic lattice (ie 1/48 of the 4 960 608 conformations including rotations and inversions confined to a $3 \times 3 \times 3$ cube first enumerated by Chan and Dill 1990), Tang and coworkers were able to identify the ground-state structures (within the restricted conformational space) of all $2^{27} = 134\,217\,728$ HP sequences of length $n = 27$. Analogous computations were performed for a 6×6 square lattice. Most recently, in a computational *tour de force*, the approach has been extended by Tang and coworkers (Cejtin et al 2002) to the $2^{36} = 68\,719\,476\,736$ HP sequences of length $n = 36$ configured within a $4 \times 3 \times 3$ cube. The total number of such maximally compact conformations is 84 731 192, as was first enumerated by Pande et al (1994b).

To increase interaction heterogeneity (Wolynes 1997; Micheletti et al 1998; Chan 1999; Wang and Wang 1999), the HP model has been extended to more complex potentials such as the HPNX and hHYX potentials (Bornberg-Bauer 1997a; Backofen et al 1999; Renner and Bornberg-Bauer 1997; Buchler and Goldstein 1999). Some of these generalisations were motivated by contact potentials derived from experimental data (Crippen 1991). However, sequence spaces become exponentially larger and their searches computationally much more intensive when there are more letters in the model alphabet. This often makes it necessary to use more sophisticated constraints-based algorithms (Backofen et al 1999) or restrict conformational searches to maximally compact conformations (Buchler and Goldstein 1999) (cf Figure 1b).

A long standing modelling question is the extent to which the HP and other alphabets with fewer than 20 letters resembles natural proteins. It has been proposed that a relatively small alphabet of ca. 3 to 8 amino acid residues may be sufficient to construct most structural scaffolds while very sophisticated functional adaptation requires only subtle fine tuning (Lim and Sauer 1991; Dill et al 1995). It is also a common belief that primordial protein alphabets consisted of fewer than 20 amino acids (Crick 1968; Wong 1975; Knight et al 1999). During the early stages of evolution, ground-state degeneracies might not have been a severe drawback as the primordial proteins might have been stabilised by their interactions with RNAs. In that case, marginally functional proteins could have predated more specialised ones with more well-defined native structures. Even so, it is inconceivable that a functional genetic code can consist of only two types of amino acids. Furthermore, the functional selection experiment of Baker and coworkers on SH3 mutants using phage display (Riddle et al 1997) suggests that a higher degree of diversity of amino acid types may be needed rather for function than for stability. Apparently, certain thresholds of interaction heterogeneity are necessary for protein stability and function. To our knowledge, 7 is the smallest number of amino acid types needed to encode for a real functional protein (Sicheri and Yang 1995; Chan 1999).

In view of these biophysical considerations, the HP model should not be misinterpreted as representing proteins made up of only two specific amino acids such as leucine and arginine. Rather, HP and similar constructs should be viewed as intuitive 'projections' of a high-dimensional sequence space onto computationally more tractable lower-dimensional sequence spaces, with, for example, H and P

representing two *classes* of amino acids. In assessing simplified protein models, it is important to take their limitations into account and focus on whether the given model is adequate for the particular issue at hand. A certain level of simplification may be appropriate for a class of questions but not for others. The additive interaction scheme embodied by the HP model does not appear to be sufficient for the observed thermodynamic cooperativity of real proteins (Chan 1998, 2000; Shimizu and Chan 2002). Nonetheless, at our current level of understanding, the HP potential is a useful model of an exhaustive sequence-structure *map* for the study of evolution (Chan et al 2002; Cui et al 2002).

Unless noted otherwise, HP results presented below are obtained from exact enumeration data of the $2^{18} = 262\,144$ HP sequences of length $n = 18$ on square lattices. The number of conformations accessible to each such sequence is equal to 5 808 335.

Models using more complex lattice representations

Geometrically more realistic lattice models have been developed by Skolnick and Kolinski and coworkers for studying folding kinetics and structure prediction (Skolnick and Kolinski 1990; Kolinski et al 1991; Chan et al 2002). Although conformational sampling in these models are more efficient than all-atom representations, they are computationally intensive, and their interaction schemes are rather complex. As a result, they have not been applied extensively to study evolutionary issues that require a broad coverage of the mapping between amino acid sequences and their folded structures.

To address evolution-related issues, Hinds and Levitt have developed a diamond lattice based method that uses exhaustive conformational searches through compact structures to address the inverse folding problem (Hinds and Levitt 1994, 1996). This approach, however, has not been applied to construct extensive sequence-structure maps.

Models using more complex interaction schemes

Several interaction schemes have been used in simple lattice protein chain models. In the pioneering work of Gō and coworkers, a different interaction scheme is postulated for each different folded native structure, such that only contacts that occur in a given native structure are favoured and non-native intra-chain contacts that do not occur in the given native structure is either neutral or disfavoured. However, since the Gō construction simply fabricates an ad hoc interaction scheme based on the given target structure without regard to sequence information, this approach does not provide a mapping between sequence and structure. Consequently, Gō models cannot address how evolutionary changes in sequences may affect the structures they encode (Chan et al 2002).

Contact-based schemes assign interaction energies to individual pairs of contacting positions along the chain (eg the Gō construction is a contact-based scheme). Sometimes these energies are sampled according to a certain presumed distribution such as that of the random energy model (REM) (Bryngelson and Wolynes 1989; Buchler and Goldstein 1999). Contact-based interaction schemes assign energies by contacting positions without regard to the identities of the pair of residues involved. Hence contact-based interaction schemes are referred to as uncorrelated (Chan and Dill 1996). Since the energetics of such schemes are not determined by the one-dimensional sequence information, it is problematic to connect them to a sequence-structure map analogous to that of real proteins. Their utility to evolutionary investigation is thus limited.

In contrast, residue-based interaction schemes use a finite alphabet. Interaction energies are determined by the residue types. This means that the energies of different contact pairs in a model protein with residue-based schemes can be correlated. Residue-based interaction schemes are useful for addressing evolutionary issues because they provide heteropolymer-based sequence-structure maps.

Among residue-based interaction schemes, aside from the two- and few-letter constructs described above, models with 20 residue types can sometimes be useful for studying evolution. One apparent advantage of these models is that they have the same number of letters as the natural protein alphabet. However, because of the tremendous computational cost of exploring their huge sequence spaces, in evolutionary applications of these models, conformational exploration is sometimes restricted to maximally compact structures so as to make these models tractable. For example, for chains of 27 units configured on simple cubic lattices, restricting conformations only to those that can be configured within a $3 \times 3 \times 3$ cube drastically reduces shape space (number of conformations unrelated by inversion or rotations) from $\approx 10^{16}$ for unrestricted self-avoiding 27mers to only $\approx 10^{5}$ maximally compact 27mer conformations (see above). But the artifacts from adopting such an approach

can be problematic because often it fails to identify the true optimal structure for a given sequence among all its physically accessible conformations. This is evident from the fact that neither the optimal solutions determined by rigorous lattice computations (Yue et al 1995; Backofen et al 1999; Irbäck and Troein 2002) nor the native structures of real proteins (Goodsell and Olson 1993) are always maximally compact. Indeed, exact enumerations on two-dimensional square lattices show that restricting searches to maximal compact conformations effectively changes the energy function of the model (Chan and Dill 1996). Thus, extra caution should be used to interpret results from such studies.

Interaction schemes more complex than lattice-based additive pairwise contact interactions have also been used to address evolutionary questions. For instance, Ebeling and Nadler introduced an 8-letter two-dimensional model with cooperatively folding units (mimicking either helices or sheets) (Ebeling and Nadler 1995, 1997). In a separate effort, empirical fold recognition programs have been used to investigate neutral evolutionary paths (Babajide et al 1997, 2001).

Comparisons with RNA models

Comparisons between protein and RNA evolutionary behaviours are relevant for understanding primordial evolution, in particular for ascertaining whether there are fundamental physical differences influencing the evolution pathways of these two distinct classes of biomolecules. Apparently, the RNA sequence-to-structure mapping is more amenable to current theoretical analyses than that of proteins. The development of RNA secondary structure models that can efficiently and reliably predict thermodynamically stable structures from their sequences (Nussinov and Jacobson 1980; Zuker and Stiegler 1981) is particularly facilitative to RNA evolutionary studies. Details of these methods can be found in recent reviews (Schuster et al 1997; Higgs 2000).

Robustness of predictions

One of the major unresolved questions in this research area is the extent to which SEM model observations reflect behaviours of real proteins. In other words, how much of the results are merely modelling artefacts? For instance, the prevalence of multiple-conformation ground states might have arisen from using small alphabets. Would this phenomenon, or modified forms of it, persist for larger alphabets with more realistic representations of chain interaction and geometry? Clearly, the simplifications involved in SEM approaches to evolution are drastic. Nonetheless, there is evidence from RNA studies (Tacker et al 1996), lattice considerations (Bornberg-Bauer and Chan 1999) and approximation algorithms (Renner and Bornberg-Bauer 1997) that general SEM conclusions regarding sequence-to-structure maps can be informative about real biomolecules and rather robust.

In the remaining parts of this review, we will focus primarily on two representative SEM modelling approaches: (1) the 2-letter HP model with full conformational enumeration, and (2) lattice models with larger alphabets but with conformational enumerations restricted to maximally compact conformations (note the caveat above). The main reason for our choice of topics is that a large number of protein evolutionary questions have been addressed using these constructs. Here their results are used as illustrative examples. Their relationship with experiments as well as other modelling approaches will be discussed.

Protein evolution

Proteins, as we know them today, are the results of eons of evolutionary optimisation processes. Phylogenetic comparisons allow us to 'look back' for a certain length of time to decipher optimisation at the level of single-residue changes and module recombinations. However, if one wants to understand at a more fundamental level the evolution of proteins from their prebiotic origins, we cannot rely on the analysis of today's proteins alone. This is because modern biopolymers have already evolved for a sustained period of time within a biological setup that they themselves helped to create. Prebiotic evolution, however, must have significantly shaped proteins before the advent of cellular organisms that have amazingly complex machineries such as ribosomes. It would be difficult to shed light on – let alone reconstruct – the history of prebiotic evolution by focusing exclusively on modern proteins.

Assuming that RNA preceded proteins, the recent elucidation of the inner workings of the ribosome suggest that most ancestral proteins might have been quite different from their modern descendants. In particular, the functions of ancestral proteins could have been only supplementary. It follows that proteins could have evolved from biopolymers with less specific structures and functions, and had only a supportive role (Anantharaman et al 2002), quite different from their present role as highly specific 'major players' in a cellular context. Therefore, it is not at all clear if, for example, the obvious preference of modern proteins

for certain common fold families is a result of evolutionary pressure or if it is due to fundamental physical laws governing the distribution of sequences over folded structures. Seeking understanding of such issues is not just of academic interest since it would eventually enable us to ascertain the limitations of rational and combinatorial protein design methods (see Yomo et al 1999; Saven 2001), as well as when standard bioinformatics techniques such as functional (or structural) inference by sequence comparison are likely to fall short. In other words, theoretical and computational efforts including SEM approaches are furthering our understanding at a basic level. They are capable of suggesting answers to questions that can't be addressed by 'traditional' bioinformatics.

Mechanisms of protein evolution

The major mechanisms of biomolecular variation are point mutations (substitutions), insertions, deletions, duplications and recombinations. The consensus view has been that duplication events allow one of two offspring to spread in both genotype and phenotype space. However, often the function of the protein has to be maintained for the organism to survive. Therefore, evolutionary selection/variation cycles often keep a significant population of the protein arrested in phenotype space, but neutral mutations can still cause a drift in sequence space. Only relatively recently has the importance of neutral evolution for reaching otherwise inaccessible regions been recognised and underpinned by theoretical considerations (Maynard-Smith 1970; Kimura 1983).

More recently, in vitro evolution under artificial selection pressures has emerged as a powerful laboratory technique in protein engineering. The theory of in vitro evolution and its implications can be found in the comprehensive review by Voigt et al (2001a).

Protein structures and their determination

At least for relatively small globular proteins, the shapes of their folded structure are encoded by their specific sequences of amino acids. Proper folding is essential for a protein's function. Conversely, mis-folding can lead to disorders such as Alzheimer's, prion and other amyloid diseases (Harrison et al 1999, 2001; Chan et al 2002). A protein's folded structure often provides extremely important information regarding its function. Traditionally, folded structures of proteins are determined by X-ray crystallography and NMR spectroscopy. Recently, new structure determination initiatives are being undertaken to keep up with the unprecedented accumulation of genome sequence information (see review by Šali and Kuriyan 1999). To better understand protein function, new bioinformatics techniques (often characterised as part of 'functional genomics') are also being developed to predict protein-protein interactions, and glean the relevant information from the growing genome databases (Vajda et al 2002). At the same time, mass spectrometry emerges as a promising experimental tool of proteomics for genome-wide analyses of biological processes including protein interactions and biochemical reaction pathways. All of these techniques are relatively demanding. Thus, large scale measurements of mutational effects remain very costly and would not be undertaken in the absence of viable hypotheses about the effectiveness of specific mutations. Theoretical and experimental advances are beginning to address the link between SEM and experimental results (Saven and Wolynes 1997; Kono and Saven 2001; Voigt et al 2001b; Koehl and Levitt 2002). We are hopeful that more extensive experimental evaluations of SEM predictions may soon be possible.

Theoretical models of evolution

Fitness/mortality landscapes

In a seminal paper in 1932 (Wright 1932), Sewall Wright coined the term fitness landscape to describe the dynamics of evolutionary optimisation. In this picture, evolution is viewed as an adaptive (uphill) walk of populations in a multidimensional genotype space towards higher fitness, whereby fitter offspring are selected according to phenotypic criteria. Building on these ideas, several formalisms were developed to describe evolving populations over genotypic landscapes (Kauffman and Levin 1987; Macken and Perelson 1989; Derrida and Peliti 1991; Bak et al 1992; Schuster and Stadler 1994; Stadler 1999). Concepts similar to that applied to graph theory, physics and physical chemistry problems such as spin glasses and potential energy surfaces have been developed to characterise fitness landscapes. Of particular mathematical interest is the average hardness of the optimisation problem with a given model phenotypic fitness criterion and a definition of distance measures on both genotype and phenotype spaces (Kauffman and Levin 1987; Schuster and Stadler 1994; Stadler 1995). Some of these models can be treated analytically if the assignment of fitness to genotypes (with or without explicit treatment of phenotypes) is assumed to take relatively simple mathematical forms.

Locally optimised sequences are depicted as peaks on the fitness landscape. Alternately, to underscore the

mutational stability of these sequences and their role as local evolutionary attractive basins (rather than highly unstable pinnacles of achievement), they have been identified as minima on 'negative-fitness', 'inverse-fitness' or 'mortality' landscapes (Bornberg-Bauer and Chan 1999; Cui et al 2002). The latter representation conforms better to prevailing practices in physics and physical chemistry. It also avoids the juxtapositioning of somewhat contradictory geographical imageries in evolutionary discourses such as referring to a valley on the fitness landscape as a 'barrier' (van Nimwegen and Crutchfield 2000; Voigt et al 2001a).

Fitness criteria

Evolutionary modelling requires assuming a certain connection between model and real sequence and conformational spaces. Generally, such a correspondence between model sequences with real genotypes is much more straightforward than that between model and real phenotypes and their fitness measures. As far as protein structure is concerned, the basic assumption in most models is that one is dealing with single-domain proteins of short to moderate length. Thus, they may represent primordial entities that are directly read (ie without splicing) from the nucleotide carrier of genetic information.

How to define and compute fitness is a general question in biophysical models with explicit chain representations. In the broadest consideration, biological selection takes place under multiple constraints, many of which are tightly interwoven, contradicting or synergistic. Examples abound: often many genes contribute to one feature (known as a multiple trait); sometimes a gene's contribution to fitness depends on other genes (epistasis of fitness); or in some instances a gene affects more than one trait (pleiotropy).

Among SEM studies of molecular evolution conducted thus far, the general approach has been to apply only one selection criterion, which may vary from study to study. The most widely used among the criteria are based on either (1) native structure, (2) foldability or (3) native thermodynamic stability. More complex fitness criteria that might, in a sense, be more relevant to biological function would require more complicated setups, such as taking into account multiple chain interactions. Consideration of such criteria would likely lead to big increases in the number of model parameters and computational intensity. Thus, at present they are not commonly tackled by SEM approaches.

Various structural criteria are often used in computational and experimental optimisation of proteins. But structure by itself can hardly be an appropriate selection criterion in broad-brush evolutionary considerations because it is difficult to assign a scalar fitness value directly to a given structure over a broad range of different putative native structures. However, one can compare structures and assign a scalar value for their similarity, such as assigning a maximum score when the backbone coordinates of two lattice proteins are identical. In bioinformatics, a correlation between structural similarity with functional similarity is a common working assumption in comparative functional analyses. Here, 'structure' or more specifically the ability of a sequence to adopt a unique ground-state structure will be used as an implicit functional selection criterion below, although it should be noted that several experiments have suggested that structure is necessary but not always a sufficient criterion for functional selection (see Motivation section). Measures of structural similarities have been developed for lattice proteins (Renner and Bornberg-Bauer 1997) as well as RNA (Shapiro 1988; Fontana et al 1989, 1993b).

Clearly, most modern proteins have to be *kinetically* foldable in order to function. Therefore, it is conceivable that foldability might have been a stringent selection criterion in pre-biotic evolution when many of the polypeptides in the primordial soup were not foldable, with many proteins perhaps having only flexible, not very well-defined 'ground-state' structures (see above). However, it is probably the case that once a protein is capable of folding at a reasonably fast rate such that it can function under physiological conditions, there is no compelling reason for it to evolve to fold even faster (see Larson et al 2002 and references therein). Indeed, this view is supported by several experiments showing that for some proteins it is relatively easy to construct mutants that fold more rapidly than wild types (Munson et al 1997; Viguera et al 1997; Kim et al 1998; Martinez et al 1999). This scenario has also been conjectured by Goldstein and coworkers based on their SEM studies (Govindarajan and Goldstein 1997a) (see also Protein evolution section).

Protein native thermodynamic stability is governed by the Boltzmann weight of the folded versus that of the unfolded states. In model studies, native stability is determined by exact conformational enumeration or by extensive conformational sampling. For instance, the free energy of folding ΔG has been computed from the full density of states in evolutionary studies using two-dimensional lattice models with HP and MJ-like contact potentials (Govindarajan and Goldstein 1998; Bornberg-Bauer and Chan 1999; Williams et al 2001). Besides methods for accurate determination of ΔG, approximate

measures such as the *Z*-score (Bowie et al 1991) and the closely related average energy gap between the ground-state and a random ensemble of unfolded conformations have also been used to characterise native stability.

While some proteins such as eye crystalline appear to have been optimised for stability, the general observation is that wild-type proteins are only marginally stable. Most single-point mutants are slightly destabilising or neutral but some can be stabilising. The experimental observation that proteins are thermodynamically marginally stable suggests that while the stabilities of functional proteins might be locally close to being optimal, ie more stable than most of their single-point mutants (Wang et al 2002), stabilities are not evolutionarily maximised for the given native conformations. In fact, the conformational flexibility that results from marginal stability is often crucial for a protein's biological function. From an evolutionary standpoint, marginal thermodynamic stability may sometimes also be beneficial because it implies that a given functional sequence is relatively close to sequences that encode for other phenotypes. Hence, marginal thermodynamic stability also entails evolutionary flexibility. More detailed computer experiments addressing these questions are described below.

Based on modelling assumptions, it has been argued in some studies that certain equilibrium (non-kinetic) native stability related measures can be used to predict kinetic foldability. An early example of such proposed measures is the energy gap between the ground state and first excited state among the maximally compact conformations in a class of lattice models (Abkevich et al 1994a; Šali et al 1994; Chan 1995). Ebeling and Nadler have noted, however, that folding sequences in a different model do not necessarily have a wide gap between the lowest and second lowest possible energies (Ebeling and Nadler 1995, 1997).

Extensive evolutionary studies have been conducted by Govindarajan and Goldstein (1997a, 1997b) using a REM-like model with chains restricted to a $3 \times 3 \times 3$ simple cubic or 4×4, 5×5 and 6×6 square lattices. Direct kinetic simulations were not used in their investigations to ascertain the actual folding rates. A 'sequence' (ie a set of interactions, see below) is (thermodynamically) defined as foldable if its 'foldability' $\mathcal{F} > \mathcal{F}_{crit}$, where $\mathcal{F} = \Delta/\Gamma$, with Δ being the difference between the energy of the native structure and the average energy of the ensemble of random conformations, Γ being the standard deviation of the distribution of energy values among the random structures,[1] and $\mathcal{F}_{crit}$ being a tunable 'critical' value for $\mathcal{F}$ (Govindarajan and Goldstein 1996). This earlier model is contact-based rather than residue-based. Thus, it does not have an explicit sequence space (see above). Instead, genotypes are represented by normalised vectors in a multidimensional interaction space, whereby mutational effects are characterised by the angles between interaction vectors. More recently, the Goldstein group introduced a residue-based model with a slightly modified MJ potential. Using this construct, they have conducted explicit simulations of adaptive walks on the fitness landscape, modelled the properties of duplicating genes (Taverna and Goldstein 2000a), as well as the influence of a range of selection criteria such as ligand binding and native compactness (Taverna and Goldstein 2000a; Williams et al 2001).

Several SEMs have been adapted to tackle function in terms of ligand binding by using refined HP models with appropriate considerations of potentials and shapes (Hirst 1999; Blackburne and Hirst 2001) or by computing the energy of model protein-ligand complexes (Williams et al 2001). The inferences regarding fitness landscape provided by these innovative applications are very similar to those deduced from simple structure-based SEM considerations. For example, structural designability emerges as a feature in dynamic optimisation of populations (Williams et al 2001), and the extended mutational networks observed (Hirst 1999) are analogous to that from more simple SEM considerations.

One modelling constraint that may also be viewed as an implicit assumption on fitness is the algorithm of Shakhnovich and coworkers that keeps the amino acid composition constant during a Monte Carlo type sequence optimisation procedure (Abkevich et al 1996). There is no obvious biological counterpart to this mechanism, though prevention of protein aggregation was offered as a possible justification (Abkevich et al 1996). Because of space limitations, we do not consider this somewhat artificial mutational constraint in the discussion below.

Computations of biopolymer evolution

Various models aiming to elucidate the structure or the folding behaviour of biopolymers, such as the RNA secondary structure model and lattice proteins were already in place since the 1980s. A survey of average properties of the two-dimensional HP model sequence-to-structure map was provided by Lau and Dill (1990). In a seminal paper that appeared soon after, Lipman and Wilbur (1991) took up a simple but strikingly convincing argument by Maynard-Smith (1970) that functional routes of molecular

evolution (corresponding to networks and/or paths, see below) must (1) exist and (2) be taken if the hosting organism of a biopolymer (and thus the biopolymer itself) is not to die out. To address this question, Lipman and Wilbur exhaustively enumerated sequences in an HP model to determine their ground states and found evidence that such networks exist. (A functional sequence in their study is defined as one that has a unique lowest-energy contact map within a restricted set of conformations.) Aspects of neutral molecular evolution were also elucidated using the same model (Lipman and Wilbur 1991). These results impacted studies on RNA (Schuster et al 1994; Huynen et al 1996; Fontana and Schuster 1998a, 1998b), which in turn triggered further SEM studies on protein evolution.

The principal merit of this investigative methodology is a marriage between a theoretical framework on evolution and a tractable chain-based biophysical model of sequence-structure mapping, which enables an exact description of mutations and evolving populations in the model. As a result, new concepts and new terminology such as *neutral set, neutral net, shape space covering, superfunnels* have emerged to characterise novel features observed in these models. In particular, two genotypes (sequences) are neutral when they have the same phenotype (defined by structure or foldability or other criteria). They are (sequence-space) neighbours when they differ by only one edit operation, for example a single point mutation or a recombination event. A Hamming distance between two sequences is defined as the minimum number of point mutations that must be applied to convert one sequence into the other. In many applications, sequence space is identified with the set of all possible sequences of a given length, where the Hamming distance is an easily computable metric in that space. Single paths of neutral neighbours are termed neutral paths and a collection of all such paths that are interconnected is a neutral net. The set of all neutral mutants, whether connected or not, is termed a neutral set.

Addressing evolutionary questions using SEMs

Conceptually, it is useful to distinguish three major types of computations which are used to elucidate different aspects of the evolutionary behaviour of SEM populations:

- 'Static average properties' that characterise the sequence-to-structure map in terms of average properties.
  - A useful measure is *convergence*, which describes how many sequences are mapped onto a limited set of phenotypes. Naturally, convergence depends on the notion of phenotype. Among all possible structures, there is generally only a small fraction that are mapped onto at least one sequence that has the given structure as its unique ground state. Such a structure is termed *encodable* (Chan and Dill 1996) or *designable* (Li et al 1996). Typically, there are many more sequences than encodable structures. Hence most neutral sets have more than one sequence. In this regard, of considerable interest is the distribution of encodablity. For instance, whether there are only a few highly encodable structures (with large neutral sets) and many rare structures (with small neutral sets), or instead all compact structures have essentially the same encodability?
  - Average landscape properties (or complex combinatory maps (Fontana et al 1993a, 1993b)) such as the dependence of average structural similarity on mutational distance from a given reference genotype. In particular, the *correlation length* of a sequence-space property may be viewed as an approximate sequence-space distance between two local optima of the given property.
- 'Static topological properties' characterise neutral nets, their topology and their arrangement in sequence space. The distribution of, and separation between nets are of importance for two main reasons: (1) For understanding how biopolymers could have evolved from a possibly small set of ancestral proteins. In other words, whether there has been a structural 'lock-in' such that all of today's proteins were descended from a relatively small set of proteins which had independently arisen. (2) For understanding how difficult it is to convert a protein encoding for one phenotype into one that encodes for another phenotype.
  - Typical extent of neutral nets: whether a large fraction of sequence positions can be changed while keeping the folded structure unchanged.
  - Sequence plasticity means resistance of structure against mutations (Lau and Dill 1990; Govindarajan and Goldstein 1996, 1997b; Bornberg-Bauer and Chan 1999). A protein sequence is plastic if many other sequences, especially its sequence-space neighbours, are phenotypically neutral in that they encode for the same structure as itself. In other words, a plastic protein sequence is mutationally stable. Of interest is whether differences in sequence plasticity can be readily predicted from sequence or structural features.

  - On the other hand, phenotypic plasticity refers to the multiplicity of structures a given sequence can fold into (Ancel and Fontana 2000). In other words, a sequence with phenotypic plasticity does not encode uniquely.
  - Shape space covering refers to the property that, given a genotype with an associated structure, genotypes that map onto almost all other structures can be found within a relatively small mutational distance. In other words, when shape space covering holds, if one 'walks in the right direction' in sequence space for a little while, one will find mutants mapping onto almost all phenotypes.
  - Transition points and switches in the sequence-space regions between neutral nets are important for the evolution from one phenotype into another.
- 'Explicit dynamics,' which refers to the stirred-flow reactor based simulation of replication–selection cycles.
- The robustness of all these results with respect to changes in energy parameters, alphabet size, mutation type, fitness criteria etc is of critical importance in ascertaining which model predictions are likely to be realistic and generalisable.

As we have argued above, to study broad-brush aspects of evolution, it is more effective to study large but simple systems for general principles than to study a few selected cases using high-resolution structural representations for the proteins. The rationale is that if several properties of a given model are consistent with experiments, it is not unlikely that other predictions of the model would also be valid and therefore are worthy of new experimental designs for their verification or falsification. To facilitate this effort, the following are brief descriptions of SEM-computed evolutionary properties and their relevant experimental results, given in roughly the same order as the list above.

Average properties

Landscapes

Both approximate treatments and SEMs have been used to characterise evolutionary landscapes. RNA models have been compared to spin glasses and tunable *NK* models. Apparently, the RNA alphabet and its associated interactions might have been well tuned to yield a fairly smooth and correlated landscape (Fontana et al 1993a). For proteins, computations using HP-like models and an approximate folding algorithm (Renner and Bornberg-Bauer 1997) suggest that a reasonable amount of neutrality and correlation exist in their evolutionary landscapes. Significantly, the likelihood of a model protein sequence to encode uniquely increases with alphabet size. This may explain why proteins have an alphabet larger than 5–8 amino acid types, even though such a reduced alphabet could have been readily encoded by nucleotide duplets instead of triplets.

The utility of information provided by average landscape properties is often limited by the fact that sequence space is not isotropic with respect to structural and evolutionary properties. In other words, sequence neighbourhoods of different sequences can be very different. This may best be illustrated by the RNA case which has been shown to have sparse but extended neutral networks (Schuster et al 1994). For example, a very small set of continuous mutations can maintain a given structure while at every point of that evolutionary path the vast majority of mutations would yield a sequence coding for a different structure. This phenomenon allows evolving populations to drift along these networks and has been described as exhibiting 'smoothness within ruggedness' (Huynen et al 1996). The terms 'neutral network' and 'neutral net' are synonymous in the present discussion.

Convergence and distribution of convergence

It is well known that most proteins belong to a relatively small set of structures (known as 'folds'), with little variations within a given fold family, mostly only in loop regions (Chothia 1992; Orengo et al 1994). Theoretical considerations (see Wang et al 1996; Zhang 1997; Govindarajan and Goldstein 1998 and references therein) suggest that the maximum number of possible protein folds of reasonable sizes is between 5000 and 8000, while the number of probably realised ones is around 1000 to 2000, and approximately 500 folds are sufficient to cover virtually all protein structures known to date.

Using SEMs, convergence has been reported for the HP model first by Lau and Dill (1990), Chan and Dill (1991), and Lipman and Wilbur (1991), with slightly different definitions for a sequence to be considered having a unique fold. As in earlier work on RNA (Schuster et al 1994), distribution of convergence has subsequently been shown to follow approximately a power-law behaviour known as Zipf's law (Li et al 1996; Bornberg-Bauer 1997b). In essence, this means that there are a relatively small number of 'dominating' folded structures that are encoded uniquely by many sequences whereas many other folded structures are 'rare' in that they are encoded uniquely by few sequences.

Using the two-dimensional HP model, Chan and Dill (1991, 1996) found that on average convergence increases with the compactness of the putative native conformation. The three-dimensional study of Li et al (1996) using a modified two-letter HP potential also indicates a highly uneven distribution of convergence among structures, notwithstanding that searches were restricted only to maximally compact two- and three-dimensional conformations in their approach. Subsequently, in a comparative study of residue-based models by Buchler and Goldstein (1999) that also restricted conformational searches to maximally compact structures (5×5 square-lattice conformations in this case), distribution and ranking of convergence were found to be dependent on the fitness criterion as well as alphabet size. In their approach, structures that are highly designable for the two-letter alphabet are not particularly designable with larger alphabets. Thus, the robustness of more detailed aspects of SEM predictions against variations in alphabet size and interaction schemes remains to be further elucidated (see also Shahrezaei and Ejtehadi 2000 and references therein). From a polymer physics standpoint, however, it is obvious that the modelling procedure of restricting to maximally compact conformations tends to increase convergence over the corresponding model that does not impose such a restriction, and it often changes the ground-state conformation(s) of a given sequence (Chan and Dill 1996). A case in point is the recent extensive study of the two-dimensional HP model with chain length $n = 25$ by Irbäck and Troein (2002). They find that if only the 5×5 maximally compact conformations were considered, 6 181 800 HP sequence would be identified as unique sequences. But full conformational searches indicate that there are only 765 147 truly unique sequences, and among them only an insignificant number of 605 have maximally compact native states. This implies that 99.99% of the sequences determined to be unique by the maximally-compact-only method are in fact *not* unique.

It has been apparent since the early study of short HP sequences ($n = 13$) by Lau and Dill (1990) that mutations in the protein core (typically an H to P) have a strongly destabilising effect, whereas surface mutations (especially P to H) are mostly neutral. These model observations are consistent with experiments on real proteins (Cordes and Sauer 1999). Most mutations that happen to take one unique sequence to another unique sequences are neutral, but in general mutations on unique HP sequences have a high probability of producing sequences with gemisch ground states (Chan and Dill 1994). A tiny fraction (~ 2%) of double mutations in the HP model studied by Lau and Dill (1990) was 'revertants' in that the two mutations were structurally compensating. Naturally, a folded conformation (shape) with a larger set of converging sequences is more tolerant towards sequence variations. This implies that more frequent structures (those with large convergence sets) are more isolated in shape space, as stipulated by the 'designing out' hypothesis (Yue and Dill 1992; Li et al 1998; Bornberg-Bauer 2002). This hypothesis is based on the simple recognition that a good folding sequence not only has to design in its putative native structure, but also has to design out competing compact structures. Similar trends have also been observed in the contact-based (non-residue-based) modelling of Govindarajan and Goldstein (1995, 1997a), in which structures with a contact map very dissimilar to others tend to have larger convergence sets. Convergence sets of more dissimilar structures are also more separated in their interaction space. Similar trends have also been observed in simple off-lattice model considerations (Nelson and Onuchic 1998).

Closer examinations of the HP model indicates that the overwhelming variation within a neutral set is due to surface mutations. Core hydrophobic residues tend to be conserved because they are critical for native stability (Bornberg-Bauer 2002; Chan et al 2002). These SEM predictions are consistent with studies using empirical potentials and continuum atomic representations of proteins (where fitness is measured by the *Z*-score of a target structure) showing that hydrophobic residues have a high tendency to be conserved along neutral paths (Babajide et al 1997). In interpreting these and related model results, it is important to recognise that the distribution of conserved residues is sensitive to the model interaction scheme. This is highlighted by the evolutionary landscape of a non-proteinlike two-letter (like-attracts-like) 'AB' model that is much more rugged than that of the HP model (Bornberg-Bauer and Chan 1999).

In the contact-based treatment of Govindarajan and Goldstein (1997a), it is found that positions in the interaction space that map onto similar structures and structures with similar optimal foldability tend to cluster together. On the other hand, using a 20-letter 36mer model, Bastolla et al (1999) found that the degree of mutability was quite uniform throughout their model protein except two core positions (residue numbers 16 and 27) which were more conserved. However, it is problematic to relate this particular study to real proteins because the protein core in their model appears to be stabilised not by hydrophobic (as in real proteins) but

by (nominally) charged residues (Chan et al 2002). In all three 20-letter sequences designed by Abkevich et al (1994b) for the target structure used by Bastolla et al, the core positions 16 and 27 are invariably stabilised by an ion pair: (aspartic acid, lysine), (arginine, aspartic acid), and (glutamic acid, arginine).

Topological aspects of neutral evolution

Topology of neutral nets

Earlier investigations on RNA showed sparse but extended neutral nets (Schuster et al 1994). That is, for any given sequences only a small fraction of mutations are neutral, and for a small fraction of sequences there are none. However, it is possible to traverse (percolate) the extent of entire sequence space by steps of neutral point mutations (or pair-mutations), with a significant probability of ending in an RNA sequence that has not a single common base at any given position with the starting sequence. Using expectation values for neutrality, a model based on random graphs have been constructed to describe this and the cognate shape space covering phenomena in RNA sequence space (Reidys et al 1997). Natural proteins, on the other hand, are known to allow many neutral mutations, ie they exhibit sequence plasticity, but show little phenotypic (structural) plasticity.

In their model simulations, Ebeling and Nadler (1997) found sequence-space patches encoding for the same target structure. Folding sequences are distributed approximately uniformly within a patch, with a relatively small average pairwise Hamming distance between them: they are thus homologous.[2] For a given target structure, they also observed nonhomologous converging sequences belonging to different patches.

HP neutral nets appear to be rather dense with a fairly high average but uneven distribution of neutral neighbours (Figure 2a). Among $n = 18$ sequences, the Hamming distance between two sequences that belong to the same neutral net can be as large as 7. In other words, for these 'homologous' sequences, $7/18 = 38.9\%$ of their positions are distinct (Bornberg-Bauer 1997b). For real proteins, however, owing to the vastly larger sizes of 20-letter sequence spaces, neutral nets are expected to consist of sparsely occupied regions of sequence space (Eigen 1987). Nonetheless, in real sequence database searches, it has been shown in several separate occasions that intermediate sequences can be found to bridge 'gaps' and thus connect family members that would

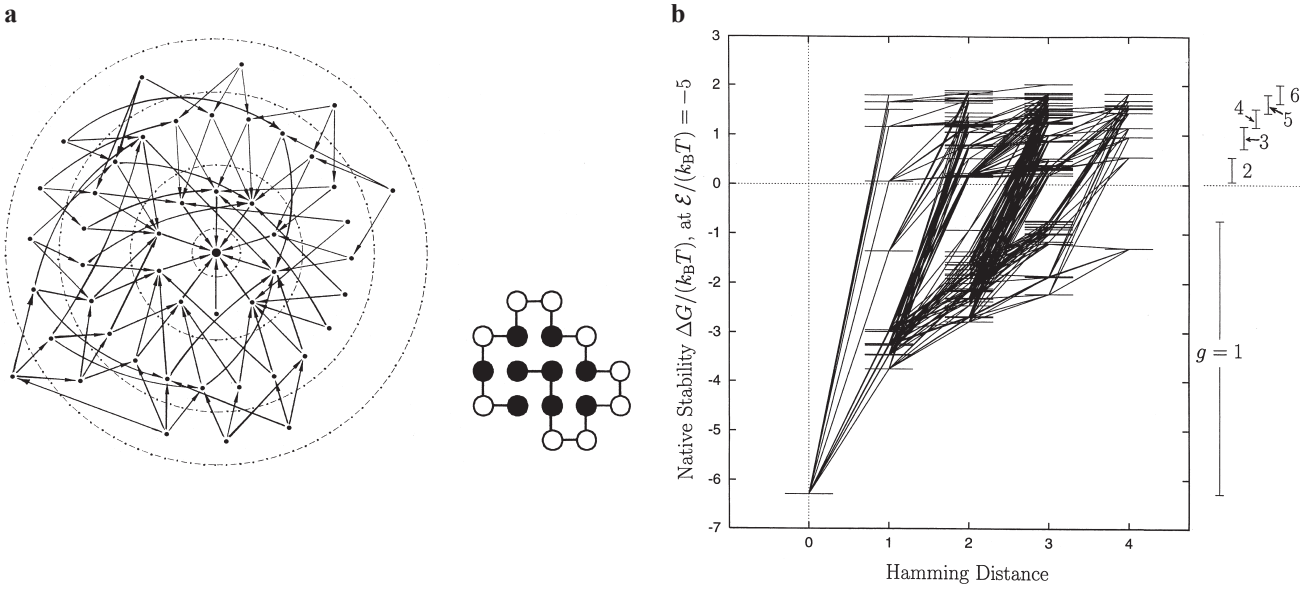

Figure 2 The superfunnel paradigm. **(a)** The largest neutral net in the HP model of chain length $n = 18$, showing 48 sequences converging onto the conformation shown. The sequences are given in Figure 16.7 of Chan et al (2002). Here each sequence is symbolised by a dot. An arrow denotes a single-point mutation, pointing towards the sequence whose native structure is thermodynamically more stable. The sequence in the centre is the prototype sequence with the largest number (ten) of neutral neighbours and whose native structure is thermodynamically most stable. It is identical to the consensus sequence of the neutral net. Hamming distances from the prototype sequence are marked by the concentric circles. **(b)** An extended neutral net including the 48 unique sequences ($g = 1$, g is ground-state degeneracy) as well as non-unique HP sequences that have the given structure (in (a)) as one of its multiple ($g > 1$) ground-state conformations. For each sequence at a given Hamming distance from the prototype sequence (horizontal scale), thermodynamic stability of the target structure (vertical position of the horizontal bar) is the free energy of folding ΔG in units of k_BT (Boltzmann constant times temperature) calculated from the sequence's density of states. Horizontal bars for a pair of sequences separated by one single-point mutation are connected by a line. Unique sequences have $\Delta G < 0$, whereas non-unique sequences have $\Delta G > 0$. Source: Bornberg-Bauer E, Chan HS. 1999. Modeling evolutionary landscapes: mutational stability, topology and superfunnels in sequence space. *Proc Natl Acad Sci USA*, 96:10689–94. Copyright 1999. By permission of the National Academy of Science, USA.

otherwise appear unrelated (Neuwald et al 1997; Park et al 1997; Krause and Vingron 1998). This observation lends credence to the picture of net-like arrangements of homologous sequences in real protein sequence space.

In the HP model, because only two letters are considered, sequence variability in neutral nets are in general lower than that for real proteins. Therefore, it is important to recognise that Hamming distances in the HP model may correspond to much larger Hamming distances for real proteins, because a conserved H or P monomer may mask mutations in the corresponding 20-letter alphabet. For example, underlying a conserved H, there can be mutations among hydrophobic (H) residues (leucine, valine, isoleucine etc) that are not registered in the reduced description of the HP model. Therefore, HP model neutral nets should be viewed as representing real protein neutral nets that are much more extended.

The most striking prediction from HP neutral nets is the emergence of prototype sequences, which have the largest number of neutral neighbours on a given neutral net. Often the prototype sequence is essentially identical to the consensus sequence of the neutral net (Bornberg-Bauer 1997b). Most interestingly, for most neutral nets, there appears to be a funnel-like (term 'superfunnel') topological arrangement of native thermodynamic stabilities (ΔG) of neutral sequences centering around the prototype sequence, whereby the prototype sequence is the one that has the highest native stability (Bornberg-Bauer and Chan 1999) (Figure 2b).

Neutral nets emerged not only in residue-based models of the protein sequence-to-structure map. Using a contact-based approach, Govindarajan and Goldstein (1997a, 1997b) also found that interaction vectors for similar structures tend to cluster together (see above) (Figure 3). In simulations of evolutionary dynamics (see more detailed discussions below), they further found that under increased selective pressure evolutionary paths are increasingly confined to a localised region of interaction space, which may be identified as a neutral network (Figure 4). In general, the part of interaction space that is well-populated is much larger when a less stringent fitness criterion is imposed. Thus, most of the sequence population is to be found in region of relative low fitness. Native thermodynamic stability is often a requirement for functional fitness, thus there is a tendency for sequences to have marginal native stability. A similar perspective is suggested by the HP model

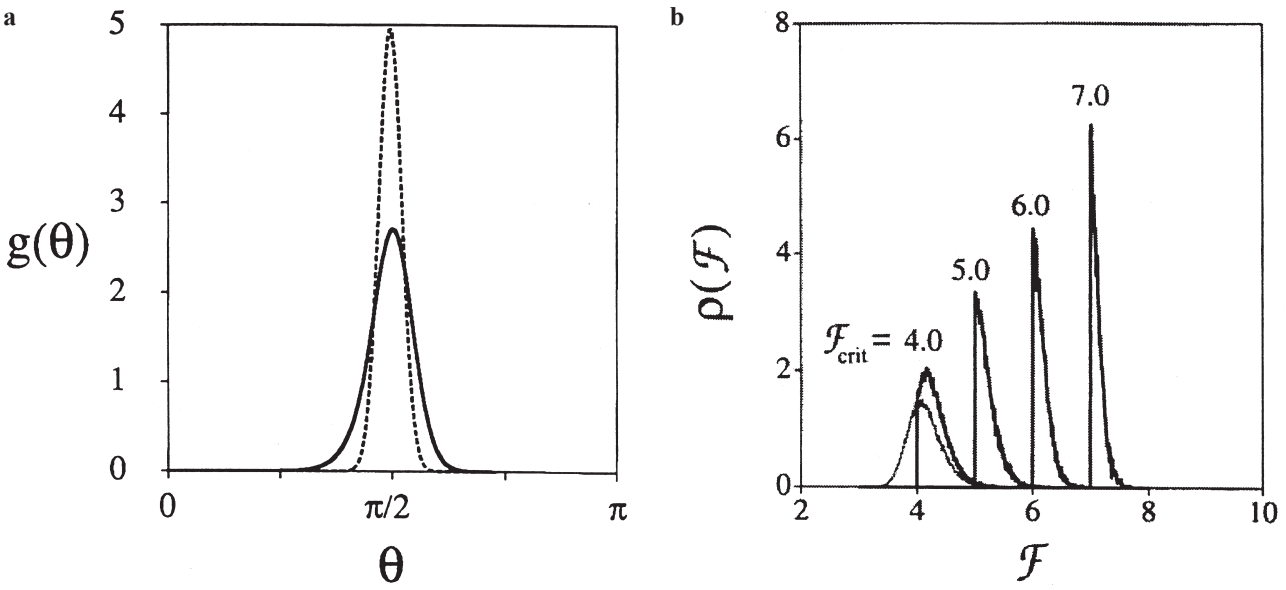

Figure 3 Protein distributions in model sequence spaces. **(a)** Nonrandomness among optimised sequences in the contact-based model of Govindarajan and Goldstein (1997a). Here θ (horizontal axis) is an interaction-space measure, larger θ values correspond to larger sequence dissimilarity. The pair correlation function $g(\theta)$ between optimised sequences – ie each of which is optimal for folding to a different structure — (solid curve) is different from that between random sequences (dotted curve). Optimised sequences that are similar tend to be more clustered than random (spread of $g(\theta)$ to the left). Presumably these sequences are optimised for similar structures. But at the same time, there are more dissimilar optimised sequence pairs than random (spread of $g(\theta)$ to the right). This may be interpreted by the fact that an optimised sequence not only need to design in the target structure but also has to design out competing structures. Thus, its interaction would be quite different from that of sequences optimised for dissimilar structures. **(b)** Distribution of foldability $\mathcal{F}$ in the evolutionary dynamics of the same model. Sequences are simulated under different selective pressure by adopting different $\mathcal{F}_{crit}$ (= 4.0, 5.0, 6.0 or 7.0) as the lower bound on $\mathcal{F}$ for a sequence to be viable. In each case, $\mathcal{F}$ values of most sequences in the evolutionary population are either exactly the same as the input $\mathcal{F}_{crit}$ or only slightly higher, implying that the evolutionary dynamics in this model naturally leads to a population dominated by sequences that are marginally foldable. Included for comparison is the distribution in the absence of selective pressure (leftmost dotted curve). Source: Govindarajan S, Goldstein RA. 1997a. The foldability landscape of model proteins. *Biopolymers*, 42:427–38.

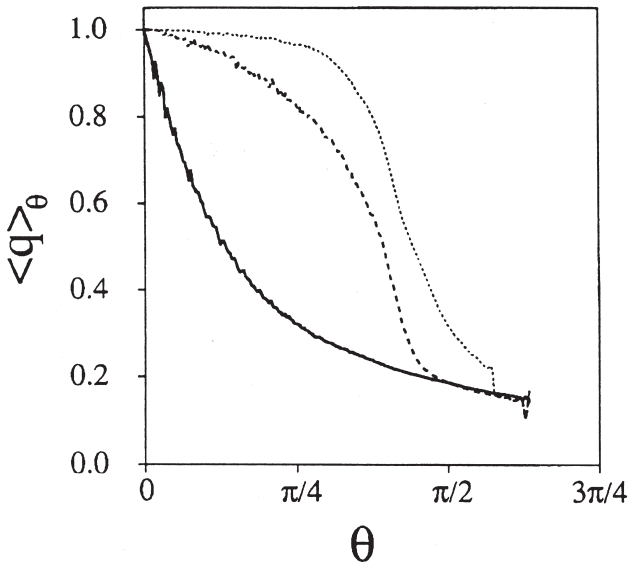

Figure 4 Structural similarity among native conformations, in the protein evolutionary dynamics model of Govindarajan and Goldstein (1997b). Larger θ values (horizontal axis) correspond to larger sequence dissimilarity, as in Figure 3a. Here $\langle q \rangle_\theta$ is an average structural similarity measure for a given θ, $q = 1$ when two structures are identical. When the folding criterion imposed on an evolving population are stringent eg $\mathcal{F}_{crit} = 6.0$ (rightmost, dotted line), native structures of a wide range of sequences are essentially identical (slow decay of $\langle q \rangle$ as θ increases), implying that the evolutionary walk among these sequences consists of extended neutral paths. On the other hand, if the folding criterion is more tolerant eg $\mathcal{F}_{crit} = 3.0$ (leftmost, solid line), even small changes in sequence would result in quite dissimilar native structures (rapid decay of $\langle q \rangle$). The middle dashed line is for an intermediate folding criterion $\mathcal{F}_{crit} = 5.0$. Source: Govindarajan S, Goldstein RA. 1997b. Evolution of model proteins on a foldability landscape. *Proteins Struct Funct Genet*, 29:461–6. Copyright 1997. By permission of Wiley and Sons Inc.

superfunnel picture (Bornberg-Bauer and Chan 1999). As pointed out above, highly designable structures are better separated in sequence space from other structures. This suggests that once protein sequences enter the sequence-space region of a highly designable structure, it is not easy for them to escape by point mutations. It is therefore reasonable to expect that once a structure with a certain threshold fitness appears, it would become fixed and the part of sequence space accessible by it would develop into a sparsely connected network.

Using an inverse folding model with empirical potentials for the full amino acid alphabet and continuum atomic chain representations, Babajide et al conducted searches for neutral paths for several small proteins (Babajide et al 1997). In the context of the same model, they have also investigated several two-, three- and four-letter alphabets with restricted amino acid types. Some restricted alphabets (eg alanine, aspartic acid, leucine and glycine) retain the ability to encode for the target structures but not others (eg alanine, aspartic acid).

In their 20-letter studies, Babajide found neutral networks that span essentially the entire sequence space. However, when the 20 amino acid types in their neutral networks are grouped into two classes of hydrophobic (H) and hydrophilic (P) residues, the 2-letter Hamming distances of the resulting HP patterns for the neutral net of a given protein structure are relatively localised. Therefore, consistent with our discussion above about HP neutral nets, these authors observed that sequences belonging to a neutral set or a neutral net are 'very flexible at the level of individual amino acids but requires a significant level of conservation of amino acid classes' (Babajide et al 1997). Similar conclusions have also been drawn by Hinds and Levitt (1996).

In summary, neutral nets of considerable extent exist in every model just described. Thus, it is reasonable to believe that neutral nets are a very robust feature in protein sequence spaces. Therefore, it is likely that structurally neutral evolutionary changes have contributed considerably to the development of today's proteins, notwithstanding the possibility that they might have arisen originally from 'locked-in' ancestral precursors. The phenomenon of neutral nets also rationalises, at least for larger protein families, the wide diversity of sequences in sequence databases. As will be further discussed below, neutral net topologies help explain the marginal stability of most observed proteins, their robustness against single-site mutations, and the possibility that sometimes the stability and folding speed of wildtype proteins can be increased by mutations.

Connectivities among neutral nets

The prospect of converting one protein structure through a series of mutations into another structure has always generated considerable experimental (Bowie et al 1990; Kamtekar et al 1993; Rose and Creamer 1994; Dalal et al 1997) and theoretical (Bornberg-Bauer 1996; Flamm et al 1999; Newman and Engelhardt 1998; Babajide et al 2001) interest. In a recent HP model study, it was shown that a majority of structures are encoded by neutral nets that are interconnected, 71.7% of model protein sequences for 52.7% of encodable structures form a big network (Cui et al 2002). But the connection between neutral nets is rather sparse (Bornberg-Bauer 1997b; Bornberg-Bauer and Chan 1999). Therefore, although it is possible to pick the correct single-point mutational paths to achieve interconversion of structures (Cui et al 2002), such evolutionary routes are unlikely (ie there are sequence-space 'entropy barriers' to conversion, see van Nimwegen and Crutchfield 2000). A similar argument holds also for SEM evolutionary landscapes with ligand-binding-like fitness criteria (Blackburne and Hirst 2001).

However, if some proteins can survive without being able to fold uniquely, other routes of structural innovation become available. For example, when non-unique HP sequences with up to six iso-energetic ground-state conformations are examined (Bornberg-Bauer and Chan 1999), *neutral paths* of uniquely folding sequences are seen to be embedded in *neutral patches* of sequences that (sometimes) fold to the same structure (as part of its multiple-conformation ground state) but not uniquely (Figure 2b). As for neutral nets, there is a general increase in mutational stability directed toward the prototype sequence of the neutral patch (Bornberg-Bauer 2002). In the context of neutral patches, sequences folding degenerately into two or more encodable structures play the role of switches that can serve as bridges between neutral nets. In some instances, a sequence with two iso-energetic ground-state conformations is exactly one point mutation away from one member each of two different neutral nets, with its degenerate ground-state structures corresponding to the unique ground-state structures of the two neutral nets (Bornberg-Bauer 1996; Trinquier and Sanejouand 1999; Chan et al 2002). For real proteins, some polypeptide fragments have been known to fold to two alternative structures such as α-helices or β-sheets depending on context (Kabsch and Sander 1984; Minor and Kim 1996). Some of these sequences may serve as evolutionary bridges. That such a role is possible for some sequences has been demonstrated experimentally by careful re-designs of HP patterns in real proteins (Murray et al 1995; Cordes et al 1999, 2000). In this connection, it is interesting to note the suggestion that prion molecules might represent evolutionary intermediates of membrane proteins on route to becoming cytoplasmic proteins (Tompa et al 2001). Some HP sequences have been shown to exhibit prion-like behaviour (Harrison et al 1999; Giugliarelli et al 2000), but their relationship with evolutionary switches has yet to be investigated.

As mentioned above, one conspicuous feature of SEM-predicted sequence-space structure is a lack of shape space covering, suggesting that protein and RNA neutral networks are organised significantly differently. For RNA, neutral nets are relatively sparse but so extended that they have transition points that come close (in terms of point mutations) to at least one member of most other neutral nets (Fontana et al 1993b; Fontana and Schuster 1998b). Of course, the fact that shape space covering was not observed in SEMs of protein evolution per se does not preclude it for real proteins. Indeed, in a recent non-SEM study that used empirical knowledge-based potentials and *Z*-scores to mimick the sequence-structure map, Babajide et al (2001) found that highly dissimilar protein structures could be encoded in their model by sequences that were only a few point mutations apart. Nonetheless, in view of the high-dimensionality of real proteins' sequence space and the clustering observed for real protein sequences as well as in several SEM models (see above), it is reasonable to expect that shape space covering does not exist for real proteins (Bornberg-Bauer 1997b). This hypothesis is further buttressed by (1) the fact that it is difficult to meet the Paracelsus challenge experimentally[3] (Rose and Creamer 1994; Dalal et al 1997), and (2) the importance of HP sequence pattern on protein structure (Davidson et al 1995; Cordes et al 1996, 2000; Marshall and Mayo 2001). These observations suggest strongly that the lack of shape space covering predicted by the HP model should apply to real proteins.

From an applied bioinformatics standpoint, an understanding of network separation in protein sequence space has direct implications on the development of database search strategies. Based on HP model results, it has been shown that the reliability of statistical protein folding potentials can be improved by taking neutral set or neutral net information into account (Cui and Wong 2000). In accordance with network separation, in an iterative application of standard algorithms for locating homologues, protein families of homologous sequences have been shown to consistently cluster together, independent of the member sequence used to seed the search. This has provided a method to sort sequences into clusters, a large fraction of which is pairwise disjoint (Krause and Vingron 1998). It follows that search strategies among evolutionary related sequences are more reliable when well chosen representatives (eg in a tree topology) rather than average representations such as profiles are used for iterative queries.

Evolutionary simulations by population dynamics

To better understand evolutionary processes, one has to consider population dynamics in addition to the static evolutionary landscape properties discussed above (Figure 5). Dynamics simulations of SEM evolution were first reported by Fontana and Schuster on RNA using a recursive folding algorithm (Fontana and Schuster 1987; Fontana et al 1989). Their simulation corresponds to a stirred-flow reactor experiment. A constant population of 2000 molecules of length $n = 70$ was simulated to undergo

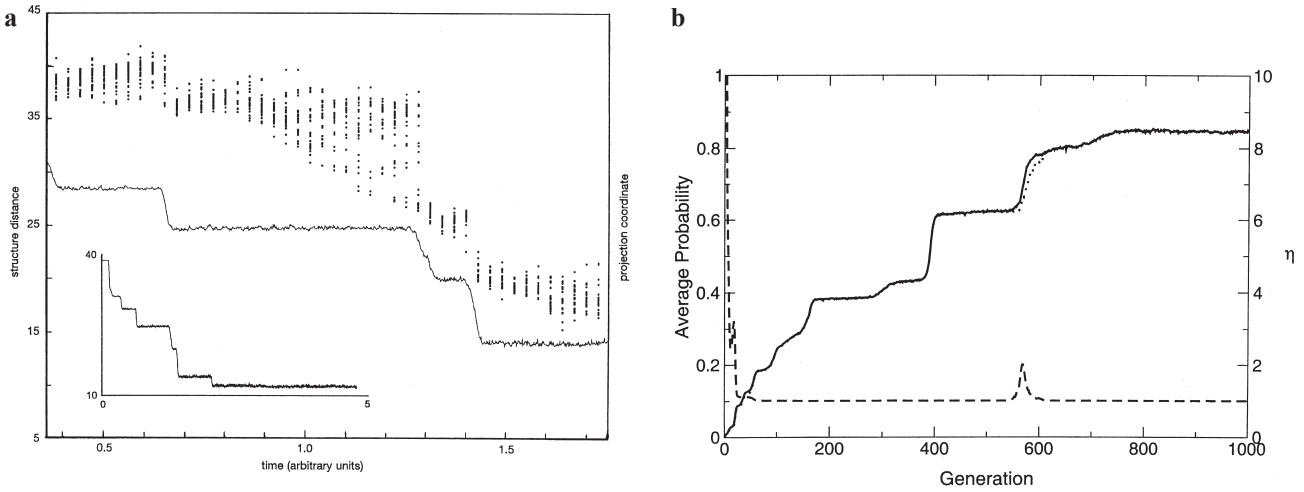

Figure 5 Evolving populations in a virtual stirred flow reactor. **(a)** Populations of RNA molecules undergoing optimisation for a target secondary structure show punctuated equilibrium like behaviour. The inset shows a complete trajectory of the average structure distance (left vertical scale) between the target structure and a collection of 1000 RNA molecules. The full line of the main graph is a zoom-in of part of this trajectory; the corresponding genetic or sequence variation in the evolving population (right vertical scale) is also monitored. Here a wider spread in the vertical column of dots indicates a higher degree of genetic diversity at the given time. The results suggest that genetic diversity increases as the population drifts on neutral nets (plateaus of essentially constant average structure distance). But at each 'jump' to a new neutral net with smaller average structural distance from the target, the genotypes are first confined to small variations around the 'entry point' sequence of the new neutral net. Source: Huynen MA, Stadler PF, Fontana W. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci USA*, 93:397–401. Copyright 1996. By permission of the National Academy of Science, USA. **(b)** Evolution of conformational compactness in a population of 1000 lattice model proteins also exhibits punctuated equilibrium like behaviour. Here conformational compactness is used as a selection criterion. Average quantities over the collection of protein sequences are shown as functions of the number of generations in the evolutionary simulations. The (i) solid trace and the (ii) (nearly identical) dotted trace are, respectively, the average probability (left vertical scale) that (i) a sequence is in a compact conformation and that (ii) a sequence adopts a compact conformation whose energy is most favourable for the given sequence. The dashed curves shows the effective number η of compact states in the population (right vertical scale). In this example, after the initiation of evolutionary dynamics, the population rapidly locks in to a single compact conformation. During the entire simulation, there is only one switch-over between two compact conformations (around generation 575). Source: Williams PD, Pollock DD, Goldstein RA. 2001. Evolution of functionality in lattice proteins. *J Mol Graphics and Modelling*, 19:150–6. Copyright 2001. By permission of Elsevier Science.

error-prone replication cycles. The model molecules' fitness was related to the RNA's stability against hydrolytic degradation. Among the most important findings from these pioneering studies were the existence of error thresholds, quasi-species like clustering of populations around 'master sequence,' and punctuated-equilibrium-like plateaus of evolutionary optimisation. These issues have also been addressed by analytical approaches (van Nimwegen and Crutchfield 2000).

In later SEM investigations, structural similarity was used as a fitness criterion and transitions between fitness plateaus were related to transitions to a new neutral net in sequence space (Huynen et al 1996). Long periods of diffusion on neutral nets without functional improvement are contrasted by sudden innovations (Figure 5a). In a follow-up study that focused on these transitions, Fontana and Schuster (1998a) found that such concerted structural rearrangements in RNA often require preceding neutral drifts. Owing to computational limitations, a comparable level of sophistication is yet to be achieved by SEM simulations of protein evolutionary dynamics.

For the HP model, analytical considerations suggested that network topology can have an effect on the distribution of steady-state evolutionary population among sequences on a neutral net. Assuming that every sequence in a neutral net has the same constant isotropic rate of mutation, *individual* sequences with more neutral neighbours, ie prototype or near-prototype sequences, would tend to be more populated than individual sequences situated at the fringe of the neutral net (Bornberg-Bauer and Chan 1999). However, because the fringe of a neutral net encompasses a larger sequence-space volume (ie has a higher sequence entropy, cf van Nimwegen and Crutchfield 2000), in the absence of additional selective pressure there is a 'centrifugal' effect resulting in more marginally stable sequences populating the fringe of the neutral net *as a whole* (Figure 6). Balance of these effects have been addressed by simulations of evolution dynamics (see below).

A number of simulations of population dynamics have been conducted by Goldstein and coworkers to address an extended range of evolutionary issues (Govindarajan and Goldstein 1997a, 1997b, 1998; Taverna and Goldstein 2000b, 2002a, 2002b; Williams et al 2001) (see sections: Simple exact models and related theoretical constructs; and Protein evolution). For example, to address the thermodynamic hypothesis of protein folding, which stipulates that a protein's native structure corresponds to the given sequence's global free energy minimum,

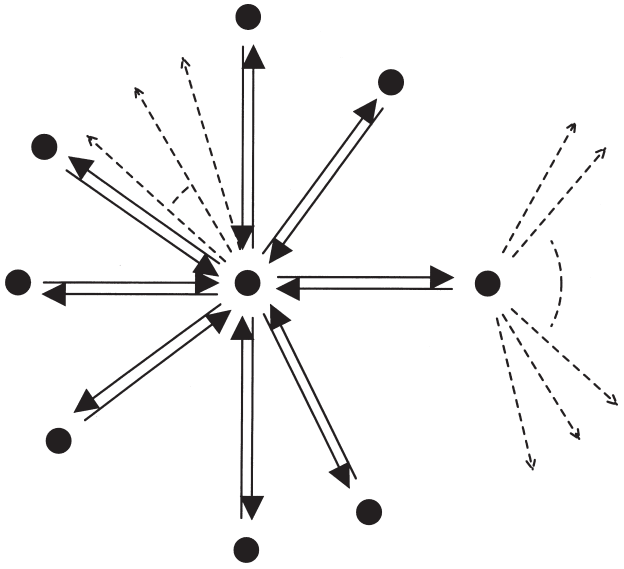

Figure 6 A simple illustration of the effect of sequence-space topology on evolutionary population distribution. As in Figure 2, viable sequences are denoted by dots. Here single-point mutations between viable sequences are represented by double solid arrows. Dotted arrows represent lethal mutations to nonviable mutants from viable sequences (depicted for two viable sequences as examples). We assume that every sequence has a total of N single-point mutational neighbours (including both viable and non-viable sequences), and that the mutation rate μ is uniform. For any sequence with population P, the population loss per unit time by undergoing single-point mutations to all N neighbouring sequences is given by $-\mu NP$. On the other hand, a sequence can only gain population from the mutations of its viable neighbours. Now let A_0 be the number of viable single-point neighbours for the centre sequence in this figure, and all other viable sequences surrounding the centre sequence have it as their only viable single-point neighbour. If P_0/P_1 is the steady-state population ratio of the centre sequence to each of the other A_0 sequences, this ratio must be maintained through mutational changes. This consideration implies that $-\mu NP_0 + \mu A_0 P_1 = cP_0$ and $-\mu NP_1 + \mu P_0 = cP_1$ for some constant c. Thus $P_0 = \sqrt{A_0}P_1$, ie a given sequence with more viable neighbours tends to have higher steady-state evolutionary populations. However, it should be noted that the total steady-state population of *all* sequences with only one viable neighbour in this example is equal to $A_0P_1 = \sqrt{A_0}P_0$, which is larger than the steady-state population of the centre sequence. (See Bornberg-Bauer and Chan (1999) for further discussions.)

Govindarajan and Goldstein (1998) simulated the evolution of a target structure's thermodynamic stability by point mutations. In this study, they considered two-dimensional 16mer chains, using all accessible conformations to compute thermodynamic stability (conformational enumerations *not* restricted to maximally compact 4×4 structures). Interestingly, they found that even if a model protein's folded structure is initially not its ground state and the sequence only folds to the target structure for kinetic reasons (ie under kinetic control), the model protein is likely to evolve toward sequences that have the given structure as its unique ground-state conformation, ie sequences that would fold to the target structure under thermodynamic control.

Using 25mer chains confined to the 1081 maximally compact 5×5 square-lattice conformations (Chan and Dill 1989), Taverna and Goldstein (2000a, 2002a, 2002b) performed population dynamics simulations. Their main findings are: (1) Evolution dynamics have an enhancing effect on the uneven distribution of structures above and beyond that predicted by static considerations of encodability or designability. The relative frequency of different structures has a greater-than-linear dependence on designability (Taverna and Goldstein 2000a). (2) Evolution dynamics and sequence entropy (see above) may provide a rationalisation (aside from the functional selection argument) for the observed marginal thermodynamic stability of real proteins. Marginally stable model proteins tend to dominate the population as evolution progresses (Taverna and Goldstein 2002a). (3) Evolution dynamics lead to a higher average mutational stability (ie enhanced sequence plasticity) in a sequence population than the average mutational stability among a random set of viable sequences in which each sequence is weighted equally (Taverna and Goldstein 2002b). This effect is consistent with that observed by Bornberg-Bauer and Chan as noted above. A similar conclusion about mutational robustness has also been reached by a non-SEM approach (van Nimwegen et al 1999). The underlying principle is rather straightforward since a sequence that has more neutral neighbours tend to have a higher chance of receiving population by back-mutation (Figure 6). These phenomena underscore the importance of population flux in and out of neutral networks during evolution, and are consistent with the concept of quasi-species (Eigen 1971; Eigen et al 1988), which posits that the survivability of a genotype is enhanced by being surrounded by fit genotypes. Similar features have also emerged in RNA simulations (Fontana and Schuster 1987; Huynen et al 1996) and for digital organisms (Wilke et al 2001).

Extending this modelling setup for population dynamics to the unconstrained conformational space of 16mers with a modified MJ potential on square lattices (using ΔG computed from all 802 075 conformations as the stability criteria), Williams et al (2001) performed evolutionary runs using the relative probability of each sequence to form a compact structure (but not the similarity to a pre-defined target structure) as a selection criterion. Punctuated-equilibrium-like behaviour was observed, but the number of compact states in the population rapidly reduced to one (Figure 5b). Thus, an additional effect of early 'freezing-in' comes into play in that the population decides early for a particular compact shape and rarely overturns its choice afterward.

The interpretation given by the authors is based on their group's previous work (Govindarajan and Goldstein 1997a,

1997b) (see Protein evolution section). In short, as the evolutionary dynamics progresses, selection pressure on each individual sequence increases as the population as a whole improves its average fitness. Consequently, 'evolutionary bridges' (regions between two structures) become rapidly depopulated, and thus increasingly difficult to facilitate switching of a population from one structural region in sequence space to another. Remarkably, the additional ligand-binding criterion in this model does not significantly affect the resulting distribution of structures. However, the early kinetic 'freezing in' strongly biases the distribution of structures in the evolutionary population relative to the underlying 'steady-state' or static distribution, suggesting that a similar effect might have influenced the distribution of folded structures among real proteins.

Finally, evolutionary considerations would not be complete without an understanding of recombination. Recombinations are important for genomic variations as they account for the development of pathogenicity islands, exon shuffling, and so on (Otto et al 1993; Barton and Charlesworth 1998). It is commonly held that the short-term effect of recombinations is rather malicious but it is outweighed by the long-term benefits for a population. The advantages include faster adaptation and more efficient linkage fixation. Recombination is clearly an important mechanism for protein evolution (Marcotte et al 1999; Patthy 1999; Voigt et al 2001a).

Exact enumerations and dynamic simulations (that mimic a fluctuating time-dependent environment) of the HP model have recently been used to investigate the exploratory effectiveness of point mutations alone versus point mutations with recombinations (Cui et al 2002). Perhaps not too surprisingly, this study confirms that with recombinations, new structures can be found much faster than without recombination because crossovers help to overcome 'evolutionary traps' (Figure 7a). More remarkably, certain local structural motifs reminiscent of autonomous folding units seem to help overcome sequence-space frustration, and hint at a hierarchical organisation of folding units (Figure 7b) and modular evolution (Figure 8). Subsequently, using a similar method based on HP-like model sequences of length $n = 25$ confined to a 5×5 square lattice, Xia and Levitt made the insightful observation that within a neutral net, recombination acting in concert with single-point mutations can lead to a much higher concentration of steady-state population around the

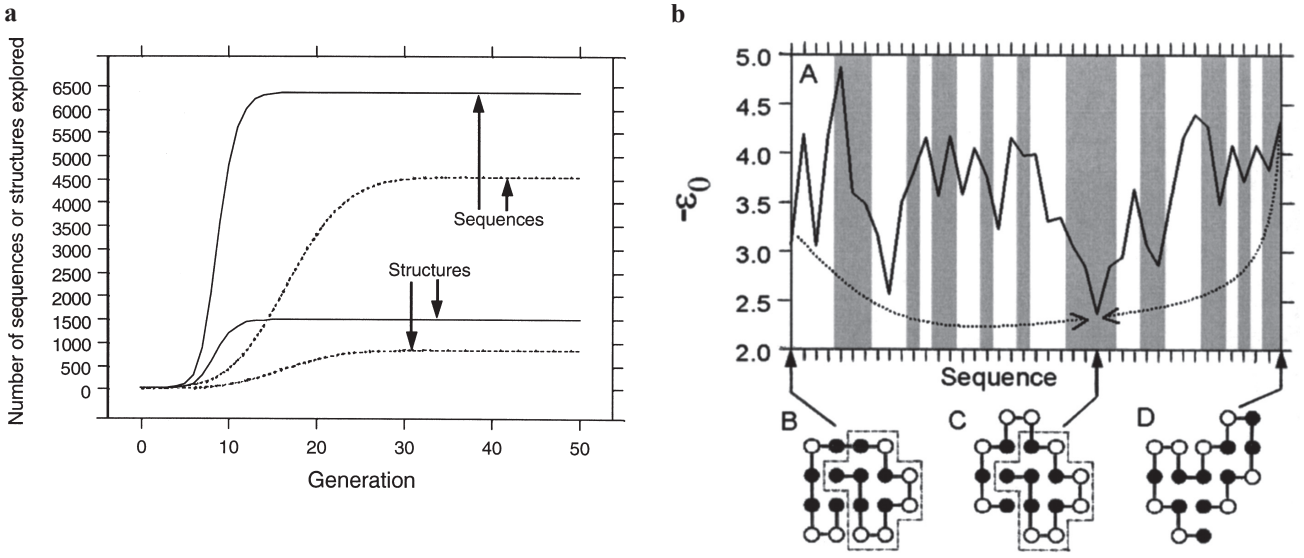

Figure 7 Recombination enhances genotypic and phenotypic innovation. **(a)** Number of sequences and structures visited as functions of number of generations in the n = 18 HP model. Evolutionary exploration by point mutations alone (dotted curves) is compared with that by point mutations plus crossovers (solid curves). This result shows that innovations in sequence and structure are significantly enhanced when point mutations are complemented by recombination. **(b)** Each graduation along the horizontal axis in (A) represents an HP model sequence (n = 18); sequences next to each other on the axis differ by one single-point mutation. A vertical stripe (white and grey shadings) contains sequences encoding for the same structure. The vertical scale $-\varepsilon_0$ is a mortality (negative fitness) measure. An optimised point-mutation path from each of the sequence/structure pairs (B or D) to the target structure (C) is shown. These sequence-space paths are one of the shortest among point-mutation paths with minimised total mortality climbs between the pair of given structures. (A) shows that even these paths need to traverse sequence space regions encoding for many other structures and overcome many mortality barriers before the target structure is reached. (A total of 19 different structures are encoded by the paths as marked by vertical stripes; two of the stripes encode for the same structure.) In contrast, recombination of the two starting sequences (B and D) yields the target structure (C) in one evolutionary step. In this crossover, the folded structure of a sequence fragment (dotted boxes in B and C) is maintained. This sequence fragment acts like an 'autonomous folding unit' because as an independent sequence it also folds uniquely to the same structure as that in the dotted boxes. Source: Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA*, 99:809–14. Copyright 2002. By permission of the National Academy of Science, USA.

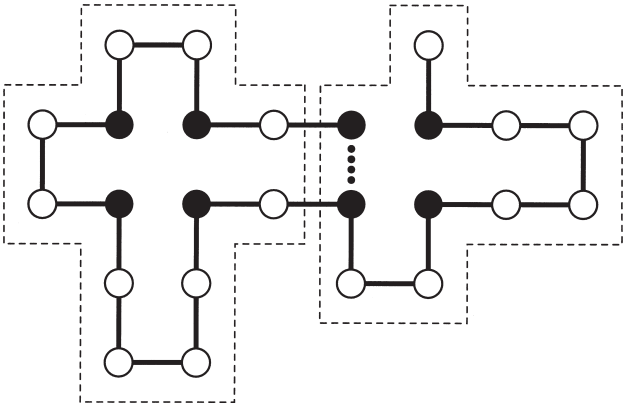

Figure 8 Modularity and autonomous folding units. An $n = 25$ HP sequence is shown in its unique native structure determined by Irbäck and Troein (2002). This structure may be divided into two domains with $n = 14$ (left) and $n = 11$ (right). We found that, as an independent $n = 14$ sequence, the subsequence on the left folds uniquely to the structure in the left dotted box. Similarly, as an independent sequence, the $n = 11$ contiguous sequence formed by joining (dotted bond) the singlet monomer on the right with the rest of the subsequence folds uniquely to the structure in the right dotted box. Thus, the two domains of this $n = 25$ folded structure may be viewed as lattice analogues of autonomous protein folding units.

prototype sequence (Xia and Levitt 2002) than that afforded by single-point mutations alone (Bornberg-Bauer and Chan 1999).

Some of these predictions may bear on recent protein engineering experiments using recombinations (Zhang et al 1997; Riechmann and Winter 2000; Voigt et al 2002). For instance, an interesting possibility suggested by the Cui et al (2002) results is that prototype sequences can be transformed into one another by swapping sequence fragments (Bornberg-Bauer 2002). Indeed, HP model enumerations indicate that for a pair of sequences with unique ground states, and especially so for prototype sequences, there is a significantly higher than random probability to recombine to form another unique or prototype sequence. This is because some sequence fragments are more popular than others among unique sequences, and even more so among prototype sequences (Cui et al 2002). These features share certain similarities with recent analyses of real protein sequences, in which a 'link' is defined between a pair of structural domains (or motifs at the sequence level) if they co-occur in at least one protein sequence. The resulting topology is found to have a small world network structure. In other words, a few domains are so popular that they co-occur with almost all other domains at least once in a protein while most domains occur either alone or are only rarely linked (Apic et al 2001; Wuchty 2001; Bornberg-Bauer 2002). In conjunction with knowledge gained from database analyses, we expect SEM-based conceptual development (Cui et al 2002; Xia and Levitt 2002) and other theoretical considerations (see Panchenko et al 1997; Bogarad and Deem 1999; Ancel and Fontana 2000; Voigt et al 2001a) will be useful for clustering sequence databases in more meaningful ways, and for deriving new and more sensitive sequence analysis methods by taking prior knowledge about domains (eg from precompiled domain signatures such as profiles or patterns) into account.

The SEM finding of uneven distributions of sequence fragments among unique and prototype sequences (Cui et al 2002) is closely related to questions about the nature and degree of statistical nonrandomness in the existing databases of real protein sequences (Pande et al 1994a; Strait and Dewey 1996; Irbäck et al 1996; Irbäck and Sandelin 2000; Schwartz et al 2001). This is a subject of an ongoing debate (cf the early study of White and Jacobs 1990) that has potentially important ramifications for sequence search methodologies. Hydrophobicity patterns in real proteins and proteinlike unique sequences in the HP model appear to possess non-negligible degrees of order or nonrandomness (Irbäck and Sandelin 2000). Thus, at least in the HP model, increasing randomness is correlated with increasing ground-state degeneracy of a sequence (Bornberg-Bauer 2002); and on average prototype sequences exhibit more order than non-prototype unique sequences (Cui et al 2002). Whether a similar trend exists in databases of real proteins remains to be explored.

Concluding remarks

We have described a number of SEM findings relevant to protein evolution, and in some cases have contrasted them with corresponding results for RNA. Based on this theoretical perspective, we have also discussed implications of these investigations on bioinformatics, especially with respect to sequence analysis. The following is a recapitulation of the main observations.

- Neutral nets and convergence are robust features across almost all protein models, including a recent SEM construct that uses a ligand-binding-like fitness criterion.
- The difference in the biophysics of intrachain interactions in proteins and RNA suggests strongly that their evolutionary strategies are different, making modular evolution necessary for proteins. In view of the topological properties of the model protein fitness landscape, early lock-in is likely.
- Hence, it is also likely that each module of protein evolution has arisen independently.

- The overrepresentation of certain folded structures in protein sequence space appears to be a generic feature of the sequence-structure map even in the absence of selection pressures favouring certain folded structures. However, the overrepresentation as we know it today might have been compounded by evolutionary dynamics.
- Sequence pattern nonrandomness arises naturally from the nature of protein sequence-structure map. Additional nonrandomness could arise from the above-noted evolutionary optimisation that might have led to enhanced structural overrepresentation.
- On average, protein sequence-space separations are significant between regions that encode for different folded structures.
- This phenomenon is closely related to the 'designing out' feature, which is robust across a variety of protein models. A sequence optimised to fold to one given native structure tends to be far away from sequence-space regions encoding for other folded structures.
- Therefore, though bi-structure switches between neutral nets are feasible, such mechanisms are probably not a common evolutionary route for protein structural innovation.

In closing, it is noteworthy that SEMs, originally developed for the study of general statistical mechanics principles of protein folding and structure, have been remarkably versatile in raising and addressing fundamental evolutionary questions. With a wider dissemination of their results, we expect that the novel theoretical perspectives they are uniquely positioned to provide will have an increasing productive impact on bioinformatics research and development.

Acknowledgments

We thank Matt Cordes, Yan Cui, Alan Davidson, Walter Fontana, Richard Goldstein, Chinlin Guo, Anders Irbäck, Kevin Plaxco, Chao Tang, Martin Vingron, Chris Voigt, Peter Wolynes, Wing Hung Wong and Tetsuya Yomo for helpful discussions. We thank Chris Voigt for his critical reading of an earlier version of this review and his insightful comments, and Hüseyin Kaya for his help with some of the figures. HSC thanks the Canadian Institutes of Health Research for financial support (CIHR grant no. MOP-15323). HSC is a Canada Research Chair in Biochemistry. EB acknowledges support from an MRC (UK) international recruitment grant.

Notes

[1] We note that a higher value for the thermodynamic parameter $\mathcal{F}$ does not necessarily imply a higher actual kinetic folding rate. For example, in a model kinetics study comparing an HP and its corresponding 'HP+' model (Chan and Dill 1998), the HP+ model has a significantly higher folding speed than the HP model. But the $\mathcal{F}$ value determined from the exact density of states of the HP+ model ($\mathcal{F} = 4.29$) is smaller than that of the HP model ($\mathcal{F} = 6.80$).

[2] Strictly speaking, homology means descent from a common ancestor. Since selection in most models considered here is based on a fitness criterion that defines the neutral nets, descendants arising from point mutations tend to be confined mostly to the same neutral net (see below). Thus, sequences belonging to the same neutral net may be referred to as homologous. Cui et al (2002) defined homology by neutral set instead of neutral net. Because an overwhelming majority of neutral sets in their model have only one neutral net per neutral set, their conclusions would be essentially unchanged if the present definition of homology was adopted.

[3] The challenge to transform one fold into another with less than 50% mutations.

References

Abkevich VI, Gutin AM, Shakhnovich EI. 1994a. Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J Chem Phys*, 101:6052–62.

Abkevich VI, Gutin AM, Shakhnovich EI. 1994b. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, 33:10026–36.

Abkevich VI, Gutin AM, Shakhnovich EI. 1996. How the first biopolymers could have evolved. *Proc Natl Acad Sci USA*, 93:839–44.

Anantharaman V, Koonin EV, Aravind L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res*, 30:1427–64.

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. *J Exp Zool*, 288:242–83.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310:311–24.

Babajide A, Hofacker IL, Sippl MJ, Stadler P. 1997. Neutral networks in protein space. *Fold Des*, 2:261–9.

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

Backofen R, Will S, Bornberg-Bauer E. 1999. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics*, 15:234–42.

Bak P, Flyvbjerg H, Lautrup B. 1992. Co-evolution in a rugged fitness landscape. *Phys Rev A*, 46:6724–30.

Barton NH, Charlesworth B. 1998. Why sex and recombination? *Science*, 281:1986–90.

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

Blackburne BP, Hirst J. 2001. Evolution of functional model proteins. *J Chem Phys*, 115:1935–42.

Bogarad LD, Deem MW. 1999. A hierarchical approach to protein molecular evolution. *Proc Natl Acad Sci USA*, 96:2591–5.

Bornberg-Bauer E. 1996. Structure formation of biopolymers is complex, their evolution may be simple. In Hunter L, Klein T, eds. Proceedings of the Pacific Symposium on Biocomputing; 1996 Jan 3–6; Hawaii. Singapore:World Scientific. p 97–108.

Bornberg-Bauer E. 1997a. Chain growth algorithms for HP type lattice proteins. In RECOMB proceedings; 1997 Jan 20–23; Santa Fe, NM, USA. New York: ACM Pr. p 47–55.

Bornberg-Bauer E. 1997b. How are model protein structures distributed in sequence space? *Biophys J*, 73:2393–403.

Bornberg-Bauer E. 2002. Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. *Z Phys Chem*, 216:139–54.

Bornberg-Bauer E, Chan HS. 1999. Modeling evolutionary landscapes: mutational stability, topology and superfunnels in sequence space. *Proc Natl Acad Sci USA*, 96:10689–94.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–70.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, 247:1306–10.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins Struct Funct Genet*, 21:167–95.

Bryngelson JD, Wolynes PG. 1989. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J Phys Chem*, 93:6902–15.

Buchler NEG, Goldstein RA. 1999. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins Struct Funct Genet*, 34:113–24.

Cejtin H, Edler J, Gottlieb A, Helling R, Li H, Philbin J, Wingreen N, Tang C. 2002. Fast tree search for enumeration of a lattice model of protein folding. *J Chem Phys*, 116:352–9.

Chan HS.1995. Kinetics of protein folding. *Nature*, 373:664–5.

Chan HS. 1998. Protein folding: matching speed and locality. *Nature*, 392:761–3.

Chan HS. 1999. Folding alphabets. *Nature Struct Biol*, 6:994–6.

Chan HS. 2000. Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins Struct Funct Genet*, 40:543–71.

Chan HS, Dill KA. 1989. Compact polymers. *Macromolecules*, 22: 4559–73.

Chan HS, Dill KA. 1990. The effects of internal constraints on the configurations of chain molecules. *J Chem Phys*, 92:3118–25 [Erratum: *J Chem Phys*, 107:10353 (1997)].

Chan HS, Dill KA. 1991. Sequence space soup of proteins and copolymers. *J Chem Phys*, 95:3775–87.

Chan HS, Dill KA. 1993. The protein folding problem. *Physics Today*, 46(2):24–32.

Chan HS, Dill KA. 1994. Transition states and folding dynamics of proteins and heteropolymers. *J Chem Phys*, 100:9238–57.

Chan HS, Dill KA. 1996. Comparing folding codes for proteins and polymers. *Proteins Struct Funct Genet*, 24:335–44.

Chan HS, Dill KA. 1998. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins Struct Funct Genet*, 30:2–33.

Chan HS, Kaya H, Shimizu S. 2002. Computational methods for protein folding: scaling a hierarchy of complexities. In Jiang T, Xu Y, Zhang MQ, eds. Current topics in computational molecular biology. Cambridge, Massachusetts: MIT Pr. p 403–47.

Chothia C. 1992. Proteins – 1000 families for the molecular biologist. *Nature*, 357:543–4.

Cordes MHJ, Burton RE, Walsh NP, McKnight CJ, Sauer RT. 2000. An evolutionary bridge to a new protein fold: interconversion of two native structures in a single mutant protein. *Nature Struct Biol*, 7:1129–32.

Cordes MHJ, Davidson AR, Sauer RT. 1996. Sequence space, folding and protein design. *Curr Opn Struct Biol*, 6:3–10.

Cordes MHJ, Sauer RT. 1999. Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Sci*, 8:318–25.

Cordes MHJ, Walsh NP, McKnight CJ, Sauer RT. 1999. Evolution of a protein fold in vitro. *Science*, 284:325–7.

Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M. 1998. On the complexity of protein folding. *J Comp Biol*, 5:423–65.

Crick FHC. 1968. Origin of genetic code. *J Mol Biol*, 38:367–79.

Crippen GM. 1991. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, 30:4232–7.

Cui Y, Wong WH. 2000. Multiple-sequence information provides protection against mis-specified potential energy functions in the lattice model of proteins. *Phys Rev Lett*, 85:5242–5.

Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA*, 99:809–14.

Dalal S, Balasubramanian S, Regan L. 1997. Protein alchemy: changing β-sheet into α-helix. *Nature Struct Biol*, 4:548–52.

Davidson AR, Lumb KJ, Sauer RT. 1995. Cooperatively folded proteins in random sequence libraries. *Nature Struct Biol*, 2:856–64.

Derrida B, Peliti L. 1991. Evolution in a flat fitness landscape. *Bull Math Biol*, 53:355–82.

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

Dill KA, Chan HS. 1997. From Levinthal to pathways to funnels. *Nature Struct Biol*, 4:10–19.

Ebeling M, Nadler W. 1995. On constructing folding heteropolymers. *Proc Natl Acad Sci USA*, 92:8798–802.

Ebeling M, Nadler W. 1997. Protein folding: optimized sequences obtained by simulated breeding in a minimalist model. *Biopolymers*, 41:165–80.

Eigen M. 1971. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10):465–523.

Eigen M. 1987. Stufen des Lebens. München: Piper.

Eigen M, McCaskill J, Schuster P. 1988. Molecular quasi-species. *J Phys Chem*, 92:6881–91.

Flamm C, Hofacker IL, Stadler PF. 1999. RNA in silico – the computational biology of RNA secondary structures. *Adv Complex Syst*, 2:65–90.

Fontana W, Konings DAM, Stadler PF, Schuster P. 1993a. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–404.

Fontana W, Schnabl W, Schuster P. 1989. Physical aspects of evolutionary optimization and adaptation. *Phys Rev A*, 40:3301–21.

Fontana W, Schuster P. 1987. A computer model of evolutionary optimization. *Biophys Chem*, 26:123–47.

Fontana W, Schuster P. 1998a. Continuity in evolution. On the nature of transitions. *Science*, 280:1451–5.

Fontana W, Schuster P. 1998b. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J Theor Biol*, 194:491–515.

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P. 1993b. RNA folding and combinatory landscapes. *Phys Rev E*, 47:2083–99.

Giugliarelli G, Micheletti C, Banavar JR, Maritan A. 2000. Compactness, aggregation and prion-like behavior of protein – a lattice model study. *J Chem Phys*, 113:5072–7.

Goodsell DS, Olson AJ. 1993. Soluble proteins: size, shape and function. *Trends Biochem Sci*, 18:65–8.

Govindarajan S, Goldstein RA. 1995. Searching for foldable protein structures using optimized energy functions. *Biopolymers*, 36:43–51.

Govindarajan S, Goldstein RA. 1996. Why are some protein structures so common? *Proc Natl Acad Sci USA*, 93:3341–5.

Govindarajan S, Goldstein RA. 1997a. The foldability landscape of model proteins. *Biopolymers*, 42:427–38.

Govindarajan S, Goldstein RA. 1997b. Evolution of model proteins on a foldability landscape. *Proteins Struct Funct Genet*, 29:461–6.

Govindarajan S, Goldstein RA. 1998. On the thermodynamic hypothesis of protein folding. *Proc Natl Acad Sci USA*, 22:5545–9.

Harrison PM, Chan HS, Prusiner SB, Cohen FE. 1999. Thermodynamics of model prions and its implications for the problem of prion protein folding. *J Mol Biol*, 286:593–606.

Harrison PM, Chan HS, Prusiner SB, Cohen FE. 2001. Conformational propagation with prion-like characteristics in a simple model of protein folding. *Protein Sci*, 10:819–35.

Higgs PG. 2000. RNA secondary structure: physical and computational aspects. *Q Rev Biophys*, 33:199–253.

Hinds DA, Levitt M. 1994. Exploring conformation space with a simple lattice model for protein structure. *J Mol Biol*, 243:668–82.

Hinds DA, Levitt M. 1996. From structure to sequence and back again. *J Mol Biol*, 258:201–9.

Hirst JD. 1999. The evolutionary landscape of functional model proteins. *Protein Eng*, 12:721–6.

Huynen MA. 1996. Exploring phenotype space through neutral evolution. *J Mol Evol*, 43:165–9.

Huynen MA, Stadler PF, Fontana W. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci USA*, 93:397–401.

Irbäck A, Petterson C, Potthast F. 1996. Evidence for nonrandom hydrophobicity structures in protein chains. *Proc Natl Acad Sci*, 93:9533–8.

Irbäck A, Sandelin E. 2000. On hydrophobicity correlations in protein chains. *Biophys J*, 79:2252–8.

Irbäck A, Troein C. 2002. Enumerating designing sequences in the HP model. *J Biol Phys*, 28:1–15.

Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA*, 81:1075–8.

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–5.

Karplus M, Šali A. 1995. Theoretical studies of protein folding and unfolding. *Curr Opn Struct Biol*, 5:58–73.

Kauffman S, Levin S. 1987. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol*, 128:11–45.

Kim DE, Gu HD, Baker D. 1998. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc Natl Acad Sci USA*, 95:4982–6.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge Univ Pr.

Knight RD, Freeland SJ, Landweber LF. 1999. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci*, 24:241–7.

Koehl P, Levitt M. 2002. Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci USA*, 99:1280–5.

Kolinski A, Milik M, Skolnick J. 1991. Static and dynamic properties of a new lattice model of polypeptide chains. *J Chem Phys*, 94:3978–85.

Kono H, Saven JG. 2001. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol*, 306:607–28.

Krause A, Vingron M. 1998. A new paradigm in sequence database searching and clustering. *Bioinformatics*, 14:430–8.

Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW. 2002. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J Mol Biol*, 316:225–33.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–97.

Lau KF, Dill KA. 1990. Theory for protein mutability and biogenesis. *Proc Natl Acad Sci USA*, 87:638–42.

Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–9.

Li H, Tang C, Wingreen NS. 1998. Are protein folds atypical? *Proc Natl Acad Sci USA*, 95:4987–90.

Lim WA, Sauer RT. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol*, 219:359–76.

Lipman DJ, Wilbur WJ. 1991. Modelling neutral and selective evolution of protein folding. *Proc R Soc London B*, 245:7–11.

Macken CA, Perelson AS. 1989. Protein evolution on rugged landscapes. *Proc Natl Acad Sci USA*, 86:6191–5.

Marcotte E, Pellegrini M, Eisenberg D. 1999. A census of protein repeats. *J Mol Biol*, 293:151–60.

Marshall SA, Mayo SL. 2001. Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol*, 305:619–31.

Martinez JC, Viguera AR, Berisio R, Wilmanns M, Mateo PL, Filimonov VV, Serrano L. 1999. Thermodynamic analysis of α-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry*, 38:549–59.

Maynard-Smith J. 1970. Natural selection and the concept of a protein space. *Nature*, 225:563.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998. Design of proteins with hydrophobic and polar amino acids. *Proteins Struct Funct Genet*, 32:80–7.

Minor DL, Kim PS. 1996. Context dependent secondary structure formation of a designed protein sequence. *Nature*, 380:730–4.

Miyazawa S, Jernigan RL. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18:534–52.

Munson M, Anderson KS, Regan L. 1997. Speeding up protein folding: mutations that increase the rate at which Rop folds and unfolds by over four orders of magnitude. *Fold Des*, 2:77–87.

Murray AJ, Lewis SJ, Barclay AN, Brady RL. 1995. One sequence, 2 folds: a metastable structure of CD2. *Proc Natl Acad Sci USA*, 92:7337–41.

Nelson ED, Onuchic JN. 1998. Proposed mechanism for stability of protein to evolutionary mutations. *Proc Natl Acad Sci USA*, 95:10682–6.

Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. 1997. Extracting protein alignment models from the sequence database. *Nucleic Acids Res*, 25:1665–77.

Newman MEJ, Engelhardt R. 1998. Effects of neutral selection on the evolution of molecular species. *Proc R Soc London B*, 265:1333–8.

Nussinov R, Jacobson AB. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci USA*, 77:6309–13.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature*, 372:631–4.

Otto SP, Feldman MW, Christiansen FB. 1993. Some advantages and disadvantages of recombination. *SFI Preprint*, 93-03-012.

Panchenko AR, Luthey-Schulten Z, Cole R, Wolynes PG. 1997. The foldon universe: a survey of structural similarity and self-recognition of independently folding units. *J Mol Biol*, 272:95–105.

Pande VS, Grosberg AYu, Tanaka T. 1994a. Non-randomness in protein sequences: evidence for a physically driven stage of evolution? *Proc Natl Acad Sci USA*, 91:12972–5.

Pande VS, Joerg C, Grosberg AYu, Tanaka T. 1994b. Enumerations of the Hamiltonian walks on a cubic sublattice. *J Stat Phys A*, 27:6231–6.

Pande VS, Grosberg AYu, Tanaka T. 1997. Statistical mechanics of simple models of protein folding and design. *Biophys J*, 73:3192–210.

Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*, 273:349–54.

Patterson M, Przytycka T. 1995. On the complexity of string folding. In Istrail S, Pevzner P, Shamir R, eds. Special issue on computational molecular biology. *Discrete Appl Math*, 73:217–30.

Patthy L. 1999. Protein evolution. Oxford: Blackwell.

Reidys C, Stadler PF, Schuster P. 1997. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull Math Biol*, 59:339–97.

Renner A, Bornberg-Bauer E. 1997. Exploring the fitness landscapes of lattice proteins, In Altman RB, Dunker AK, Hunter L, Klein TE, eds. Proceedings of the 1997 Pacific Symposium on Biocomputing; 1997 Jan 6–9; Honolulu, Hawaii. Singapore: World Scientific. p 361–72.

Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol*, 4:805–9.

Riechmann L, Winter G. 2000. Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc Natl Acad Sci USA*, 97:10068–73.

Rose GD, Creamer TP. 1994. Protein folding: predicting predicting. *Proteins Struct Funct Genet*, 19:1–3.

Šali A, Kuriyan J. 1999. Challenges at the frontiers of structural biology. *Trends Biochem Sci*, 24:M20–4.

Šali A, Shakhnovich E, Karplus M. 1994. Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J Mol Biol*, 235:1614–36.

Saven JG. 2001. Designing protein energy landscapes. *Chem Rev*, 101:3113–30.

Saven JG, Wolynes PG. 1997. Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J Phys Chem B*, 101:8375–89.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc London B*, 255:279–84.

Schuster P, Stadler PF. 1994. Landscapes: complex optimization problems and biopolymer structures. *Comput Chem*, 18:295–324.

Schuster P, Stadler PF, Renner A. 1997. RNA structures and folding: from conventional to new issues in structure predictions. *Curr Opin Struct Biol*, 7:229–35.

Schwartz R, Istrail S, King J. 2001. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci*, 10:1023–31.

Shahrezaei V, Ejtehadi MR. 2000. Geometry selects highly designable structures. *J Chem Phys*, 113:6437–42.

Shakhnovich EI. 1996. Modeling protein folding: the beauty and power of simplicity. *Fold Des*, 1:R50–4.

Shapiro BA. 1988. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci*, 4:387–93.

Shimizu S, Chan HS. 2002. Anti-cooperativity and cooperativity in hydrophobic interactions: three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. *Proteins Struct Funct Genet*, 48:15–30 [Erratum: *Proteins Struct Funct Genet*, 49:294 (2002)].

Sicheri F, Yang DSC. 1995. Ice-binding structure and mechanism of an antifreeze protein from winter flounder. *Nature*, 375:427–31.

Skolnick J, Kolinski A. 1990. Simulations of the folding of a globular protein. *Science*, 250:1121–5.

Stadler PF. 1995. Towards a theory of landscapes. In Lopéz-Peña R, Capovilla R, García-Pelayo R, Waelbroeck H, Zertuche F, eds. Complex systems and binary networks. New York: Springer-Verlag. p 77–163.

Stadler PF. 1999. Fitness landscapes arising from the sequence-structure maps of biopolymers. *J Mol Struct (THEOCHEM)*, 463:7–19.

Strait BJ, Dewey GT. 1996. The Shannon information entropy of protein sequences. *Biophys J*, 71:148–55.

Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P. 1996. Algorithm independent properties of RNA secondary structure predictions. *Eur Biophys J*, 25:115–30.

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

Taverna DM, Goldstein RA. 2000b. The evolution of duplicated genes considering protein stability constraints. In Altman RB, Dunker AK, Hunter L, Klein TE, eds. Proceedings of the 2000 Pacific Symposium on Biocomputing; 2000 Jan 5–9; Honolulu, Hawaii. Singapore: World Scientific. p 66–77.

Taverna DM, Goldstein RA. 2002a. Why are proteins marginally stable? *Proteins Struct Funct Genet*, 46:105–9.

Taverna DM, Goldstein RA. 2002b. Why are proteins so robust to site mutations? *J Mol Biol*, 315:479–84.

Thirumalai D, Woodson SA. 1996. Kinetics of folding of proteins and RNA. *Acc Chem Res*, 29:433–9.

Tompa P, Tusnády GE, Cserzö M, Simon I. 2001. Prion protein: evolution caught en route. *Proc Natl Acad Sci USA*, 98:4431–6.

Trinquier G, Sanejouand YH. 1999. New proteinlike properties of cubic lattice models. *Phys Rev E*, 59:942–6.

Vajda S, Vakser IA, Sternberg MJE, Janin J. 2002. Modeling of protein interactions in genomes. *Proteins Struct Funct Genet*, 47:444–6.

van Nimwegen E, Crutchfield JP. 2000. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull Math Biol*, 62:799–848.

van Nimwegen E, Crutchfield JP, Huynen, M. 1999. Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA*, 96:9716–20.

Viguera AR, Villegas V, Aviles FX, Serrano L. 1997. Favourable native-like helical local interactions can accelerate protein folding. *Fold Des*, 2:23–33.

Voigt CA, Kauffman S, Wang Z-G. 2001a. Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv Protein Chem*, 55:79–160.

Voigt CA, Martinez C, Wang Z-G, Mayo SL, Arnold FH. 2002. Protein building blocks preserved by recombination. *Nature Struct Biol*, 9:553–8.

Voigt CA, Mayo SL, Arnold FH, Wang Z-G. 2001b. Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci USA*, 98:3778–83.

Wang J, Wang W. 1999. A computational approach to simplifying the protein folding alphabet. *Nature Struct Biol*, 6:1033–8.

Wang JTL, Marr TG, Shasha D, Shapiro BA, Chirn GW, Lee TY. 1996. Complementary classification approaches for protein sequences. *Protein Eng*, 9:381–6.

Wang X, Minasov G, Shoichet, BK. 2002. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol*, 320:85–95.

White SH, Jacobs RE. 1990. Statistical distribution of hydrophobic residues along the length of protein chains: implications for protein folding and evolution. *Biophys J*, 57:911–21.

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412:331–3.

Williams PD, Pollock DD, Goldstein RA. 2001. Evolution of functionality in lattice proteins. *J Mol Graphics Modelling*, 19:150–6.

Wolynes PG. 1997. As simple as can be? *Nature Struct Biol*, 4:871–4.

Wong JTF. 1975. Co-evolution theory of genetic code. *Proc Natl Acad Sci USA*, 72:1909–12.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Jones DF, ed. Proceedings of the Sixth Intl Congress on Genetics. Volume 1. Brooklyn Botanic Gardens, New York, USA. p 356–66.

Wuchty S. 2001. Scale-free behavior in protein domain networks. *Mol Biol Evol*, 18:1694–702.

Xia Y, Levitt M. 2002. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA*, 99:10382–7.

Yomo T, Saito S, Sasai M. 1999. Gradual development of protein-like global structures through functional selection. *Nature Struct Biol*, 6:743–6.

Yue K, Dill KA. 1992. Inverse protein folding problem: designing polymer sequences. *Proc Natl Acad Sci USA*, 89:4163–7.

Yue K, Fiebig M, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA*, 92:325–9.

Zhang C-T. 1997. Relations of the numbers of protein sequences, families and folds. *Protein Eng*, 10:757–61.

Zhang J-H, Dawes G, Stemmer WPC. 1997. Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc Natl Acad Sci USA*, 94:4504–9.

Zuker M, Stiegler P. 1981. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9:133–48.