4

The Structurally Constrained Neutral Model of Protein Evolution

U. Bastolla, M. Porto, H.E. Roman and M. Vendruscolo

The observation that protein sequences accumulate substitutions in time at an almost regular rate [1] created a great interest in molecular evolution, suggesting that substitutions in protein sequences can be used as an effective 'molecular clock' for estimating the time elapsed from the last common ancestor among genes [1–5]. This approach opened a new avenue for reconstructing the tree of life by analyzing the sequences of orthologous genes, whose evolutionary tree coincides with the tree of the species containing them. The practical importance of the study of molecular evolution became therefore evident as a way to reconstruct natural histories.

In addition, the molecular clock hypothesis sparked a lively debate about the mechanisms of molecular evolution. Kimura [6, 7] and King and Jukes [8] proposed that most substitutions in protein sequences are fixed in evolving populations not because they offer a selective advantage but, rather, because they are effectively neutral and therefore invisible to natural selection. The 'neutral theory' could account for the regular rate in time of the accumulation of amino acid substitutions. It failed, however, to predict correctly other features of the evolutionary process, among which the variance of the number of substitutions [9].

One is now starting to understand the reasons for this apparent limitations of neutral theories, thanks to the recent progress in structural biology. This progress has begun to make possible the use of structural information in evolutionary studies, starting with the pioneering works of the Vienna group on the RNA model [10–12] (see also the chapter by Schuster and Stadler in this book), whereas the study of molecular evolution was initially almost entirely based on the analysis of macromolecular sequences [3, 4, 7]. It appears that a paradigm shift is taking place in the field of molecular evolution, from coding symbols (sequence) to coded meaning (structure and function). This book investigates this new approach at several levels of biological organization.

In this chapter, we review some results that were obtained through approaches in which the structural stability of the native state of proteins is taken explicitly into account as a constraint on the evolutionary process [13–30], and

in particular through the Structurally Constrained Neutral (SCN) model of protein evolution [31, 32].

We will also show that several results of SCN simulations can be rationalized and rederived analytically by considering a vectorial representation of protein sequences and structures. In this approach, protein sequences are represented as hydrophobicity profiles HPs [33] and protein structures are represented through the principal eigenvector (PE) of the contact matrix [34–37]. As we have shown that the optimal HP and the structural profile are strongly correlated [38], an 'optimal' HP can be derived, i.e. the profile best compatible with a given protein structure. In simulations of SCN evolution, sequence vectors move around this optimal one. This scheme provides us with a framework that can be used to predict, by analytical calculations, site-specific conservation due to structural constraints and site-specific amino acid distributions [39, 40].

4.1 Aspects of Population Genetics

First of all, we need to state some terminology. A mutation is a microscopic event in which the sequence of a gene is altered in a single individual. At the population level, a substitution is a macroscopic event in which the representative, or wild-type, gene changes as a result of the fixation of a mutant gene.[1] Natural selection mediates this transition from the microscopic to the macroscopic level. In physical sciences, a similar role is played by statistical mechanics, which explains macroscopic phenomena in terms of the behaviour of their microscopic components. One of the aims of this chapter is to explore this analogy further.

Three main factors influence the fixation of a mutant allele in a population: the size of the population, M; the selective effect of the mutation, measured through its fitness relative to the wild-type, s; and the rate at which mutations occur, measured in mutations per gene and generation, μ.

4.1.1 Population Size and Mutation Rate

In most of this chapter, we will consider the limit of very small mutation rates, $M\mu \ll 1$, as it is customary in classical population genetics. For $M\mu \ll 1$, the time scale for the appearance of a new mutant ($1/\mu$) is much larger than the time scale for fixation of a neutral allele, which spans on the average M generations. This limit implies that the population is fairly homogeneous genetically, and at any generation there is at most one mutant arising. This has been termed the 'blind-ant' regime [41] because the population can only test a very small neighbourhood in genotype space at any time. The opposite regime, $M\mu \gg 1$, is assumed to hold in the 'quasispecies' model [42, 43], which

[1] Fixation of a mutation takes place when all individuals in the population are descendent of one individual bearing that allele.

considers infinite population sizes (concerning this regime, see the chapters by Jain and Krug and by Lázaro in this book).

To justify the choice of the blind-ant regime, we note that the mutation rate in mammalian genomes was estimated to be 5×10^{-9} per nucleotide per year [3], which, for a species with generation time of two years and a protein of 600 nucleotides (i.e. 200 amino acids) yields $\mu = 6 \times 10^{-6}$. An even smaller value of μ would have resulted by considering that many mutations are synonymous. For a population of effective size $M = 10^5$ (already a quite large estimate)[2] one obtains $M\mu = 0.6$. Although this value is not so small, numerical studies reveal that the results valid in the blind-ant regime continue to be valid qualitatively for $M\mu$ of order one (see Sect. 4.1.5).

It has been argued that the opposite regime of large $M\mu$ is valid for RNA viruses (see the chapter by Lázaro in this book), which have very high mutation rates, of the order of one nucleotide per genome per year [44], corresponding to $\mu \approx 10^{-1}$. Their effective population size is, however, quite reduced because of the bottlenecks that the population suffers when transferred from one host to the other (in these cases, the effective population size essentially coincides with the population at the bottleneck [4]).

4.1.2 Natural Selection

The other relevant parameter for the evolutionary dynamics is the difference in fitness between competing alleles. Since reproduction is inherently stochastic, there is a chance that the less fit allele is fixed even starting as a single individual. Different stochastic models of the reproductive process give qualitatively similar results. We illustrate them through the Moran's birth and death process [45]. According to this model, the probability that a mutant allele B with fitness $F(B)$, arising as a single individual in a haploid[3] population of size M, substitutes the wild-type A with fitness $F(A)$, is given by

$$P_{\text{fix}}(A \to B) = \frac{1 - \mathrm{e}^{f(B)-f(A)}}{1 - \mathrm{e}^{M[f(B)-f(A)]}}\,, \tag{4.1}$$

where $f(x) = \log[F(x)]$ with $x = A, B$. We will define in the following $s = f(B) - f(A)$. Notice that if $|Ms|$ is small there is a significant probability that even deleterious mutations ($s < 0$) are eventually fixed in the population.

Berg et al. [46] and Sella and Hirsh [47] have recently noticed that the above formula has an interesting analogy with the stochastic processes used to simulate statistical mechanical systems, since it satisfies the condition of

[2] The effective population size is the effective number of breeding adults in a population after adjusting for diverse factors, including reproductive dynamics. The effective population size is usually much less than the actual number of living or reproducing individuals [7].

[3] Haploid organisms carry one single copy of each chromosome, in difference to diploid organism carrying two copies of each chromosome.

detailed balance, $\pi(A)\,P(A \to B) = \pi(B)\,P(B \to A)$, with respect to a stationary distribution $\pi(A)$ that is analogous to a Boltzmann distribution in statistical physics (see the chapter by Lässig in this book). If the mutation process satisfies detailed balance with respect to a stationary distribution $\pi_{\mathrm{mut}}(A)$, as it is assumed in many models of molecular evolution [4], then the stationary distribution of the substitution process is

$$\pi(A) = \frac{1}{Z}\,\pi_{\mathrm{mut}}(A)\,\mathrm{e}^{Mf(A)}\,. \tag{4.2}$$

This equation is formally identical to a Boltzmann distribution in statistical physics if one identifies the logarithmic fitness $f(A)$ as the energy and the population size M as the inverse temperature (Z is a normalization constant). Smaller populations evolve at higher temperature, in the sense that the evolution is more dominated by stochastic events, and their mean fitness is lower than for corresponding larger populations.

The above result is valid for the small mutation rate regime. It is interesting that a formal analogy between evolving systems and statistical mechanical systems can be derived also for the quasi-species regime, where the infinite population limit is considered. In this case, the mutation rate μ, considered to be vanishingly small in the previous approach, plays the role of the temperature [48, 49]. For a treatment of this subject (see Chap. 14 by Jain and Krug).

4.1.3 Mutant Spectrum

We now go back to classical population genetics. It is customary to divide mutations into four classes, depending on their fitness effect (for a deeper discussion of this topic, see Chap. 13).

1. *Strongly deleterious mutations*: $Ms \ll -1$. These mutations decrease significantly the fitness of the individuals carrying them and they are soon removed from the population through purifying selection.
2. *Nearly neutral mutations*: $-\log(M) \leq Ms \leq \log(M)$. The fitness effect of these mutations is of the same order of importance as are reproductive fluctuations, and their fate is determined both by selection and by random drift [50, 51] (see also the chapter by Ohta in this book). Deleterious mutations in this range have a non-vanishing probability to lead to substitutions. The detailed balance condition, satisfied by several models of the substitution process, including the one presented above, implies that the frequency of mildly deleterious and mildly advantageous substitutions must be equal on average [47], as also previously noted by several authors, which is in contrast with the emphasis of some studies on mildly deleterious substitutions. The advantageous compensatory substitutions play an important role in the dynamics of viral populations, as discussed in the chapter by Lázaro in this book. For small $|Ms|$, the average time required for fixation of these substitutions is of the order of the population size M.

3. *Neutral mutations*: They have negligibly small effects on the fitness, $Ms \approx 0$ and can spread in the population through random genetic drift. The probability of fixation of a neutral mutation is $1/M$, and the expected time for fixation is of order M.
4. *Advantageous mutations*: $Ms \gg 1$. These mutations are efficiently fixed in the population through natural selection with probability close to one, and the time for fixation increases only logarithmically with the population size as $\log(M)/s$.

This classification is useful for distinguishing between different evolutionary scenarios, as advantageous, neutral and nearly neutral mutations can lead to substitutions. In the early years of population genetics, the emphasis was placed on the positive selection of advantageous mutations as the dominant force acting on the substitution process [52]. However, the accumulation of protein sequences eventually changed this view. To explain the very high amount of heterozygosity found in natural populations, as well as the molecular clock hypothesis, at the end of the 1960s Kimura [6] and King and Jukes [8] proposed that most substitutions are selectively neutral. This hypothesis, provocative and controversial at that time, lead to a simple mathematical model of the substitution process that will be discussed in Sect. 4.1.4. The neutral model is now considered by many as the null model of molecular evolution, and distinguishing positive selection from a neutral background is the subject of a vast area of evolutionary sequence analysis [53,54]. Subsequently, Ohta and Kimura [50] introduced the concept of nearly neutral substitutions, and Ohta [51] proposed that most substitutions belong to this class.

As more specifically discussed in the chapter by Ohta in this book, there are testable differences between neutral and nearly neutral substitutions, in particular: (a) The rate of nearly neutral substitutions, especially non-synonymous ones, is expected to decrease with population size.[4] This dependence can explain the discrepancies observed between various mammalian groups in the substitution rates per generation [55]. (b) The presence of nearly neutral substitutions implies that compensatory substitutions must be positively selected. This might explain the surprisingly high level of positive selection detected recently [54] using the McDonald and Kreitman test [53]. (c) In nearly neutral, but not in neutral, evolution, macromolecular properties are expected to be less optimized in smaller populations. Studies of endosymbiotic bacteria, which have small effective populations because of the bottleneck in the transmission from one host to its offsprings, have predicted that r-RNA molecules coded in the genomes of endosymbiotic bacteria have lower thermodynamic stability [56] and that their proteins are less stable with respect to misfolding [57]. These findings are consistent with the high expression of chaperones, which are proteins that assist the folding of other proteins, observed in endosymbiotic

[4] In principle, also the neutral substitution rate should decrease with the population size since the condition for a mutation to be neutral is $Ms \approx 0$. This effect, however, is usually neglected in mathematical models.

bacteria [58], and that can favour fitness recovering in a bacterial population subject to strong bottlenecks [59] (see Chap. 7).

4.1.4 Neutral Substitutions

The neutral theory of Kimura is based on the assumption that the fitness effect of a mutation with respect to the wild-type, s, has a bimodal distribution, with the most likely effects corresponding either to strongly disadvantageous ($Ms \ll -1$) or to neutral mutations ($Ms \approx 0$). Advantageous mutations are not considered because they are expected to be rare, at least for proteins that maintain the same function and evolve in the rather stable cellular environment [60]. The neutral theory therefore applies to families of orthologous proteins, whose evolutionary tree coincides with the species tree, and whose function and structure is expected to be conserved in evolution. On the other hand, paralogous proteins, which diversified after an event of gene duplication specializing into different functions (as for instance myoglobin and the two hemoglobin chains), undergo several positively selected substitutions in the process of developing a new function, as it is witnessed by the acceleration of the substitution rate after gene duplication [3]. Nearly neutral mutations are not considered for the sake of mathematical simplicity. From the point of view of the neutralist–selectionist controversy that was discussed for several decades in the molecular evolution literature, nearly neutral substitutions were often considered on the same ground as strictly neutral one, despite the differences discussed in the previous section.

In Kimura's model, neutral mutations undergo a diffusion process that in the population genetics literature receives the name of 'random genetic drift'. The rate at which neutral mutations occur in individual genes is μx, where μ is the mutation rate and x is the probability that a mutation is neutral. This probability is considered to be independent of population size M, even though, strictly speaking, the condition that a mutation is neutral is $s \ll 1/M$. The connection between the population size and the substitution rate lays at the heart of the nearly neutral theory and distinguishes it from the original neutral theory.

The number of neutral mutations arising in one generation is therefore $M\mu x$ and, since the probability that one of them substitutes the wild-type is $1/M$ (all the M genes have the same selective value), the neutral substitution rate per generation is given by

$$\frac{\mathrm{E}[S_t]}{t} = \mu x \tag{4.3}$$

and it is independent of M. Here, S_t is the number of accepted neutral mutations in a time interval t. This provides a sort of molecular clock, in agreement with the earliest empirical observations [7], but in worse agreement with the so-called generation time effect (see the chapter by Ohta in this book).

Another assumption, which we call the 'homogeneity hypothesis', is that the neutral mutation rate $x(\mathbf{A})$,[5] which in principle may be different for all sequences $\mathbf{A}$, is constant throughout evolution, $x(\mathbf{A}) \equiv x$. As shown later, this hypothesis implies that the number of neutral substitutions has a Poissonian distribution in the low mutation limit $M\mu \ll 1$. The population, as we mentioned above, is fairly homogeneous in this limit and there is at most one mutant arising at each generation. The number of mutations taking place in time t in an individual lineage is a Poissonian variable with mean value μt. For a population, the number of mutations is the sum of M Poissonian variables, and it is still Poissonian with mean $M\mu t$. The probability that one of these mutants become fixed is the product of the probability that the mutation is neutral, x, times $1/M$. Since at every generation there is at most one mutant, the probability of n out of m mutants becoming fixed is $\binom{m}{n}(x/M)^m(1-x/M)^{m-n}$. Therefore, the probability that there are n neutral substitutions within a time interval t is given by

$$\begin{aligned} \mathrm{P}\{S_t = n\} &= \sum_{m=n}^{\infty} \mathrm{e}^{-M\mu t} \frac{(M\mu t)^m}{m!} \binom{m}{n} \left(\frac{x}{M}\right)^n \left(1 - \frac{x}{M}\right)^{m-n} \\ &= \mathrm{e}^{-\mu x t} \frac{(\mu x t)^n}{n!} \,. \end{aligned} \tag{4.4}$$

As one can see, the result is a Poissonian variable with average value $\mu x t$. The homogeneity hypothesis seems at first sight very plausible since the neutral fraction x results from the average over a large number of sites in a gene. If the evolving sites are uncorrelated, the law of large numbers implies that the fluctuations of x vanish. However, as we shall see later, stability constraints introduces global correlations between the sites of protein coding genes, so that the homogeneity hypothesis is violated in models that take into account such stability constraints.

4.1.5 Beyond the Small $M\mu$ Regime: Neutral Networks

In the next sections, we shall consider the small $M\mu$ limit (the blind-ant regime). In this regime, the substitution process can be represented through the evolution of a single wild-type sequence. It should be emphasized that this set-up does not correspond to a one-individual population, but rather to a large population with a small mutation rate $\mu \ll 1/M$, so that most individuals have the same genotype. The population maintains the wild-type genotype until one of the possible neutral mutations is fixed. One time step in this set-up corresponds to the typical time for the fixation of a neutral mutation, M.

[5] We adopt a notation in this chapter where bold-face mathematical symbols such as $\mathbf{A}$ indicate vectors (sequences) or matrices, whereas A_i indicates the i-th component of $\mathbf{A}$.

When the mutation rate is not small, however, the fate of a genotype depends not only on its fitness $F(\mathbf{A})$, as indicated in (4.2) but also on the fitness of its neighbours in sequence space that can be connected to it through point mutations. An important quantity in this regime is the mutation load, i.e. the fraction $\mu(1 - x(\mathbf{A}))$ of offsprings of individuals with genotype $\mathbf{A}$ that undergo lethal mutations. If the homogeneity hypothesis does not hold and $x(\mathbf{A})$ fluctuates in sequence space, the population dynamics may favour genotypes with large neutrality fraction $x(\mathbf{A})$ and hence small mutation load. The parameter that controls whether this is the case is the product $M\mu$. As discussed earlier, a population with very small $M\mu$ can be represented through a single effective sequence evolving in the blind-ant regime. In the opposite limit of very large $M\mu$ (the quasi-species regime [42]), the distribution of the population in sequence space can be obtained analytically for a neutral model in which all viable sequences have the same fitness $F(\mathbf{A})$.

The result can be cast into a simple form [41]: Define the neutral connectivity matrix $x(\mathbf{A}, \mathbf{A}')$ to be 1 if $\mathbf{A}$ and $\mathbf{A}'$ are two viable sequences that can be connected through one point mutation and 0 otherwise. This matrix describes a neutral network of viable sequences interconnected through point mutations [10]. The stationary distribution of the fraction of individuals with genotype $\mathbf{A}$, $\rho(\mathbf{A})$, has to satisfy the stationarity condition $\rho(\mathbf{A}) = \sum_{\mathbf{A}'} \rho(\mathbf{A}')\, x(\mathbf{A}', \mathbf{A})$ and therefore it is proportional to the component of the PE of the neutral connectivity matrix for genotype $\mathbf{A}$. This component constitutes a sort of effective neutral connectivity of sequence $\mathbf{A}$ and it is positively correlated with the fraction of neutral neighbours $x(\mathbf{A})$ (see Sect. 4.4.1). Therefore, sequences with large $x(\mathbf{A})$ are more populated, and the mutation load is reduced.

Van Nimwegen et al. [41] simulated population dynamics on a neutral network $x(\mathbf{A}, \mathbf{A}')$, obtained from the predicted folding properties of a small RNA molecule. They found that the blind-ant regime is a good approximation up to $M\mu \approx 10$ and the large $M\mu$ regime is approached at $M\mu \approx 200$. Similar results were obtained by Wilke [61] using the neutral network obtained through the predicted folding thermodynamic properties of a model protein. We argue that the value of $M\mu$ at which the cross-over of the two regimes takes place depends on the correlation length of $x(\mathbf{A})$ in sequence space, ℓ_x. In fact, in neutral evolution the population occupies a region in sequence space around the wild-type with radius of order $M\mu$ mutations [45]. If this radius is smaller than ℓ_x, then all values of $x(\mathbf{A})$ in the population are fairly similar and the small differences in the mutation load can not be fixed in the population.

For animal and plant populations, characterized by small mutation rate and effective population sizes of tens of thousands of individuals, $M\mu$ is of order one and one would expect that the blind-ant regime is still a good approximation to the neutral dynamics. On the contrary, viral populations have large $M\mu$, compatible with the cross-over region towards the quasi-species regime.

We end this section with a summarizing comparison between the two limiting regimes of population genetics. Population genetics models can be

simplified in two opposite regimes: very small (blind-ant regime) and very large (quasi-species regime) $M\mu$. In both cases, a formal analogy with statistical mechanical systems can be established. For $M\mu \ll 1$, when the population is fairly homogeneous, the negative of the logarithmic fitness plays the role of the energy function and the inverse of the population size plays the role of temperature. For $M\mu \gg 1$, when the population is very spread in sequence space, a combination of the negative of the logarithmic fitness with a mutation term plays the role of the energy and the mutation rate plays the role of temperature [48, 49] (see the chapter by Jain and Krug in this book). As the simulations by van Nimwegen et al. [41] and by Wilke [61] show in this case, even when mutant alleles are completely neutral under the point of view of the fitness, they may not be neutral under the point of view of mutation resistance. In the following, only the small $M\mu$ regime will be examined, since this is the relevant regime for many biological populations, most notably higher eukaryotes.

4.2 Structural Aspects of Molecular Evolution

4.2.1 Neutral Theory and Protein Folding Thermodynamics

The thermodynamic stability of the native state is a strong constraint on molecular evolution, and a consequence of the more general requirement of maintaining the biological function [62]. The native state of a protein must be stable with respect to both unfolding and misfolding [63]. However, the stability against unfolding and stability against misfolding are anticorrelated [57, 64]. Therefore, natural selection cannot achieve simultaneously the optimal value for both stability requirements and has to trade off between them.

Natural selection eliminates mutations that reduce folding stability and favors the fixation of more stable proteins. Nevertheless, natural proteins are only marginally stable against unfolding [65], and it is not difficult to engineer protein mutants to improve their stability. Moreover, a large number of mutations do not alter significantly the measured thermodynamic stability or the function of the protein. In the framework of the neutral theory of molecular evolution [6], these results can be interpreted, assuming that changes increasing folding stability are selectively neutral above some specific thresholds. According to this hypothesis, the threshold values are most frequently realized in protein evolution, because they correspond to an overwhelming portion of sequence space. This framework provides a possible explanation for the relatively low stability of native states of proteins [22] and for the fact that the observed amino acid occurrences are very close to the ones predicted from nucleotide occurrence frequencies [66, 67].

4.2.2 Structural Conservation and Functional Changes in Protein Evolution

It has since long been established that protein structures evolve much more slowly than protein sequences [68,69]. Methods of protein structure prediction on the basis of sequence homology are therefore quite successful [70]. Algorithms for comparing protein structures typically reveal distant evolutionary relationships between proteins having low sequence similarity [68]. Although these observations can be attributed to both sequence divergence and structure convergence, careful analysis of specific cases and more accurate methods for detecting sequence homology [71] suggest that sequence divergence beyond the limits of detectable homology is rather common (see e.g. [72] and the chapter by Dokholyan and Shakhnovich in this book). This prevalence of structural conservation has made it possible to create databases in which protein structures are classified into distinct structural groups with the same overall architecture (folds) [68, 73, 74]. For example, proteins classified in the same fold in the FSSP database [68] show a distribution of sequence identity comparable to that of random pairs of sequences [69]. Nevertheless, other indicators of structural changes often show a regular behaviour. For instance, within a given fold, the root mean square deviation between homologous proteins increases as sequences diverge [75].

Protein function, instead, is not as much conserved as the underlying structure, making its prediction rather difficult [76]. New functions are often created through gene duplication followed by differential regulation and recruitment of one of the copies to a new function [3]. In the transition to a new function, proteins accumulate substitutions, which may be fixed through positive selection, in a process that usually does not change significantly the overall fold.

Despite these general rules, several examples of proteins with detectable homology and yet different folds have been provided [77]. In these cases, the evolutionary changes are usually mediated through large scale mutations, such as insertion or deletions of entire secondary structure elements and circular permutations. As a consequence, the concept of protein fold has been reconsidered, and it has been suggested that insertions or deletions of secondary structure elements can provide a mechanism to connect many known folds [78]. Significant similarities between folds previously classified as distinct, possibly pointing at distant evolutionary relationships, were identified by Orengo and colleagues through an algorithm of protein structure comparison at the level of secondary structure [79] (see also the chapter by Ranea et al. in this book). In the majority of cases, however, point mutations and insertions or deletions of single residues do not seem to have produced evolutionary transitions to different protein folds. Therefore, in particular in the evolution of proteins that retain their function, the concept of protein fold can still be considered useful.

4.2.3 Models of Molecular Evolution with Structural Conservation

Structural stability was first considered in models describing the molecular evolution of RNA structures [10]. Schuster and co-workers described neutral networks in sequence space, associated to specific macromolecular structures (see also the chapter by Schuster and Stadler in this book).

In this view, structurally constrained molecular evolution proceeds along neutral networks, whose properties have a large impact on the evolutionary process. Schuster et al. showed that, in the case of some common RNA secondary structures, the neutral networks are dense in sequence space, and that networks of different common structures can be connected through a small number of point mutations [10]. These results suggest a view of RNA structural evolution as adaptation through neutrality, in which evolution proceeds along a neutral network until a crossing point to a fitter structure is found [11, 12].

Inspired by these studies, several authors introduced models of protein evolution with structural conservation. In this section, we shortly review some of these models. These models differ in the way the molecular structure is represented and the requirement of thermodynamic stability of the target structure is implemented. In the case of RNA, efficient algorithms can determine, approximately but reliably, the secondary structure of minimal energy for a given sequence [80]. Equivalent algorithms do not exist for protein tertiary structures. Therefore, several groups represented protein structures as self-avoiding walks on the simple cubic or square lattice, studying them by means of Monte Carlo simulations. The idea behind this approach is that qualitative properties of the evolution of lattice models can be transferred to real proteins. Other groups also adopted simplified off-lattice representations of protein structures, which were studied through effective energy functions, analogous to those used for lattice models. The two approaches usually yield qualitatively similar results. One should also distinguish between the approaches that impose only the requirement that the target structure has minimal energy, from those that further require that the energy landscape is well correlated. In the latter, all structures that are very different from the native one are energetically separated by a large energy gap from it, therefore favouring stability against misfolding.

Bornberg-Bauer and co-workers [13, 14] studied lattice polymers by imposing the condition that, for sequences in the neutral network, the energy of the target structure should be lower than that of all alternative structures, thus following closely the original RNA model. They studied the structures on a two-dimensional lattice and represented the sequences by a two-letter (hydrophobic-polar) code. Such a simplified protein model is amenable to exact enumeration of both conformations and sequences, and enabled Bornberg-Bauer and co-workers to establish that in the case of lattice proteins, neutral networks are disconnected in sequence space. They also discovered that these neutral networks are centred around the so-called *prototype sequence*,

which is the sequence of maximal stability for a given structure, both mutationally and thermodynamically. Furthermore, these studies indicated that protein structures can be changed through point mutations, analogously to what was previously found for the RNA model.

Babadje et al. [15] adopted simplified representations of real protein structures, evaluating how well test sequences fit the target structure through a measure (the Z-score [81]) of the energy difference with respect to a set of alternative structures. They found that protein sequences can diverge almost as much as random pairs of sequences despite maintaining a high compatibility with the original structure.

Shakhnovich and Gutin [16] proposed an evolutionary model in which selection for fast folding is imposed in the framework of a lattice model, but without requiring the conservation of a particular structure. Later, Dokholyan and Shakhnovich [17] extended this approach considering sequences of fixed composition for which the target structure was required to have low energy. Evolution was modelled as a Monte Carlo process in sequence space, and large entropy barriers were found to separate clusters in sequence space. Mirny and Shakhnovich [18] analysed amino acid conservation in five of the most populated protein folds, identifying structural features correlated with conservation.

Dokholyan and Shakhnovich [19] modelled the process of gene duplication followed by structural divergence, showing that it can account for some of the statistical features of observed protein folds, most notably the almost power law distribution of the number of proteins per fold, and in addition that the model provides useful predictions concerning protein function (see also the chapter of Dokholyan and Shakhnovich in this book).

Goldstein and colleagues [20, 21] used lattice polymers to study a fitness landscape where the fitness of protein structures is given by their foldability, a concept borrowed from the spin-glass model of protein folding. They found that foldability can vary broadly, where structures with similar and large foldabilities are clustered together in structure space. When the selective pressure is increased, evolutionary trajectories become increasingly confined to 'neutral networks', where the sequence can be significantly changed while a constant structure is maintained. In a subsequent work, Taverna and Goldstein [22] showed that the marginal stability of proteins is a direct consequence of the hypothesis that changes in stability are neutral above some threshold and also of the high dimensionality of the sequence space.

Bussemaker et al. [23] obtained the interesting prediction that, in the lattice model they studied, the stability of small proteins is rather insensitive to random mutations. Tiana et al. [24] performed an exhaustive study of the effects that single mutations have on the stability of the native structure of a lattice protein, simulating the folding dynamics through a Monte Carlo approach. They classified protein sites into three types according to their robustness to mutations: 'green' sites, where mutations do not produce any relevant effect on stability (typically at the surface of the structure), 'yellow'

sites for which the structure is slightly modified and 'red' sites (typically at the core of the structure) where mutations have a disruptive effect.

Parisi and Echave [25,26] studied the impact of structural conservation on protein evolution, in a similar spirit to the SCN model that will be described in next section; the main difference is that they did not impose conditions on the stability of alternative structures. They simulated site-specific amino acid transition matrices, which were used in the calculation of the likelihood of families of protein sequences given their phylogenetic tree. In this way, they showed that the use of structural information can improve notably the likelihood of evolutionary models, and their ability to distinguish between different phylogenies.

Xia and Levitt [27,28] used a two-dimensional lattice model and performed an exhaustive enumeration of the space of all sequences and the space of all structures. They found that, when evolution is dominated by mutation, the preference of the prototype sequence is not strong enough to offset the huge size of sequence space, so that most native sequences are located near the boundary of the fitness region and are marginally compatible with the native structure, in agreement with the results by Taverna and Goldstein [22]. On the other hand, when evolution is dominated by recombination events, the evolutionary preference for the prototype sequence is strong enough so that most native sequences are located near the centre of sequence–structure compatibility.

Aita et al. [29] identified amino acid sequences that fold into a target structure, imposing that the energy of the target must be much lower than that of alternative structures. They found that the neutral networks of different structures are separated by 5–30 mutations in sequence space, with separation increasing with the required threshold stability. Bloom et al. [30] studied the impact of random mutations on the stability of a wild-type structure, and found that the probability that a protein retains its structure declines exponentially with the number of mutations.

4.3 The SCN Model of Evolution

The SCN model is based on the observation that evolution conserves protein structure much more than protein sequence (see e.g. [68,69]). It assumes that all mutations that maintain protein stability above a predefined threshold are selectively neutral, and all other mutations are strongly deleterious, thus resulting in a neutral model. These assumptions are consistent with the observation that many mutations do not significantly modify the activity of a protein and its thermodynamic stability, while mutations that improve substantially protein functionality are rare [60].

4.3.1 Representation of Protein Structures

In the SCN model, the structure of a protein of N residues is represented through an $N \times N$ contact matrix $\mathbf{C}$. This matrix is defined as $C_{ij} = 1$ if sites i and j are in contact, and $C_{ij} = 0$ otherwise. Two sites are considered in contact if any two of their heavy atoms are closer than a given cut-off distance, which we take as 4.5 Å. The effective free energy associated to a sequence of amino acids $\mathbf{A}$ in the configuration $\mathbf{C}$ is, in this type of approach, assumed to have the form of a sum of pairwise contact interactions,

$$E(\mathbf{A}, \mathbf{C}) = \sum_{i<j} C_{ij} U(A_i, A_j) \,, \tag{4.5}$$

where A_i labels one of the 20 amino acid types and $\mathbf{U}$ is a 20×20 symmetric interaction matrix, so that $U(a, b)$ is the interaction energy between amino acids a and b when in contact. A useful choice for the latter is the matrix derived in [82] in such a way to assign high thermodynamic stability to the native states of a large set of monomeric proteins [83].

Three remarks need to be made here: (a) The effective energy parameters take implicitly into account the effect of the solvent and they depend on temperature, thus they express free energies rather than energies. (b) The effective energy of a structure is defined with respect to a completely extended reference structure where no contacts are formed and which sets the zero of the energy scale. (c) The chain entropy sN is not included into the effective energy, as it is constant for constant chain length N.

4.3.2 Stability Against Unfolding

The stability of the native state against unfolding can be estimated from the negative of the native contact energy, $-E(\mathbf{A}, \mathbf{C}^*)$, neglecting changes of conformational entropy with the protein sequence. In the SCN model, we impose that $-E(\mathbf{A}, \mathbf{C}^*)$ is larger than a positive threshold $-E_{\rm thr}$ for sequences $\mathbf{A}$ belonging to the neutral network.

As an alternative measure of stability, one can also use the Z-score of the native energy, $Z(\mathbf{A}, \mathbf{C}^*)$ [81,84], which gives the difference between the energy of sequence $\mathbf{A}$ in configuration $\mathbf{C}^*$ and its average energy in a set of alternative configurations, $\{\mathbf{C}\}$, in units of the standard deviation of the energy

$$Z(\mathbf{A}, \mathbf{C}^*) = \frac{E(\mathbf{A}, \mathbf{C}^*) - \langle E(\mathbf{A}, \mathbf{C}) \rangle_{\{\mathbf{C}\}}}{\sqrt{\langle E(\mathbf{A}, \mathbf{C})^2 \rangle_{\{\mathbf{C}\}} - \langle E(\mathbf{A}, \mathbf{C}) \rangle^2_{\{\mathbf{C}\}}}} \,. \tag{4.6}$$

When a sequence $\mathbf{A}$ folds into a structure $\mathbf{C}^*$, the corresponding Z-score is negative and very large in absolute value. This measure is, however, better suited for estimating the stability against misfolding (see Sect. 4.3.3).

4.3.3 Stability Against Misfolding

For a given sequence $\mathbf{A}$, the energy landscape is defined to be well correlated if all configurations of low energy are very similar to the configuration of minimal effective energy, $\mathbf{C}^*$. Structure similarity is measured by the overlap $q(\mathbf{C}, \mathbf{C}^*)$, which counts the number of contacts that two structures have in common. This number is normalized by the maximal number of contacts, so that q ranges between zero and one. In a well-correlated energy landscape, the inequality

$$\frac{E(\mathbf{A}, \mathbf{C}) - E(\mathbf{A}, \mathbf{C}^*)}{|E(\mathbf{A}, \mathbf{C}^*)|} \geq \alpha(\mathbf{A}) \left(1 - q(\mathbf{C}, \mathbf{C}^*)\right) , \tag{4.7}$$

with a large $\alpha(\mathbf{A})$ holds. This inequality indicates that the energy gap between the ground state $\mathbf{C}^*$ of sequence $\mathbf{A}$ and any alternative structure $\mathbf{C}$, measured in units of the ground state energy, is larger than a quantity $\alpha(\mathbf{A})$ times the structural distance $1 - q(\mathbf{C}, \mathbf{C}^*)$. The dimensionless quantity $\alpha(\mathbf{A})$, which is the largest quantity for which the above inequality holds, can be used to evaluate the folding properties of sequence $\mathbf{A}$. For random sequences, the lowest energy configurations are structurally different and have similar energy, hence $\alpha(\mathbf{A})$ is close to zero. In this case, the energy landscape is rugged, the folding kinetics is very slow, and the thermodynamic stability is low. In contrast, computer simulations of well-designed sequences have shown that, when $\alpha(\mathbf{A})$ is large, the folding kinetics is fast and the stability with respect to changes in the energy parameters as well as mutations in the sequence is very high [16, 31]. In the SCN model, we impose that $\alpha(\mathbf{A})$ is larger than a positive threshold α_{thr} for sequences $\mathbf{A}$ belonging to the neutral network.

Further, it is assumed that the ground state structure $\mathbf{C}^*$ coincides with the target structure defining the neutral network. Indeed, in all the simulations performed using the SCN model, it was never found a sequence whose ground state structure was different than the target one and simultaneously had a sufficiently large energy gap. Therefore, imposing a well-correlated energy landscape through a condition on the normalized energy gap makes it very difficult to change the native structure into a new structure, which is also stable against misfolding. This result agrees qualitatively with the simulations of Aita et al. [29]. It illustrates the difference between RNA and proteins, since it is in contrast with the findings of Schuster et al., who showed that the neutral networks of two different RNA secondary structures can be separated by just one point mutation [10].

4.3.4 Calculation of $\alpha(\mathbf{A})$

Candidate structures for a protein sequence were generated from all possible alignments of the sequence with structures in the PDB. This procedure is called *gapless threading*. To speed up the computation, we considered a non-redundant subset of the PDB in which proteins with homologous sequences

are excluded [85]. About 10^6 alternative structures were obtained for proteins of 100 amino acids, with this number decreasing for longer proteins. The energy function correctly assigns the lowest energy to the native structure for most proteins of known structure, and it generates a well-correlated energy landscape in which structures very different from the native have high energy gaps, so that $\alpha(\mathbf{A})$ is large.

Most of the computer time of these simulations is spent in the calculation of $\alpha(\mathbf{A}')$ for all possible point mutants of the actual sequence $\mathbf{A}$. To speed up this calculation, we note that $\alpha(\mathbf{A})$ is obtained from the configuration $\mathbf{C}$ with the highest destabilizing power, i.e. the highest value of the energy gap divided by the structural distance from the native configuration. This structure changes through evolution, but it is expected that the set of high scoring structures remains the same for neighbouring sequences. Therefore, for each actual sequence, we store a sufficiently large number of configurations with the highest destabilizing powers (typically 50, see [86]), and we compute their destabilizing power in the mutated sequences $\mathbf{A}'$. This procedure may slightly overestimate $\alpha(\mathbf{A}')$, since not all configurations are used, but the fraction of sequences for which $\alpha(\mathbf{A}')$ crosses the acceptance threshold is below 0.1% [86].

One drawback of the computation of $\alpha(\mathbf{A})$ based on gapless threading is that the number of alternative structures generated in this way decreases with the length of the sequence, N. Therefore, the actual value of $\alpha(\mathbf{A})$ is overestimated for longer sequences. This is not a significant problem when, as here, one is interested in comparing values of $\alpha(\mathbf{A})$ for different sequences of the same length. Nevertheless, it can be convenient, in particular for long chains, to evaluate $\alpha(\mathbf{A})$ using a different method [87]. This method estimates the minimal energy for non-native structures through a theoretical prediction based on the random energy model (REM) [88, 89],

$$E_{\mathrm{REM}}(\mathbf{A}) \approx N_{\mathrm{c}} \langle U \rangle_{\mathbf{A}} - \sigma_{U,\mathbf{A}} \sqrt{2 N_{\mathrm{c}} \log(m_N)}\,, \tag{4.8}$$

where N_{c} is the number of native contacts, $\langle U \rangle_{\mathbf{A}}$ and $\sigma_{U,\mathbf{A}}$ are the mean and standard deviation of the interaction energy for all possible contacts, native and non-native ones, within sequence $\mathbf{A}$, and m_N is the number of independent contact matrices for a protein of length N, satisfying physical constraints of hard core repulsion, hydrogen bonding and compactness. The minimal energy estimated in this way, $E_{\mathrm{REM}}(\mathbf{A})$, is in very good agreement with the minimal non-native energy found by threading, $E_{\min}(\mathbf{A}) \approx (1.003 \pm 0.009) E_{\mathrm{REM}}(\mathbf{A}) - (0.0016 \pm 0.0012)$, when m_N is set equal to the number of structures generated through threading, with a correlation coefficient $r = 0.96$ [87]. Using this estimate, one can evaluate the normalized energy gap as

$$\alpha'(\mathbf{A}) = \frac{E(\mathbf{A}, \mathbf{C}^*) - N_{\mathrm{c}} \langle U \rangle_{\mathbf{A}} + \sigma_{U,\mathbf{A}} \sqrt{2 N_{\mathrm{c}} \log(m_N)}}{E(\mathbf{A}, \mathbf{C}^*)(1 - q_0)}\,. \tag{4.9}$$

The number of alternative structures m_N is expected to increase exponentially with chain length as

$$\log(m_N) \approx A\,N + B\,. \tag{4.10}$$

The parameters have been approximately estimated as $A \approx 0.1$ and $B \approx 4$ in such a way that the minimal energy coincides with the one evaluated through threading for short chains ($N < 50$) and the estimated minimal effective energy is higher than the native energy for most proteins in the PDB [87]. Finally, one sets $q_0 = 0.1$ as the typical overlap between unrelated structures, disregarding the length dependence of this quantity.

4.3.5 Sampling the Neutral Networks

The neutral network of a given protein structure is defined as the set of sequences $\mathbf{A}$ for which the stability against both unfolding and misfolding, measured through $E(\mathbf{A}, \mathbf{C}^*)$ and $\alpha(\mathbf{A}, \mathbf{C}^*)$, respectively, exceed predetermined thresholds, chosen as 98.5% of the values of those parameters for the wild-type sequence in the PDB. The threshold chosen enforces conservation of the thermodynamic stability of the native structure $\mathbf{C}^*$. We verified that the qualitative behaviour of the model does not change in the range between 95% and 100% of the values for PDB sequences.

The SCN algorithm [31,32] explores the neutral network of a given protein starting from its PDB sequence $\mathbf{A}_1 = \mathbf{A}_{\rm PDB}$ and iterating the following procedure: At iteration n, (a) the number $X(\mathbf{A}_n)$ of viable neighbours of sequence $\mathbf{A}_n$ is computed, and (b) the sequence $\mathbf{A}_{n+1}$ is extracted randomly among all the viable neighbours of $\mathbf{A}_n$. In this way, we generate a stochastic process that explores the neutral network. This process looses rather quickly the memory of the initial sequence. The total number of viable point mutations, $X(\mathbf{A})$, expresses the local connectivity of the neutral network. This number is normalized by the total number of attempted mutations, $X_{\rm tot}$,[6] thus obtaining the fraction of neutral neighbours, $x(\mathbf{A}) = X(\mathbf{A})/X_{\rm tot} \in (0, 1]$.

4.3.6 Fluctuations and Correlations in the Evolutionary Process

In contrast with the homogeneity assumption of Kimura's neutral model, the SCN model shows that stability constraints produce a broad distribution of the fraction of neutral neighbours $x(\mathbf{A})$. This distribution $P(x)$ is shown in Fig. 4.1

[6] We impose conservation of the starting cysteine residues in the sequence, and do not allow that other residues mutate into cysteine. These requirements are imposed because a mutation that changes the number of cysteine residues by one would leave the protein with a very reactive unpaired cysteine that would most likely affect its functionality and would be therefore rejected with very high probability. The maximum number of attempted mutations is therefore $X_{\rm tot} = 18(N - N_{\rm cys})$, where N is the number of residues and $N_{\rm cys}$ is the number of cysteine residues in the starting sequence.

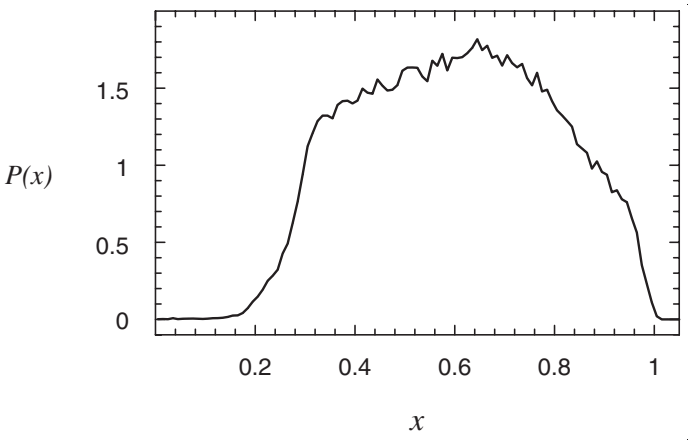

Fig. 4.1. Probability distribution $P(x)$ of the fraction x of neutral neighbours for myoglobin, as obtained by the SCN model (adapted from [90])

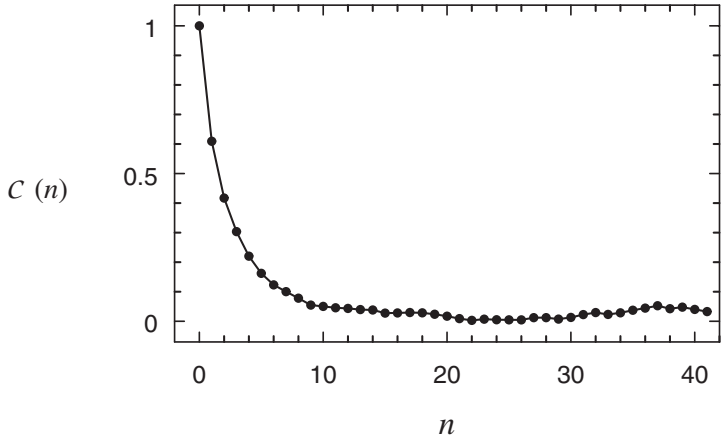

Fig. 4.2. Auto-correlation function $\mathcal{C}(n) \equiv \mathcal{C}(x(\mathbf{A}_k), x(\mathbf{A}_{k+n}))$ of neutral connectivities for sequences separated by n substitutions for myoglobin, as obtained by the SCN model (adapted from [86])

for the neutral network of myoglobin (PDB id. `1a6g`). Other proteins yield qualitatively the same results. Besides this distribution being very broad, the fraction of neutral neighbours is strongly auto-correlated along a trajectory. In Fig. 4.2, we show the auto-correlation function $\mathcal{C}(x(\mathbf{A}_k), x(\mathbf{A}_{k+n}))$ of $x(\mathbf{A}_k)$, defined as

$$\mathcal{C}(x(\mathbf{A}_k), x(\mathbf{A}_{k+n})) = \frac{\dfrac{1}{m-n}\displaystyle\sum_{k=1}^{m-n} x(\mathbf{A}_k)\, x(\mathbf{A}_{k+n}) - \overline{x}^2}{\sigma_x^2}\,, \tag{4.11}$$

where the mean value $\overline{x} = (1/m)\sum_{k=1}^{m} x(\mathbf{A}_k)$ and the variance $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ are calculated over the whole trajectory. Our results show that the auto-correlation decays exponentially as

$$\mathcal{C}(x(\mathbf{A}_k), x(\mathbf{A}_{k+n})) \approx \exp(-n/\ell_x)\,, \tag{4.12}$$

with ℓ_x of the order of three substitutions [90] and, as we shall see, it has important consequences on the statistics of the substitution process.

Broad fluctuations and strong auto-correlations of the neutral connectivity are a general feature of the SCN model, and distinguish it from the standard neutral model by Kimura. They have a rather simple explanation. Defining $x_i(\mathbf{A})$ as the fraction of neutral neighbours when mutation occurs at site i, one has $x(\mathbf{A}) = \sum_{i=1}^{N} x_i(\mathbf{A})/N$. If the fraction of neutral neighbours at different sites are not correlated, their mean $x(\mathbf{A})$ is expected to have fluctuations vanishing as $1/\sqrt{N}$. The broad distribution of $x(\mathbf{A})$ that we found indicates that this is not the case. In fact, there are significant positive correlations between almost all pairs of variables $x_i(\mathbf{A})$ and $x_j(\mathbf{A})$ [90]. These correlations are induced by the fact that the $x_i(\mathbf{A})$ at each site are significantly correlated with some global variable, for instance, the mean fraction of neutral neighbours $x(\mathbf{A})$. This is shown in Fig. 4.3 for the case of myoglobin, defining

$$\mathcal{C}_i = \frac{1}{m}\sum_{k=1}^{m} \frac{\Big(x_i(\mathbf{A}_k) - \overline{x_i}\Big)\Big(x(\mathbf{A}_k) - \overline{x}\Big)}{\sigma_{x_i}\,\sigma_x}\,. \tag{4.13}$$

The figure shows that all the correlations $\mathcal{C}_i$ are positive and significant (they were computed from order of 10^6 sequences, with significance threshold of order 10^{-3}), and moreover, they are positively correlated with the robustness of site i to mutation, measured by $\overline{x_i}$ [90].

Therefore, sequences with large $x(\mathbf{A})$ are more rebust to mutation at all sites. As also found by Bornberg-Bauer for prototype sequences [13], and as we will discuss in next section, these more robust sequences have higher thermodynamic stability, so that mutations applied to them produce more often other stable sequences. Figure 4.3 goes one step further, and shows that there are some sites with small $\overline{x_i}$ that are less tolerant to mutations both in general and in mutationally robust sequences (the correlation between $x_i(\mathbf{A})$ and $x(\mathbf{A})$ is minimal for these sequences). The structural determinants of strong structurally constrained sites will be investigated in the next section.

4.3.7 Substitution Process

Amino acid substitutions within the SCN model are controlled by two independent events: Random mutations, described by a Poissonian process, and an acceptance process, which consists in testing whether the sequence is viable. The acceptance probability for a mutation that takes place in a protein of

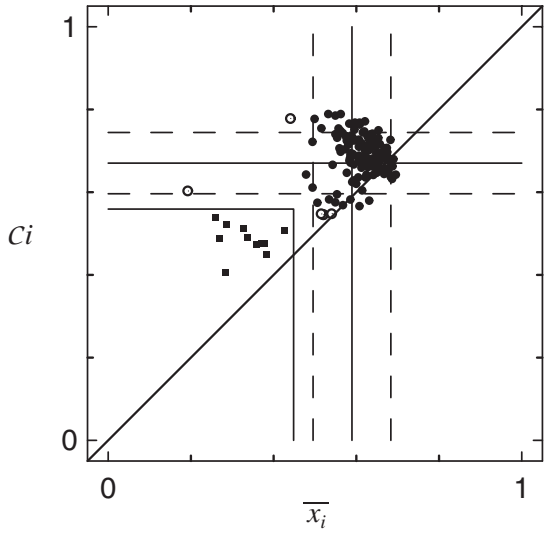

Fig. 4.3. Comparison between cross-correlations and conservations for myoglobin. The fraction of neutral mutations at site i, $\overline{x_i}$, is shown on the abscissa, and the correlation between x_i and the overall neutral connectivity x, $\mathcal{C}_i$, is shown on the ordinate. The *dashed horizontal and vertical lines* indicate one standard deviation from the mean (*full horizontal and vertical lines*). Additionally, *horizontal* and *vertical lines* at the threshold of 1.5 standard deviations below the mean are also shown. The sites above the threshold for both quantities are shown as *full circles*, the sites below the threshold for both quantities are shown as *full squares*, whereas the sites that are above the first threshold but below the second, or vice versa, are shown as *open circles* (adapted from [86])

sequence $\mathbf{A}$ is given by the neutral connectivity $x(\mathbf{A})$. As a result of the broad distribution of this variable, the resulting substitution process is not Poissonian. For a given evolutionary trajectory (i.e. for a given sequence of neutral connectivities $\{x(\mathbf{A}_1), x(\mathbf{A}_2), \ldots\}$) one can compute the probability that the number S_t of accepted mutations in a time interval t equals n. This probability is the product of the Poissonian probability that k mutations take place in the time interval t, times the conditional probability that n of these are accepted,

$$\mathrm{P}\{S_t = n\} = \sum_{m=n}^{\infty} \mathrm{e}^{-\mu t} \frac{(\mu t)^m}{m!} P_{\mathrm{acc}}(n|m)\,, \tag{4.14}$$

where the conditional acceptance probability of n mutations out of m is given by

$$P_{\mathrm{acc}}(n|m) = \left(\prod_{i=1}^{n} x(\mathbf{A}_i) \right) \sum_{\{m_j\}} \prod_{j=1}^{n+1} [1 - x(\mathbf{A}_j)]^{m_j}\ . \tag{4.15}$$

Here, the $\{m_j\}$ are all integer numbers between zero and $m-n$ satisfying $n+\sum_{j=1}^{n+1} m_j = m$. The probability that a mutation is accepted is thus $x(\mathbf{A}_1)$, as long as the protein sequence is $\mathbf{A}_1$, $x(\mathbf{A}_2)$ as long as the sequence is $\mathbf{A}_2$ and so on.

If all sequences have the same fraction of neutral neighbours $x(\mathbf{A}) = x$, (4.14) coincides with (4.4), and the number of substitutions in a branch of length t, S_t, is a Poissonian variable with mean $\mu t x$ and the substitution rate equals μx, as in Kimura's model. If the variance of the neutral connectivity is not zero, the moments of the substitution distribution can be computed in the long-time limit using the central limit theorem. Define τ_i as the time interval between the i-th and $i+1$-th substitutions. The τ_i are independent variables with exponential distribution and expectation values $\mathrm{E}[\tau_i] = \overline{1/\mu x}$, $\mathrm{E}[\tau_i^2] = \overline{2/\mu x^2}$. If S_t is large, we can apply the central limit theorem to the mean value $\sum_{i=1}^{S_t} \tau_i / S_t$, finding

$$\sum_{i=1}^{S_t} \tau_i \approx S_t \overline{\frac{1}{\mu x}} \left[1 + \frac{zB}{\sqrt{S_t}} + \frac{1}{2} \frac{z^2 B^2}{S_t} \right] \approx t \,, \tag{4.16}$$

where z is a normalized Gaussian variable, and

$$B^2 = \frac{\mathrm{E}[\tau_i^2]}{\mathrm{E}^2[\tau_i]} - 1 = \left(1 - \mathrm{e}^{-1/\ell_x}\right)^{-1} \left(\frac{\overline{1/x^2}}{\overline{1/x}^2} - 1 \right) + \frac{\overline{1/x^2}}{\overline{1/x}^2} \,. \tag{4.17}$$

The normalized variance B^2 is larger than one because (a) the distribution of x is broad, so that $\mathrm{E}[1/x^2] > \mathrm{E}^2[1/x]$ and (b) trajectories are correlated (the term $[1-\exp(-1/\ell_x)]^{-1}$ tends to one if the correlation length ℓ_x tends to zero). The first two moments of S_t can be calculated as

$$\mathrm{E}[S_t] \approx \frac{\mu t}{\overline{1/x}} \tag{4.18}$$

$$R(t) \equiv \frac{\mathrm{E}[S_t^2] - \mathrm{E}^2[S_t]}{\mathrm{E}[S_t]} \approx B^2 \left(1 - \frac{3B^2 \overline{1/x}}{4t} \right) . \tag{4.19}$$

The normalized variance $R(t)$ is called the 'dispersion index'. Notice that if the substitution process is Poissonian one has $R \equiv 1$. The asymptotic value of the dispersion index for large time is $R(t \to \infty) = B^2$, which is larger than one due to the broad fluctuations and time correlations of x. Therefore, the substitution process is overdispersed. For small t, when the process probes only one sequence, the substitution process is expected to behave as a Poissonian process with $R(t \to 0) = 1$.

We compared the above predictions to the expectation values calculated from the probability defined in (4.14). The values of the neutral connectivities were obtained from the evolutionary trajectories $\{x(\mathbf{A}_1), x(\mathbf{A}_2), \ldots\}$ simulated with the SCN model (details of the calculation are given in [90]). Averages along an evolutionary trajectory are indicated with angular brackets $\langle \cdot \rangle$,

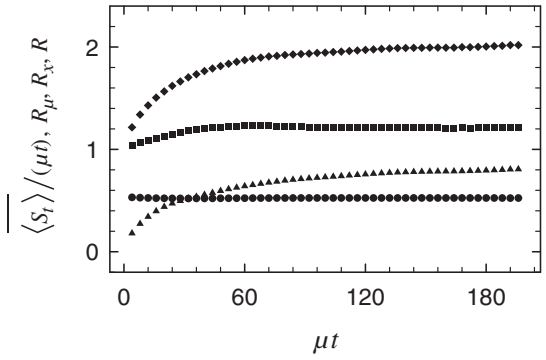

Fig. 4.4. Statistical properties of the substitution process of myoglobin, showing the average number of substitutions $\overline{\langle S_t \rangle}$ divided by μt (*circles*), the normalized mutation variance $R_\mu(t)$ (*squares*), the normalized trajectory variance $R_x(t)$ (*triangles*) and the normalized total variance $R(t) = R_\mu(t) + R_x(t)$ (*diamonds*)

whereas averages over evolutionary trajectories are indicated with an overline $\overline{\cdot}$. The mean and the normalized variance of the number of substitutions are shown in Fig. 4.4 for the case of myoglobin. In the plot, we distinguish the normalized mutation variance

$$R_\mu(t) = \frac{1}{\overline{\langle S_t \rangle}} \left(\overline{\langle S_t^2 \rangle} - \overline{\langle S_t \rangle^2} \right) , \tag{4.20}$$

the normalized trajectory variance

$$R_x(t) = \frac{1}{\overline{\langle S_t \rangle}} \left(\overline{\langle S_t \rangle^2} - \overline{\langle S_t \rangle}^2 \right) , \tag{4.21}$$

and the normalized total variance (the dispersion index) $R(t) = R_\mu(t) + R_x(t)$. Notice that if $x(\mathbf{A}) = x$, one obtains $R_\mu(t) \equiv 1$ as for all Poissonian processes, and the normalized trajectory variance $R_x(t) \equiv 0$. From the plot, it is also clear that most of the overdispersion comes from $R_x(t)$, i.e. from the variance between different evolutionary trajectories, which can generate rather different substitution rates.

The quantitative agreement of the dispersion index $R(t)$ of the SCN process with the prediction (4.19) is quite good as far as the long-time limit is concerned, but the temporal dependence is not well captured by this first-order approximation. The dispersion index of the SCN process is compatible with empirically obtained dispersion indices, which are usually in the range 1.5–5 [9, 50, 91]. Hence, these observed dispersion indices may be to a large extent due to the correlations present in the evolutionary process both in space and in time [90]. This result provides a mechanistic explanation of the fluctuating neutral space model proposed by Takahata to account for the observed statistics of the substitution process [92].

Here, we should notice a difference between the results shown in Fig. 4.4 and those presented in [86]. In [86], we reported that the average substitution rate $\overline{\langle S_t \rangle}/t$ decreases in time in the SCN model, tending to the asymptotic value $\mu\overline{1/x}$. Recently Ho et al. [93], analyzing protein sequences, observed an apparent decrease of the substitution rate through time that would match qualitatively the SCN prediction. However, in obtaining the results presented in [86], we sampled the initial sequences of the evolutionary trajectories with equal probability. This procedure is not entirely consistent, since the time spent at sequence $\mathbf{A}$ is proportional on the average to $1/x(\mathbf{A})$, so that the process spends more time in sequences with small neutral connectivity x. Taking this into account, we have sampled the initial sequence $\mathbf{A}_1$ with probability proportional to $1/x(\mathbf{A}_1)$. The initial rate is therefore $\int_0^1 P(x)\,(1/x)\,(\mu x)\,\mathrm{d}x / \int_0^1 P(x)\,(1/x)\,\mathrm{d}x$, which is equal to the final rate $\mu\overline{1/x}$, so that the rate is now constant in time. Figure 4.4 refers to this new sampling protocol. This does not modify significantly the normalized variances presented in Fig. 10 of [86], which was obtained with homogeneous sampling. Therefore, the results of [86] cannot explain the empirical observations of non-constant rate by Ho et al. [93].

4.4 Site-Specific Amino Acid Distributions

The reconstruction of phylogenetic trees from sequence alignments requires the use of a model of protein evolution [4,94] (see also the chapters by Xia and by Liò et al. in this book). In this context, the effects of both the mutational and the selection processes on protein folding and function must be taken into account. It is well known for instance that the local environment of a protein site within the native structure influences the probability of acceptance of a mutation at that site [95]. Nevertheless, such a view, which is based on structural biology, has a relatively limited impact on studies of phylogenetic reconstruction, where the corresponding models usually rely on substitution matrices that do not consider the structural specificity of different sites. The most used substitution matrices, such as JTT [96], are obtained by extrapolating substitution patterns observed for closely related sequences, and they have low performances when distant homologs are concerned [97].

To account for selection at the protein level, it is necessary to consider site-specific amino acid distributions within a protein family [98]. The use of site-specific substitution matrices improves substantially maximum likelihood methods for reconstructing phylogenetic trees [99–103]. In the studies mentioned above, site-specific constraints are obtained either through simulations of a protein evolution model or by fitting the corresponding parameters within a maximum likelihood framework. As we will discuss in the following, it is possible to deduce from the SCN model an analytical expression for site-specific amino acid distributions with no adjustable parameters. The resulting

distributions are in very good agreement with model simulations and with site-specific amino acid distributions obtained from the PDB [39, 40].

Sites in the same protein evolve in a correlated way, because they undergo global stability constraints. However, Maximum Likelihood approaches become almost computationally unfeasible unless one assumes that sites evolve independently. Here, we will define a mean-field protein evolution model with independent sites that reproduces with great accuracy the results of the SCN model with global stability constraints. The price to pay for this simplification is that we shall consider an effective selection process that depends on the mutation process. At the mean-field level, mutation and selection, that are independent processes in the Darwinian framework, become effectively entangled.

4.4.1 Vectorial Representation of Protein Sequences

The interaction matrix $\mathbf{U}$ in (4.5) can be written in its spectral form as $U(a,b) = \sum_{\alpha=1}^{20} \epsilon_\alpha \, u^{(\alpha)}(a) \, u^{(\alpha)}(b)$, where ϵ_α are the eigenvalues, ranked by their absolute value, and $\mathbf{u}^{(\alpha)}$ are the corresponding eigenvectors. The main contribution to the interaction energy is given by $\epsilon_1 \, u^{(1)}(a) \, u^{(1)}(b)$, which has a correlation coefficient 0.81 with the elements $U(a,b)$ and a negative eigenvalue ϵ_1. It is well known that hydrophobic interactions constitute the most significant contribution to pairwise interactions in proteins, the components of the main eigenvector are strongly correlated with experimental hydropathy scales [104, 105]. By considering only this main component, one can define an effective energy function, $H(\mathbf{A}, \mathbf{C})$, which provides a good approximation to the energy, (4.5), as

$$\frac{H(\mathbf{A}, \mathbf{C})}{k_{\mathrm{B}} T} \equiv \epsilon_1 \sum_{i<j} C_{ij} \, h(A_i) \, h(A_j) \, . \tag{4.22}$$

The vector $\mathbf{h}(\mathbf{A}) \equiv \mathbf{u}^{(1)}(\mathbf{A})$ is denoted as the Hydrophobicity Profile (HP) of sequence $\mathbf{A}$ [38]. This is an N-dimensional vector whose i-th component is given by $h(A_i) \equiv u^{(1)}(A_i)$. The 20 parameters $h(a) \equiv u^{(1)}(a)$, obtained from the PE of the interaction matrix, are called *interactivity* parameters, and are reported in Table 4.1.

Table 4.1. Interactivity scale used in this chapter and presented in [38]

A	R	N	D	C	Q	E	G	H	I
0.1366	0.0363	−0.0345	−0.1233	0.2745	0.0325	−0.0484	−0.0464	0.0549	0.4172
L	K	M	F	P	S	T	W	Y	V
0.4251	−0.0101	0.1747	0.4076	0.0019	−0.0432	0.0589	0.2362	0.3167	0.4083

4.4.2 Vectorial Representation of Protein Folds

A convenient vectorial representation of protein structures may be derived from the PE of the contact matrix $\mathbf{C}$, which we denote as $\mathbf{c}$. The latter maximizes the quadratic form $\sum_{ij} C_{ij}\, c_i\, c_j$ with the condition $\sum_i c_i^2 = 1$. In this sense, c_i can be interpreted as the effective connectivity at site i, since sites with large c_i are in contact with as many as possible sites j with large c_j. All the components of $\mathbf{c}$ have the same sign, which, by convention, is taken as positive. Moreover, if the contact matrix represents a single connected graph, as is the case for single-domain globular proteins, the information contained in the PE is in most cases sufficient to reconstruct the whole contact matrix [37], and consequently the full three-dimensional structure [106].

4.4.3 Relation Between Sequence and Structure

The constraint of thermodynamic stability predicts that there should be a correlation between the vectorial representations of protein sequences and structures.

For a given protein fold, we define the optimal HP, denoted as $\mathbf{h}_{\rm opt}$, as the vector that minimizes the approximate effective free energy, (4.22), under the constraints that its mean hydrophobicity, $\langle h\rangle = N^{-1}\sum_i h_{\rm opt}(A_i)$, and its mean square value, $\langle h^2\rangle = N^{-1}\sum_i h^2_{\rm opt}(A_i)$, are kept fixed.[7] These constraints imply that the mean and standard deviation of non-native interactions is also kept fixed, so that the normalized energy gap, (4.9), is also kept large. From the property of the PE that it maximizes $\sum_{ij} C_{ij}\, c_i\, c_j$ with the condition $\sum_i c_i^2 = 1$, it is clear that $\mathbf{h}_{\rm opt}$ is strongly correlated with $\mathbf{c}$ [38]. In this formulation, $\langle h\rangle$ and $\langle h^2\rangle$ are parameters not determined by the native structure, and they should guarantee a large normalized energy gap (in fact, in the approximation given by (4.22), the mean and mean square contact interactions that enter into the calculation of $\alpha(\mathbf{A})$ by (4.8) are $\langle U\rangle = \epsilon\,\langle h\rangle^2$ and $\langle U^2\rangle = \epsilon^2\,\langle h^2\rangle^2$).

The optimal HP represents an analytical solution to the problem of sequence design for the effective energy function (4.22), and thus an approximate solution for the energy function, (4.5). In the SCN evolutionary model, mutations are accepted whenever the effective free energy and the normalized energy gap overcome predefined thresholds. Thus, the optimal HP is not expected to be ever realized during evolution, since they correspond to a negligible volume in the neutral network. However, thermodynamically stable sequences compatible with the given fold are expected to have HP values not too different from the optimal one. This is indeed observed in simulations of the SCN model. The mean correlation coefficient between the PE of the fold and the HP of the sequences generated through SCN simulations is typically 0.45, which is significant. The HP averaged over all sequences compatible with

[7] Here, we denote by angular brackets the average over all positions in a given protein sequence or structure.

a given fold, $[\mathbf{h}]_{\mathrm{evol}}$, correlates much more strongly with the PE of that fold (and hence with the optimal HP), with a correlation coefficient larger than 0.95 for all of the studied folds [38]. These results show that one can recover the optimal HP through an evolutionary average of the HPs compatible with the protein fold.

Protein families represented in the FSSP [68] and in the PFAM [107] databases show qualitatively similar results. The correlation between the PE of the fold and the HP of individual sequences compares well with what was found in SCN simulations. The average HP over aligned sequences from the same protein family correlates more strongly with the PE than individual HPs: The average correlation coefficient is 0.58 for FSSP families and 0.57 for PFAM families [38]. This correlation is however much weaker than the analogous one for SCN protein families, which is 0.96. There are several explanations for this weaker correlation. First, this can be due to functional conservation, which plays an important role in protein evolution and is not represented in the SCN model. Part of the discrepancy can be also attributed to the approximate character of the effective energy function used to test the thermodynamic stability. Furthermore, real protein families are much smaller than SCN families, for which we generated of the order of 10^6 sequences. To test for such an effect, the average HP has been also computed using only few hundreds of SCN sequences, i.e. of the same order of magnitude as in FSSP or PFAM families. As a result, the correlation between the average HP and the PE was found to be reduced to values comparable to those observed for the FSSP and the PFAM sequence databases [38].

4.4.4 The PE as a Structural Determinant of Evolutionary Conservation

As showed by Bornberg-Bauer [13], thermodynamic stability and mutational stability are correlated. Sequences that are more stable can also bear a larger number of mutations. Bornberg-Bauer called the sequence of maximal mutational stability the *prototype sequence* of a fold, and showed that it has also maximal thermodynamic stability. In our model, the optimally stable sequence can be predicted analytically to have a HP that correlates very strongly with the PE. Sequences close to the optimal one, in the sense that they have a large correlation coefficient $r(\mathbf{h}(\mathbf{A}), \mathbf{c})$, are therefore expected to bear a large number of mutations and to have larger neutral connectivity $x(\mathbf{A})$. We verified this prediction using the SCN families. Although there is a significant correlation between the two quantities, the scattering of the data is very large. Thus in Fig. 4.5, we plot $x(\mathbf{A})$ averaged over protein sequences that have $r(\mathbf{h}(\mathbf{A}), \mathbf{c})$ in the same bin of width 0.02, as obtained for mesophilic rubredoxin (PDB id. `1iro`). Sequences close to the optimal one have a very large fraction of neutral neighbours, as expected.

Thus, the relation $r(\mathbf{h}(\mathbf{A}), \mathbf{c})$ between PE and HP explains a significant part of the sequence variation of the overall mutational stability $x(\mathbf{A})$. As we

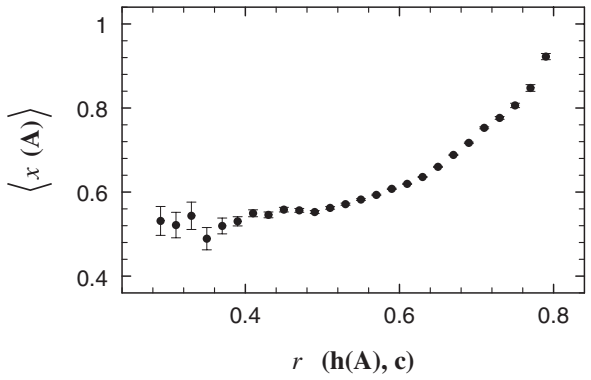

Fig. 4.5. Mean fraction of neutral neighbours $\langle x(\mathbf{A})\rangle$ as a function of the correlation coefficient $r(\mathbf{h}(\mathbf{A}), \mathbf{c})$ between the vectorial representations of sequence $\mathbf{h}(\mathbf{A})$ and structure $\mathbf{c}$ for mesophilic rubredoxin (error bars indicate the standard deviation of the mean)

will see in the next section, the PE explains also a large part of the site-specific variation of mutational stability, with sites having PE components that are smaller or larger than the mean being more conserved through evolution.

4.4.5 Site-Dependent Amino Acid Distributions

The SCN model of protein evolution generates trajectories in sequence space for which the resulting HP fluctuates around the optimal HP, the latter being strongly correlated with the PE of the protein fold's contact matrix [38]. This remarkable feature can be used to compute analytically the site-specific distribution of amino acid occurrences $\pi_i(a)$, where i indicates a protein site and a one of the 20 amino acid types [39].

To derive an analytical expression for $\pi_i(a)$, the correlation coefficient between the PE $\mathbf{c}$ of the native contact matrix and the evolutionary average of the hydrophobicity vector, $\big[\mathbf{h}\big]_{\rm evol}$, is assumed to be exactly 1, yielding that the two vectors are linearly related as

$$\big[h_i\big]_{\rm evol} \equiv \sum_{\{a\}} \pi_i(a)\, h(a) = A\,(c_i/\langle c\rangle - 1) + B\,, \tag{4.23}$$

where the sum over $\{a\}$ is over all amino acids, and

$$A = \sqrt{\frac{\langle \big[h\big]^2_{\rm evol}\rangle - \langle \big[h\big]_{\rm evol}\rangle^2}{\big(\langle c^2\rangle - \langle c\rangle^2\big)/\langle c\rangle^2}} \text{ and } B = \langle \big[h\big]_{\rm evol}\rangle\,. \tag{4.24}$$

In the above equations, two kinds of averages have been introduced: The angular brackets, denoting the average over the N sites of the protein,

$\langle f \rangle = N^{-1} \sum_i f_i$, where the corresponding standard deviation is denoted as $\sigma_f^2 = \langle f^2 \rangle - \langle f \rangle^2$, and the square brackets, denoting site-specific evolutionary averages, $\left[f_i\right]_{\rm evol} = \sum_{\{a\}} \pi_i(a)\, f(a)$.

Equations (4.23) and (4.24) represent the conditions that the stationary distributions $\pi_i(a)$ have to fulfil in order to guarantee a perfect correlation between PE and the average HP. Assuming that these conditions are the only requirement that the $\pi_i(a)$ have to meet, we require that the $\pi_i(a)$ are the distributions of maximum entropy having the given average values. It is well known that the solution of this optimization problem are Boltzmann-like (exponential) distributions, characterized by an effective 'temperature' $|\beta_i|^{-1}$ that, in this context, varies from site to site and measures the tolerance of site i to accept mutations over very long evolutionary times,

$$\pi_i(a) = \frac{\exp[-\beta_i\, h(a)]}{\sum_{\{a'\}} \exp[-\beta_i\, h(a')]}\,, \tag{4.25}$$

with the constraint, (4.23),

$$\sum_{\{a\}} \exp[-\beta_i\, h(a)]\, [h(a) - A\,(c_i/\langle c \rangle - 1) - B] = 0\,. \tag{4.26}$$

Equation (4.26) states an analytical relation between the 'Boltzmann parameter' β_i and the PE component c_i, given the two evolutionary parameters A and B. This equation indicates that β_i equals zero if $c_i/\langle c \rangle = 1 + A^{-1}\big(\sum_{\{a\}} h(a)/20 - \langle \left[h\right]_{\rm evol} \rangle\big)$, and that β_i becomes negative for larger c_i and positive for smaller c_i. The relationship between β_i and c_i is expected to be almost linear in the range $\left|c_i/\langle c \rangle - 1\right| \ll 1$. In addition, β_i tends to minus infinity when the average hydrophobicity at site i, $\left[h_i\right]_{\rm evol}$, tends to the maximally allowed value, and to plus infinity when the average hydrophobicity at site i tends to the minimum allowed value.

Equation (4.26) has a simple qualitative interpretation. Positions with large eigenvector component c_i are buried in the core of the protein structure and are therefore with high probability occupied by hydrophobic amino acids (positive $h(a)$), having a large and negative β_i. Conversely, surface sites with small c_i are more likely occupied by polar amino acids (negative $h(a)$), having large and positive β_i. Intermediate sites are the most tolerant to mutations, having a small $|\beta_i|$ corresponding to high substitution temperature.

The distributions derived here refer to very long evolutionary times, when memory of the starting sequence has been lost. We recall the three assumptions that have been made for deriving the site-specific distributions: (a) The first assumption is that selection on folding stability can be represented effectively as a maximal correlation between the HP of sequences compatible with a given fold and the optimal HP of that fold, the latter nearly coinciding with the PE. This assumption follows directly from an approximation of the effective free energy function with its principal (hydrophobic) component, (4.22). (b) The

second assumption is that the average of the HP of selected sequences over very long evolutionary times has a correlation coefficient of unity with the PE, i.e. all other energetic contributions average out. (c) The third assumption is that this correlation is the only relevant property of the site-specific amino acid distributions, indicating that these distributions are the distributions of maximum entropy whose site-specific averages have correlation one with the PE, thus fulfilling the stability requirement. From these three assumptions, the Boltzmann form of the amino acid distributions follows in a straightforward manner. To compute the site-specific Boltzmann parameters, however, one still needs to determine the positional mean and standard deviation of the site-specific HPs. These quantities depend on the mutation process and the selection parameters. They were computed directly from the data in such a way that the analytical prediction does not contain any free parameter.

The agreement between the predicted site-specific amino acid distributions and those observed in SCN simulations is very good [39], showing that this analytical approach reproduces quantitatively the statistics of the much more complex SCN process.

Boltzmann distributions have a long history in studies of protein structure and evolution. Structural properties of native protein structures, as for instance amino acid contacts, have been assumed to be Boltzmann-distributed [108], and Boltzmann statistics for structural elements was predicted in stable folds of globular proteins [109]. Our work points out to a complementary explanation for such distributions.

Shakhnovich and Gutin [110] proposed a model of sequence design through Monte Carlo optimization, which produced a Boltzmann distribution in sequence space. A mean-field approximation of this model [17, 111] results in site-specific amino acid distributions of the form

$$\pi_i(a) \propto \exp[-\beta\,\phi_i(a)]\,, \tag{4.27}$$

formally similar to (4.25). There are, however, three important differences between the present formulation and (4.27). First, (4.27) was derived as a mean-field approximation to a Boltzmann distribution for entire sequences, whereas (4.25) was derived from the relationship between average hydrophobicity at a given site and the PE component. Second, in (4.27), the Boltzmann parameter β is the same for all sites, whereas β_i, obtained here from the PE, changes along the protein structure. Third and most important, to compute (4.27), aligned families of natural proteins were used in [17, 111], whereas the present computation only requires the PE and two empirical values, the average and the standard deviation of the HP.

In [100, 101], Goldstein and co-workers assumed that the site-specific distributions of physico-chemical amino acid properties have a Boltzmann form. From this assumption they derived a protein evolution model to be used in phylogenetic reconstruction within a maximum likelihood framework. Since the properties that were used in these studies are hydrophobicity and amino acid size, the proposed distributions are a general case of those discussed here.

However, differently from [100, 101], here we classify sites according to the PE component, which is a structural indicator strongly correlated with conservation, and we compute the Boltzmann parameters analytically, whereas in [100, 101] they are fitted using a maximum likelihood framework.

4.4.6 Sequence Conservation and Structure Designability

We have shown that there is a direct relationship between a structural indicator, the PE, and site-specific measures of long-term evolutionary conservation that imposes limits to divergent evolutionary changes. This relationship also provides a link between the topology of a fold and its designability.

One convenient measure of the amino acid conservation at a given site is given by the rigidity, defined in terms of $\pi_i(a)$ as

$$R_i \equiv \sum_{\{a\}} \left[\pi_i(a)\right]^2 = \frac{\sum_{\{a\}} \exp[-2\beta_i\, h(a)]}{\left\{\sum_{\{a\}} \exp[-\beta_i\, h(a)]\right\}^2} \,. \tag{4.28}$$

The value $R_i = 1$ means that the same amino acid is present at site i in all sequences, leading to complete conservation and $\beta_i^{-1} = 0$. In general, the rigidity decreases with increasing temperature $|\beta_i|^{-1}$. One can use (4.26) and (4.28) to compute the rigidity directly from the PE.

A standard information-theoretic measure of site-specific sequence conservation is given by the entropy of the amino acid distribution

$$S_i \equiv -\sum_{\{a\}} \pi_i(a) \log\left[\pi_i(a)\right] = \log\left[Z(\beta_i)\right] + \beta_i \left[h_i\right]_{\rm evol} \,, \tag{4.29}$$

where $Z(\beta_i) \equiv \sum_{\{a\}} \exp[-\beta_i\, h(a)]$. The entropy attains its maximum value, $S_i = \log(20)$, at $\beta_i = 0$, and it decreases with increasing $|\beta_i|$. Predictions of the entropy based on a different approach, (4.27), using aligned protein families have been obtained in [17, 111].

An important property of the entropy is that its exponential, $\exp(S_i)$, provides an estimate of the average number of amino acid types acceptable at site i over very long evolutionary times. Assuming that the amino acid distributions at different sites are independent from each other, the exponential of the sum of all site-specific entropies, $\exp(\sum_i S_i)$, gives an estimate of the region of the sequence space compatible with a given fold. The size of this region represents the designability of the fold. Although the independence assumption is a clear oversimplification, the estimate of designability that can be obtained should be a valuable approximation, and the present approach allows to connect it explicitly to a topological feature of the protein native structure [112, 113].

Kinjo and Nishikawa [114] have recently pointed out the existence of a strong relationship between hydrophobicity and the main eigenvector of substitution matrices derived from protein alignments with various values of the

sequence similarities of the aligned proteins. They considered the eigenvector corresponding to the largest eigenvalue (in absolute value) of the substitution matrices. For high sequence similarities (above 35%), this eigenvector indicates the propensity of the amino acid to mutate over short evolutionary times (mutability). For low sequence identities (below 35%), corresponding to long evolutionary times, this eigenvector is very strongly correlated with hydrophobicity. This correlation is easily understood in the light of the results presented here. In fact, Kinjo and Nishikawa used Henikoffs' method [115] for deriving substitution matrices from observed frequencies of aligned amino acids at sites with various PE values. In the present notation, these substitution matrices can be indicated as $M(a,b) \approx \log\left[\langle \pi_i(a)\,\pi_i(b)\rangle / \langle \pi_i(a)\rangle\langle \pi_i(b)\rangle\right]$, where the angular brackets denote positional average. In other words, these substitution matrices measure the tendency of two residue types a and b to co-occur at the same sites. The relationship between large time substitution matrices and hydrophobicity gives therefore independent support to the results discussed here.

4.4.7 Site-Specific Amino Acid Distributions in the PDB

We tested how the predicted site-specific distributions compare to those obtained from a representative subset of the PDB [39]. For this comparison, we considered a non-redundant subset of single-domain globular proteins in the PDB, with a sequence identity below 25% [85]. Globular folds were selected by imposing that the fraction of contacts per residue was larger than a length-dependent threshold, $N_c/N > 3.5 + 7.8N^{-1/3}$. This functional form represents the scaling of the number of contacts in globular proteins as a function of chain length (the factor $N^{-1/3}$ comes from the surface to volume ratio), and the two parameters are chosen so as to eliminate outliers with respect to the general trend, which represents mainly non-globular structures. Single-domain folds were selected by imposing that the normalized variance of the PE components is smaller than a threshold, $\left(1 - N\langle c\rangle^2\right)/\left(N\langle c\rangle^2\right) < 1.5$. In fact, multi-domain proteins have PE components that are large inside their main domains and small outside them (the PE components would be exactly zero outside the main domains if the domains would not share contacts). Therefore, multi-domain proteins are characterized by a larger normalized variance of PE components with respect to single-domain ones. It has been verified that the threshold of 1.5 is able to eliminate most of the known multi-domain proteins and very few of the known single-domain proteins.

In [39], we selected 404 sequences of less than 200 amino acids, and classified sites according to the value of $c_i/\langle c\rangle$ into bins, where $\langle c\rangle$ denotes the average over a single structure. For each bin, the observed distributions $\pi_{c_i/\langle c\rangle}(a)$ were fitted with an exponential function of the hydrophobicity parameters, $\pi_{c_i/\langle c\rangle}(a) \propto \exp[-\beta_{c_i/\langle c\rangle}\, h(a)]$. As in the case of the SCN simulations, the interactivity scale derived from the effective free energy function, (4.22), was used. The exponential fit was sufficiently good, and yielded the observed

Boltzmann parameters $\beta_{c_i/\langle c\rangle}$ as a function of the normalized PE components, $c_i/\langle c\rangle$.

Next, one can calculate the predicted Boltzmann parameters $\beta_{c_i/\langle c\rangle}$ from the relation

$$c_i/\langle c\rangle = 1 + \tilde{A}^{-1}\left[\frac{\sum_{\{a\}} h(a)\exp[-\beta_{c_i/\langle c\rangle}\, h(a)]}{\sum_{\{a\}}\exp[-\beta_{c_i/\langle c\rangle}\, h(a)]} - \tilde{B}\right], \tag{4.30}$$

where $\tilde{A}$ and $\tilde{B}$ are defined as the analogous terms in (4.24), and the averages indicated by the square brackets in (4.24) now denote, instead of the evolutionary averages over a protein family, the average over all sites with $c_i/\langle c\rangle$ in the same bin, even belonging to different structures, whereas angular brackets in (4.24) now denote the average over all values of $c_i/\langle c\rangle$ weighted with the number of sites in the bins.

The observed Boltzmann parameters are compared in Fig. 4.6 to the predictions of (4.30). The agreement is remarkable, as the predictions do not involve any adjustable parameter, since $\tilde{A}$ and $\tilde{B}$ are calculated from the PDB data [39].

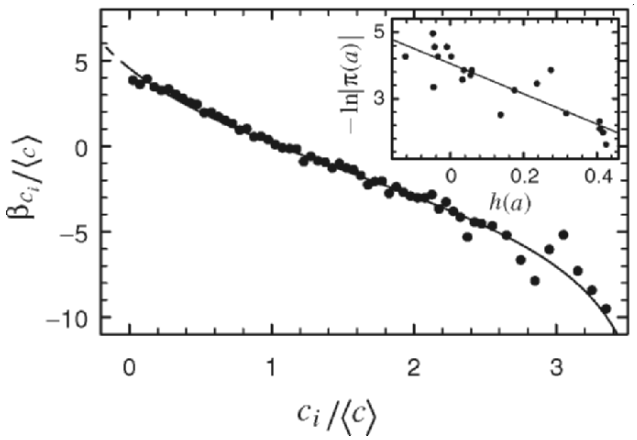

Fig. 4.6. 'Boltzmann parameter' $\beta_{c_i/\langle c\rangle}$ as a function of the normalized PE component $c_i/\langle c\rangle$ (*symbols*) obtained by analysing a subset of 404 non-redundant single-domain globular structures. The *continuous line* shows the analytical prediction, (4.30), obtained using the mean hydrophobicity $\langle [h]_{\mathrm{PDB}}\rangle = 0.128$ and the variance $\langle [h]^2_{\mathrm{PDB}}\rangle - \langle [h]_{\mathrm{PDB}}\rangle^2 = 0.009$ as obtained from this set. The *dashed part* of the curve indicates the forbidden area $c_i < 0$. The *inset* exemplifies the numerically obtained $-\ln[\pi(a)]$ vs. hydrophobicity $h(a)$ of amino acid a (*symbols*), as obtained for $2.45 \le c_i/\langle c\rangle < 2.5$, yielding via a linear fit (shown as *line*) a value of $\beta = -4.53$ for this bin (adapted from [39])

4.4.8 Mean-Field Model of Mutation plus Selection

Despite the good agreement with observations, the predicted distributions do not take into account the mutation process acting at the DNA level, but consider that all mutations from one amino acid to another are equiprobable. To incorporate the DNA level into the SCN scheme, we represent protein evolution at site i as an effective stochastic process with transition matrix

$$T(a,b) = P_\mu(a,b)\, P_{\mathrm{fix},i}(a,b) \tag{4.31}$$

for a substitution from a to $b \neq a$. The first factor represents the mutation process, and it is the same at all positions, and the second one represents the site-specific neutral fixation of mutations that conserve thermodynamic stability. Results from the SCN model show that, for what concerns the stationary distribution, the fixation term can be written as

$$P_{\mathrm{fix},i}(a,b) = \min\left\{1, \exp\left(-\beta_i\left[h(b) - h(a)\right]\right)\right\}\,, \tag{4.32}$$

where the Boltzmann parameter β_i takes the value that fulfils (4.23). The larger the absolute value of β_i is, the larger is the fraction of mutations that are eliminated by negative selection for protein stability and the larger is the mutational load.

The stationary distribution of the complete transition matrix has now the form $\pi(a,\beta) \propto w_\beta(a)\, \exp[-\beta\, h(a)]$, where $w_\beta(a)$ satisfies the equations

$$\begin{aligned} 0 = \sum_{\{b\},b\neq a} & \min\left\{\exp\left[-\beta\, h(b)\right], \exp\left[-\beta\, h(a)\right]\right\} \\ & \times \left[w_\beta(a)\, P_\mu(a,b) - w_\beta(b)\, P_\mu(b,a)\right]\,, \end{aligned} \tag{4.33}$$

for all final amino acid states b. If the mutation matrix satisfies the detailed balance equation, $w(a)\, P_\mu(a,b) = w(b)\, P_\mu(b,a)$, which is called 'reversibility' in the molecular evolution literature, then the stationary distribution of the mutation plus fixation process becomes

$$\pi_i(a) = \frac{w(a) \exp\left[-\beta_i\, h(a)\right])}{\sum_{\{a'\}} w(a') \exp\left[-\beta_i\, h(a')\right]}\,, \tag{4.34}$$

where $w(a)$ is the stationary distribution of the mutation process, which is also the stationary distribution of the protein evolution process at sites where β_i equals zero (no mutations are rejected).

Within this more general context, the case $w(a) \equiv 1$, which corresponds to $P_\mu(a,b) = 1/20$, is the mutational model that was adopted in the previous subsection. Despite its simplicity, it provides already a surprisingly good prediction of the observed amino acid frequencies. If we adopt a reversible mutational model at the nucleotide level, we find

$$w(a) \propto \sum_{\mathrm{codons}(a)} f(n_1)\, f(n_2)\, f(n_3)\,, \tag{4.35}$$

where $f(n)$ is the stationary frequency of the four nucleotides A, T, G and C. Using uniform nucleotide frequencies, $f(n) \equiv 1/4$ (or, in other words, $w(a)$ proportional to the number of codons) improves the prediction by 40% when measuring the similarity using the Jensen-Shannon (JS) divergence [40], without introducing any free parameter. By fitting the nucleotide frequencies, we can further improve significantly the prediction by 30% with only three free parameters [40]. The optimal nucleotide frequencies are $f(\mathrm{T}) = 0.24$, $f(\mathrm{A}) = 0.31$, $f(\mathrm{C}) = 0.19$ and $f(\mathrm{G}) = 0.26$. Notice that the optimal nucleotide frequencies violate Sueoka's parity 2 rule $f(\mathrm{A}) = f(\mathrm{T})$ and $f(\mathrm{C}) = f(\mathrm{G})$ [116], hinting at an asymmetric distribution of coding sequences on the two DNA strands [117].

Note that the site-specific mean hydrophobicity now depends on the parameters of the mutation process, so that

$$\left[h_i\right] \equiv \frac{\sum_{\{a\}} h(a)\, w(a)\, \exp[-\beta_i\, h(a)]}{\sum_{\{a\}} w(a)\, \exp[-\beta_i\, h(a)]} = A\,(c_i/\langle c\rangle - 1) + B\,. \tag{4.36}$$

Therefore, the selection parameters β_i, defined implicitly by the above equation, also depend on the parameters of the mutation process. This looks at first sight in contradiction with the Darwinian paradigm according to which selection and mutation are independent forces. However, the contradiction is only apparent, as shown by the fact that the predicted distributions agree very well with simulations of the SCN model with mutations at the DNA level [117], for which the Darwinian paradigm holds. In the SCN model protein sites evolve in a correlated way as a result of global stability constraints. The effective model presented here is a mean-field model in which sites evolve independently, which constitutes a considerable simplification, in particular with respect to the task of evaluating likelihoods. The price to pay is that the selection parameter has to be computed self-consistently as the result of the mean hydrophobic environment created by other residues, in which the mutation process enters. For instance, when mutations favour the T nucleotide, that in second codon positions mostly codes for hydrophobic amino acids, the β parameter vanishes at hydrophobic positions with large $c_i/\langle c\rangle$, whereas, with the opposite mutation pattern, the β parameter vanishes at hydrophilic positions with small $c_i/\langle c\rangle$. Therefore, mutation and selection, although independent processes at a mechanistic level, become effectively entangled in the mean-field model. Accordingly, the mutation load, i.e. the fraction of mutants eliminated by negative selection, depends on the mutation bias [117], and so do the properties of protein folding thermodynamics: When the bias favours hydrophobic mutations, the balance between stability with respect to unfolding and stability with respect to misfolding shifts towards the former [57, 117]. In this way, the mutation process has a deep influence on the properties of proteins.

4.5 Conclusions

We have described how the conservation of protein structures influences the statistical properties of the evolutionary process by reviewing results that were obtained by using the SCN model. We have given particular emphasis to the effects of structure conservation on the topology of the neutral networks in sequence space and on the correlations during evolutionary trajectories, including the mutual effects on connected structural sites. Additionally, we have explained how the site-specific distributions of amino acids can be derived from the SCN model are consistent with those obtained from an analytically solvable mean-field model.

As illustrated by the results that we discussed, the inclusion of structure conservation in evolutionary models represents a powerful source of insight into the rules that determine molecular evolution. With the advent of structural genomics initiatives and the constant advances in computer technology, the range of applications of this approach is expected to expand considerably in the future.

Acknowledgements

UB would like to thank Peter Grassberger for interesting discussions on the SCN model, and Maya Paczusky for interesting discussions, for the hospitality offered at the Perimeter Institute (Waterloo, Canada) where part of this chapter was written, and for pointing out an inconsistency in the previous treatment of the substitution process.

References

1. E. Zuckerkandl, L. Pauling, in *Horizons in Biochemistry*, ed. by M. Kasha, B. Pullman (Academic Press, New York, 1962), pp. 189–225
2. E. Margoliash, Proc. Natl. Acad. Sci. USA **50**, 672 (1963)
3. D. Graur, W.H. Li, *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, 2000)
4. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford University Press, 2000)
5. L. Bromham, D. Penny, Nat. Rev. Genet. **4**, 216 (2003)
6. M. Kimura, Nature **217**, 624 (1968)
7. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983)
8. J.-L. King, T.H. Jukes, Science **164**, 788 (1969)
9. J.H. Gillespie, *The Causes of Molecular Evolution* (Oxford University Press, Oxford, 1991)
10. P. Schuster, W. Fontana, P.F. Stadler, I.L. Hofacker, Proc. R. Soc. London B **255**, 279 (1994)
11. M.A. Huynen, P.F. Stadler, W. Fontana, Proc. Natl. Acad. Sci. USA **93**, 397 (1996)

12. W. Fontana, P. Schuster, Science **280**, 1451 (1998)
13. E. Bornberg-Bauer, Biophys. J. **73**, 2393 (1997)
14. E. Bornberg-Bauer, H.S. Chan, Proc. Natl. Acad. Sci. USA **96**, 10689 (1999)
15. A. Babajide, I.L. Hofacker, M.J. Sippl, P.F. Stadler, Folding Des. **2**, 261 (1997)
16. A.M. Gutin, V.I. Abkevich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **92**, 1282 (1995)
17. N.V. Dokholyan, E.I. Shakhnovich, J. Mol. Biol. **312**, 289 (2001)
18. L.A. Mirny, E.I. Shakhnovich, J. Mol. Biol. **291**, 177 (1999)
19. N.V. Dokholyan, B. Shakhnovich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 14132 (2002)
20. S. Govindarajan, R.A. Goldstein, Biopolymers **42**, 427 (1997)
21. S. Govindarajan, R.A. Goldstein, Procl. Natl. Acad. Sci. USA **95**, 5545 (1998)
22. D.M. Taverna, R.A. Goldstein, Proteins **46**, 105 (2002)
23. H.J. Bussemaker, D. Thirumalai, J.K. Bhattacharjee, Phys. Rev. Lett. **79**, 3530 (1997)
24. G. Tiana, R.A. Broglia, H.E. Roman, E. Vigezzi, E.I. Shakhnovich, J. Chem. Phys. **108**, 757 (1998)
25. G. Parisi, J. Echave, Mol. Biol. Evol. **18**, 750 (2001)
26. G. Parisi, J. Echave, Gene **345**, 45 (2005)
27. Y. Xia, M. Levitt, Proc. Natl. Acad. Sci. USA **99**, 10382 (2002)
28. Y. Xia, M. Levitt, Curr. Op. Struct. Biol. **14**, 202 (2004)
29. T. Aita, M. Ota, Y. Husimi, J. Theor. Biol. **221**, 599 (2003)
30. J.D. Bloom, J.J. Silberg, C.O. Wilke, D.A. Drummond, C. Adami, F.H. Arnold, Proc. Natl. Acad. Sci. USA **102**, 606 (2005)
31. U. Bastolla, H.E. Roman, M. Vendruscolo, J. Theor. Biol. **200**, 49 (1999)
32. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, J. Mol. Evol. **56**, 243 (2003)
33. R.M. Sweet, D. Eisenberg, J. Mol. Biol. **171**, 479 (1983)
34. L. Holm, C. Sander, Proteins **19**, 256 (1994)
35. N. Kannan, S. Vishveshwara, J. Mol. Biol. **292**, 441 (1999)
36. N. Kannan, S. Vishveshwara, Prot. Eng. **13**, 753 (2000)
37. M. Porto, U. Bastolla, H.E. Roman, M. Vendruscolo, Phys. Rev. Lett. **92**, 218101 (2004)
38. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Proteins **58**, 22 (2005)
39. M. Porto, H.E. Roman, M. Vendruscolo, U. Bastolla, Mol. Biol. Evol. **22**, 630; Mol. Biol. Evol. **22**, 1156 (2005)
40. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Gene **347**, 219 (2005)
41. E. van Nimwegen, J.P. Crutchfield, M. Huynen, Proc. Natl. Acad. Sci. USA **96**, 9716 (1999)
42. M. Eigen, Naturwissenschaften **58**, 465 (1971)
43. M. Eigen, J. Mc Caskill, P. Schuster, Adv. Chem. Phys. **75**, 149 (1989)
44. J.W. Drake, J.J. Hollandy, Proc. Natl. Acad. Sci. USA **96**, 13910 (1999)
45. R. Durrett, *Probability Models for DNA Sequence Evolution*, (Springer, Berlin Heidelberg New York 2002)
46. J. Berg, S. Willmann, M. Lässig, BMC Evol. Biol. **4**, 42 (2004)
47. G. Sella, A.E. Hirsh, Proc. Natl. Acad. Sci. USA **102**, 9541 (2005)
48. I. Leuthauser, J. Stat. Phys. **48**, 343 (1987)
49. P. Tarazona, Phys. Rev. A **45**, 6038 (1992)
50. T. Ohta, M. Kimura, J. Mol. Evol. **1**, 18 (1971)

51. T. Ohta, Nature **246**, 96 (1973)
52. R.A. Fisher, *The Genetic Theory of Natural Selection* (Dover, 1930)
53. J. McDonald, M. Kreitman, Nature **351**, 652 (1991)
54. N.G.C. Smith, A. Eyre-Walker, Nature **415**, 1022 (2002)
55. T. Ohta, J. Mol. Evol. **41**, 115 (1995)
56. D.J. Lambert, N.A. Moran, Proc. Natl. Acad. Sci. USA **95**, 4458 (1998)
57. U. Bastolla, A. Moya, E. Viguera, E. van Ham, J. Mol. Biol. **343**, 1451 (2004)
58. S. Aksoy, Insect Mol. Biol. **4**, 23 (1995)
59. M.A. Fares, M.X. Ruiz-Gonzalez, A. Moya, S.F. Elena, E. Barrio, Nature **417**, 398 (2002)
60. M.C. Orencia, J.S. Yoon, J.E. Ness, W.P. Stemmer, R.C. Stevens, Nat. Struct. Biol. **8**, 238 (2001)
61. C.O. Wilke, BMC Genet. **5**, 25 (2004)
62. A.R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman, 1999)
63. C.M Dobson, Nature **426**, 884 (2003)
64. V.N. Uversky, Cell. Mol. Life Sci. **60**, 1852 (2003)
65. K.A. Bava, M.M. Gromiha, H. Uedaira, K. Kitajima, A. Sarai, Nucl. Ac. Res. **32**, D120 (2004)
66. N. Sueoka, Proc. Natl. Acad. Sci. USA **47**, 469 (1961)
67. J.R. Lobry, Gene **205**, 309 (1997)
68. L. Holm, C. Sander, Science **273**, 595 (1996)
69. B. Rost, Folding Des. **2**, S19 (1997)
70. D. Cozzetto, A. Di Matteo, A. Tramontano, FEBS J. **272**, 881 (2005)
71. S.F. Atschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucl. Acids Res. **25**, 3389 (1997)
72. N. Nagano, C.A. Orengo, J.M. Thornton, J. Mol. Biol. **321**, 741 (2002)
73. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
74. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, Structure **5**, 1093 (1997)
75. C. Chothia, A.M. Lesk, EMBO J. **5**, 823 (1986)
76. D. Devos, A. Valencia, Proteins **41**, 98 (2000)
77. N.V. Grishin, J. Struct. Biol. **134**, 167 (2001)
78. W.R. Taylor, Nature **416**, 657 (2002)
79. A. Harrison, F. Pearl, R. Mott, J. Thornton, C. Orengo, J. Mol. Biol. **323**, 909 (2002)
80. M. Zuker, D. Sankoff, Bull. Math. Biol. **46**, 591 (1984)
81. J.U. Bowie, R. Lüthy, D. Eisenberg, Science **253**, 164 (1991)
82. U. Bastolla, M. Vendruscolo, E.W. Knapp, Proc. Natl. Acad. Sci. USA **97**, 3977 (2000)
83. U. Bastolla, J. Farwer, E.W. Knapp, M. Vendruscolo, Proteins **44**, 79 (2001)
84. R.A. Goldstein, Z.A. Luthey-Schulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992)
85. U. Hobohm, C. Sander, Protein Sci. **3**, 522 (1994)
86. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, J. Mol. Evol. **57**, S103 (2003)
87. U. Bastolla, L. Demetrius, PEDS **18**, 405 (2005)
88. B. Derrida, Phys. Rev. B **24**, 2613 (1981)

89. E.I. Shakhnovich, A.M. Gutin, Biophys. Chem. **34**, 187 (1989)
90. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Phys. Ref. Lett. **89**, 208101 (2002)
91. C.H. Langley, W.M. Fitch, J. Mol. Evol. **3**, 161 (1974)
92. N. Takahata, Genetics **116**, 169 (1987)
93. S.Y.W. Ho, M.J. Phillips, A. Cooper, A.J. Drummond, Mol. Biol. Evol. **22**, 1561 (2005)
94. J. Felsenstein, J. Mol. Evol. **17**, 368 (1981)
95. J. Overington, M.S. Johnson, A. Sali, T.L. Blundell, Proc. Roy. Soc. Lond. B **241**, 132 (1990)
96. D.T. Jones, W.R. Taylor, J.M. Thornton, Comp. Appl. Biosci. **8**, 275 (1992)
97. S. Henikoff, J.G. Henikoff, Proteins **17**, 49 (1993)
98. A.L. Halpern, W.J. Bruno, Mol. Biol. Evol. **15**, 910 (1998)
99. P. Liò, N. Goldman, Genome Res. **8**, 1233 (1998)
100. J.M. Koshi, R.A. Goldstein, Proteins **32**, 289 (1998)
101. J.M. Koshi, D.P. Mindell, R.A. Goldstein, Mol. Biol. Evol. **16**, 173 (1999)
102. J.L. Thorne, Curr. Opin. Genet. Dev. **10**, 602 (2000)
103. M.S. Fornasari, G. Parisi, J. Echave, Mol. Biol. Evol. **19**, 352 (2002)
104. G. Casari, M.J. Sippl, J. Mol. Biol. **224**, 725 (1992)
105. H. Li, C. Tang, N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997)
106. M. Vendruscolo, E. Kussell, E. Domany, Folding Des. **2**, 295 (1997)
107. A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, E.L.L. Sonnhammer, Nucl. Ac. Res. **28**, 263 (2000)
108. S. Miyazawa, R.L. Jernigan, Macromolecules **18**, 534 (1985)
109. A.V. Finkelstein, A.Ya. Badretdinov, A.M. Gutin, Proteins **23**, 142 (1995)
110. E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993)
111. N.V. Dokholyan, L.A. Mirny, E.I. Shakhnovich, Physica A **314**, 600 (2002)
112. P. Koehl, M. Levitt, Proc. Natl. Acad. Sci. USA **99**, 1280 (2002)
113. J.L. England, E.I. Shakhnovich, Phys. Rev. Lett. **90**, 218101 (2003)
114. A.R. Kinjo, K. Nishikawa, Bioinformatics **20**, 2504 (2004)
115. S. Henikoff, J.G. Henikoff, Proc. Natl. Acad. Sci. USA **89**, 10915 (1992)
116. N. Sueoka, J. Mol. Evol. **40**, 318 (1995)
117. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, BMC Evol. Biol. **6**, 43 (2006)