Miguel Arenas, Helena G. Dos Santos, David Posada, Ugo Bastolla, Protein evolution along phylogenetic histories under structurally constrained substitution models, *Bioinformatics*, Volume 29, Issue 23, December 2013, Pages 3020–3028,
https://doi.org/10.1093/bioinformatics/btt530

Original paper

Protein Evolution along Phylogenetic Histories under Structurally Constrained Substitution Models

Miguel Arenas[*], Helena G. Dos Santos[*], David Posada[#], and Ugo Bastolla[*]

[*] Centre for Molecular Biology "Severo Ochoa", Consejo Superior de Investigaciones Científicas (CSIC) and Universidad Autónoma de Madrid, Madrid, Spain.

[#] Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain.

Email addresses:

MA: marenas@cbm.uam.es

HG: hgomes@cbm.uam.es

DP: dposada@uvigo.es

UB: ubastolla@cbm.uam.es

Corresponding author:

Miguel Arenas
Centre for Molecular Biology "Severo Ochoa", CSIC.
Universidad Autónoma de Madrid
C/ Nicolás Cabrera, 1
28049 Cantoblanco, Madrid. Spain
E-mail address: marenas@cbm.uam.es
Phone: +34 911 964 633
Fax: +34 911 964 420

Running head: Structurally constrained protein evolution along phylogenies

Keywords: Molecular evolution, protein structure, energy function, protein stability, phylogenetic history

Abstract

Motivation: Models of molecular evolution aim at describing the evolutionary processes at the molecular level. However, current models rarely incorporate information from protein structure. Conversely, structure-based models of protein evolution have not been commonly applied to simulate sequence evolution in a phylogenetic framework and they often ignore relevant evolutionary processes such as recombination. A simulation evolutionary framework that integrates substitution models that account for protein structure stability should be able to generate more realistic *in silico* evolved proteins for a variety of purposes.

Results: We developed a method to simulate protein evolution that combines models of protein folding stability, such that the fitness depends on the stability of the native state both with respect to unfolding and misfolding, with phylogenetic histories that can be either specified by the user or simulated with the coalescent under complex evolutionary scenarios including recombination, demographics and migration. We have implemented this framework in a computer program called *ProteinEvolver*. Remarkably, comparing these models with empirical amino acid replacement models, we found that the former produce amino acid distributions closer to distributions observed in real protein families, and proteins that are predicted to be more stable. Therefore, we conclude that evolutionary models that consider protein stability and realistic evolutionary histories constitute a better approximation of the real evolutionary process.

Availability: *ProteinEvolver* is written in C, can run in parallel, and is freely available from http://code.google.com/p/proteinevolver/.

Contact: marenas@cbm.uam.es, ubastolla@cbm.uam.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The simulation of molecular evolution is commonly used to mimic real world processes, allowing the study of complex systems that are analytically intractable or to understand the mechanisms by which an evolutionary process is modified (Arenas, 2012; Arenas, 2013; Hoban *et al.*, 2012). A key mechanism in the simulation of molecular sequences is the substitution or replacement process. Markov substitution/replacement models are commonly used in population genetics and molecular evolution to mimic evolutionary processes at the molecular level (see for a review Liò and Goldman, 1998). Nevertheless, most substitution models assume that sites evolve independently, and they cannot incorporate information on the structural and functional role of amino acids within proteins, which is determined by the interactions of different sites. It is known that these interactions may lead to non-independent evolution since the evolutionary rate at a site is influenced by substitutions at neighboring sites (e.g., Berard and Gueguen, 2012 and references therein). Moreover, the conformational diversity of proteins may also influence their molecular evolution (Javier Zea *et al.*, 2013; Juritz *et al.*, 2013). Models of evolution that incorporate structure are therefore of increasing importance (Anisimova and Liberles, 2007; Wilke, 2012). Recently, structurally constrained models of protein evolution have been introduced to represent folding stability of a target structure as a proxy for fitness (reviewed in Liberles *et al.*, 2012). These models have been studied assuming a neutral fitness landscape (e.g., Bastolla *et al.*, 2003; Bastolla *et al.*, 2006; Bastolla *et al.*, 1999; Parisi and Echave, 2001; Rastogi *et al.*, 2006; Taverna and Goldstein, 2002a; Taverna and Goldstein, 2002b) and as a function of population size (e.g., Goldstein, 2011b; Grahnen *et al.*, 2011; Mendez *et al.*, 2010). However, these models have been seldom used to obtain evolutionary insights from real data, probably because of the lack of widely available and easy-to-use software.

Furthermore, proteins simulated under a given substitution process might be unrealistic if common evolutionary processes are ignored. For example, recombination constitutes a fundamental evolutionary force at the molecular level (Posada *et al.*, 2002) that can affect the estimation of different evolutionary parameters, like molecular adaptation (by increasing the number of false positively selected sites) (Anisimova *et al.*, 2003; Arenas and Posada, 2010a; Kosakovsky Pond *et al.*, 2008), substitution rate (lack of molecular clock) (Schierup and Hein, 2000) or ancestral states (Arenas and Posada, 2010b). Similarly, population genetics processes such as demographic changes, population structure and migration may influence evolutionary histories. For example, branches are short when the population size undergoes a bottleneck (e.g., Slatkin, 1996) and deme sizes in structured populations may influence the topology of genealogical trees since lineages are assumed to be in the same deme to coalesce (e.g., Neuhauser and Tavaré, 2001). These aspects could influence the number of substitutions and evolutionary trajectories generated in the simulations, which consequently influence the simulated data (Posada, 2001). In our view, it is important that a simulation framework allows reproducing and testing these evolutionary features in order to be able to address wider biological questions.

In this study, we have implemented structurally-constrained substitution models (hereafter, SCS models) that allow for site-dependent substitutions, under neutral and non-neutral fitness landscapes that depend on protein stability, in the freely available computer program *ProteinEvolver,* which is able to simulate the evolution of proteins and protein-coding genes along evolutionary histories such as phylogenetic trees or ancestral recombinant graphs (ARGs). These phylogenetic histories can be specified by the user or simulated through the coalescent with recombination, demographics and migration (see the reviews, Nordborg, 2000; Wakeley, 2008). Recently, Grahnen and Liberles introduced the computer program *CASS* that simulates protein sequence evolution under selection to fold

into a specific conformation (Grahnen and Liberles, 2012). Our program differs from CASS in two main aspects: the representation of the protein structure, which influences the way in which we treat misfolding stability, and the possibility to represent a broad variety of demographic and evolutionary scenarios (for instance varying recombination rates and mutation). Whereas Grahnen and Liberles adopt all-atom representations of side chains, we adopt contact matrices, which are less precise but allow a statistical mechanical treatment of the ensemble of misfolded conformations that is computationally affordable (and actually very fast with our approximation) (e.g., Bastolla *et al.*, 2005a; Bastolla *et al.*, 2005b). Misfolding stability is important because, if only the unfolded state is considered, selection tends to artificially favor very hydrophobic sequences. Therefore, the structural approach taken by the two programs is complementary, and each of them may be suited to address different kinds of biological questions.

Through extensive simulations we compared the SCS models with commonly used empirical amino-acid substitution models using as a benchmark 10 well-known protein families. We found that sequences simulated under our SCS models produce amino acid distributions closer to the observed ones. Furthermore, the folding stability of the native state, assessed by building structural models by homology and predicting their stability with a method different from the one adopted in the simulations, is significantly larger for proteins simulated under the SCS model. We conclude that substitution models that incorporate protein structure information are better approximations to the real evolutionary process and may provide more meaningful evolutionary inferences than site-independent substitution models.

2 Methods and algorithms

We simulate protein evolution in two main steps. First, the genealogy is either specified by

the user or it is simulated using the coalescent optionally modified with recombination, migration and demographics. Second, protein-coding genes and protein sequences are evolved along this genealogy under a given substitution or replacement model.

2.1 Simulation of genealogies

Genealogies are simulated according to the standard coalescent process (Kingman, 1982) modified with recombination (Hudson, 1983), demographics and migration (Hudson, 1998). Recombination can be either homogeneous or heterogeneous (recombination hotspots and coldspots) across the sequence following the algorithm developed by Wiuf and Posada (2003). Note that the simulation of recombination events leads to reticulate nodes and, consequently, to an ARG (Griffiths and Marjoram, 1997). Demographics include growth rate and demographic periods by following the algorithms implemented in Arenas and Posada (2007; 2010a). Gene flow among subpopulations can be specified under island, stepping-stone or continent-island migration models (e.g., Hudson, 1998). In addition, longitudinal samples or population/species trees, among other capabilities, can also be specified (see Table S1; supplementary material).

2.2 Evolution of proteins along phylogenies

After the phylogenetic history has been specified, a protein structure and a sequence are assigned to the root (the most recent common ancestor -MRCA-, or grand most recent common ancestor -GMRCA- in case of recombination). Then, the protein is evolved along the phylogeny going forward in time from the root to the tips (see, Yang, 2006) according to the SCS model, generating a protein for all internal and terminal nodes in the phylogeny.

Overall, for a given branch the SCS models perform five steps. (i) The number of substitutions is computed considering the branch length (number of expected substitutions

without considering structural constraints) and the length of the protein (number of amino acids). (ii) A mutation is introduced according to the instantaneous rate matrix (the relative rates of change can be used to determine mutational sites). (iii) The folding free energy of the mutated protein structure is computed. (iv) The selective effect of the mutation is evaluated; it will be accepted or rejected depending on the fixation probability associated with the change of fitness (see next section). (v) If the mutation is rejected, the process goes to "(ii)" and a new mutation is introduced. If the mutation is accepted, then it the mutation is fixed and it becomes a substitution. The five-step process is repeated until the number of mutations or substitutions (i) is completed. In this way, fixed mutations (substitutions) result in differences among proteins from different evolutionary lineages.

2.2.1 Substitution models based on the stability of the protein structure

Evaluation of the structural stability of mutated proteins.

We evaluate the folding stability of a given mutated protein taking into account the stability against unfolding and against misfolding. Initially, we estimate the stability of the mutated sequence folded into the target structure at the simulation temperature using a contact-based free energy function. The contact matrix C_{ij} takes the value 1 if residues i and j are 'close' (<4.5Å) in space and 0 otherwise. This matrix has been shown sufficient to accurately reconstruct the three-dimensional structure of the protein (Vendruscolo *et al.*, 1997). We assume that the free energy of a protein with sequence A folded into the contact matrix C is given by the sum of its pairwise contact interactions:

$$E(A,C) = \sum_{ij} C_{ij} U(A_i, A_j) \quad (1)$$

where $U(a,b)$ is the contact interaction matrix that expresses the free energy gained when amino acids a and b are brought into contact determined in (Bastolla *et al.*, 1999). For proteins that fold with two-state thermodynamics, i.e. for which only the native structure

and the unfolded structure are thermodynamically important, stability against unfolding is defined as the free energy difference between the folded and the unfolded states, estimated as $\Delta G \sim E(A,Cnat)+sL$. Here *Cnat* is the native structure, L is the protein length, s is an entropic parameter and sL is the free energy of the unfolded state for proteins with two-states thermodynamics. We use $s = 0.074$, a value that was determined fitting the above equation to a set of 20 experimentally measured unfolding free energies, yielding a correlation coefficient $r = 0.92$. The correlation coefficient between the predicted and the observed stability effect of mutations is larger than 0.8 using only two fit parameters, which is comparable to state-of-the-art atomistic methods such as Fold-X (Guerois *et al.*, 2002).

Stability against unfolding is however not sufficient to characterize protein stability. We also have to check the stability against compact, incorrectly folded conformations of low energy that can act as kinetic traps in the folding process and, in many cases, result in pathological aggregations. Stability against misfolding is achieved by natural proteins by increasing the energy of key contacts that are frequently found in alternative structures, which is termed negative design (Berezovsky *et al.*, 2007; Minning *et al.*, 2013; Noivirt-Brik *et al.*, 2009) to distinguish it from the positive design that favors protein stability by strengthening native interactions. Therefore, stability against misfolding may be influenced by mutations at positions that are distant in the native structure. Stability against misfolded structures is difficult to estimate, and most models of protein evolution do not consider it despite its importance being increasingly recognized (Krishna *et al.*, 2004; Mendez *et al.*, 2010; Zheng *et al.*, 2013). Here we do consider the set of alternative compact matrices of L residues that can be obtained from non-redundant structures in the Protein Data Bank. This procedure, called *threading*, guarantees that the contact matrices fulfill physical constraints on chain connectivity, atomic repulsion, and hydrogen bonding (secondary structure),

which are not enforced in the contact energy function. The free energy of this misfolded ensemble is often estimated with the Random Energy Model [REM, (Derrida, 1981)]:

$$G_{misfold} \approx \langle E(A,C) \rangle - \frac{\sigma^2}{2k_B T} - k_B T s_c L \quad (2)$$

where $\langle E(A,C) \rangle$ is the mean and σ^2 is the variance of the energy of alternative compact structures (Goldstein, 2011a). This formula holds for temperatures above the freezing temperature at which the entropy of the misfolding ensemble vanishes. At lower temperatures the free energy maintains the same frozen value (Derrida, 1981). A recent study showed that the third moment of the energy cannot be neglected (Minning *et al.*, 2013), so that the free energy of the misfolded ensemble can be computed as

$$G_{misfold} \approx \sum_{ij} \langle C_{ij} \rangle U_{ij} - \frac{1}{2k_B T} \sum_{ijkl} \langle C_{ij} C_{kl} - \langle C_{ij} \rangle \langle C_{kl} \rangle \rangle U_{ij} U_{kl} + \frac{1}{6(k_B T)^2} \sum_{ijklmn} \langle (C_{ij} - \langle C_{ij} \rangle)(C_{kl} - \langle C_{kl} \rangle)(C_{mn} - \langle C_{mn} \rangle) \rangle U_{ij} U_{kl} U_{mn} - k_B T s_c L$$

where we denote with $U_{ij}=U(A_i,A_j)$ the contact free energy between residues *i* and *j,* and with $\langle C_{ij} \rangle$ the contact-specific mean value of the contact between the pair of residues *i* and *j* in a large set of compact protein structures of the same length *L* as the target structure.

In the present work, we have reduced considerably the computation time approximating the above free energy (Minning *et al.*, 2013) with one that only depends on pairs of residues,

$$G_{misfold} \approx A^{(1)} \langle U \rangle - \frac{A^{(2)} \langle U \rangle^2 + \sum_{ij} B_{ij}^{(1)} U_{ij}^2}{2k_B T} + \frac{A^{(3)} \langle U \rangle^3 + \langle U \rangle \sum_{ij} B_{ij}^{(2)} U_{ij}^2 + \sum_{ij} B_{ij}^{(3)} U_{ij}^3}{6(k_B T)^2} - k_B T s_c L$$

(3)

The quantities $A^{(1)}\ A^{(2)}\ A^{(3)}\ B_{ij}^{(1)}\ B_{ij}^{(2)}\ B_{ij}^{(3)}$ only depend on the set of alternative contact matrices and on protein length *L*, and they are pre-computed before the simulation starts. In this way, we can evaluate how the misfolded free energy changes upon mutation performing only order *L* operations for computing $U(A_i = b, A_j) - U(A_i = a, A_j)$ when the residue at the mutated site *i* changes from amino-acid *a* to *b*. Thus, the stability of the native state is finally evaluated as the difference in free energy between the native, the

unfolded and the misfolded states, $\Delta G = E(A, C_{nat}) - G_{misfold} - k_B T s_u L$.

The statistical properties of alternative contact matrices are computed from a large set of protein structures, distributed with the *ProteinEvolver* package, that can be modified by the user. Supplementary Figure S5 shows the histogram of the lengths of the alternative structures.

Note that, even if the two configurational entropies per residue s_u (unfolded ensemble) and s_C (misfolded ensemble) act additively, the free energy may not simply depend on their sum, since it is only s_C that determines the freezing temperature of the misfolded ensemble.

Relationship between protein stability and fitness.

Once we have defined protein stability, for modeling protein evolution we still have to define how protein stability influences fitness. Our program provides two alternatives. The simplest possibility is a neutral fitness landscape where the fitness is a binary variable and all proteins with stability above a given threshold, i.e. $\Delta G < \Delta G_{thr}$ are considered viable and equally fit, whereas all proteins below threshold are considered lethal and therefore discarded. We choose as threshold the folding free energy of the protein sequence A_0 in the Protein Data Bank $\Delta G_{thr} = \Delta G(A_0, C_{nat})$. This choice implies that the neutral SCS model is not sensitive to variations of the entropy parameters and it is little sensitive to variations in temperature.

Alternatively, we can consider a non-neutral scenario in which the probability of mutations being fixed depends on population size. In this case, there will be segregating variation in a population**.** Here, the fitness landscape is modelled in such a way that fitness is an increasing function of stability, and in particular it is proportional to the fraction of protein that is in the native state (Goldstein, 2011a),

$$f(A) = \frac{1}{1 + e^{\Delta G(A, C_{nat})/kT}} \quad (4)$$

Note that the fitness landscape can be reduced to the neutral landscape in the low temperature limit, since in this limit the fitness tends to 1 if $\Delta G < 0$and to zero if $\Delta G > 0$. This is a neutral landscape with $\Delta G_{thr} = 0$.

We then assume that the mutation rate is small and we model selection through the Moran's birth-death process (Ewens, 1979), which yields the fixation probability,

$$P(ij) = \frac{1 - \left(\frac{f_i}{f_i}\right)^a}{1 - \left(\frac{f_i}{f_j}\right)^{2N}} \quad (5)$$

where $f_i$is the fitness of the wild-type, $f_j$is the fitness of the mutant, N is the effective population size and a = 2 or 1 for a haploid or diploid population, respectively. Given the probability of fixation, the succession of mutant fixations can be depicted as a Markov process, in which the genotype of the evolving lineage moves from one sequence to another one according to the mutation and fixation probabilities. Both the neutral and non-neutral scenarios are formally equivalent to a Monte Carlo process in statistical mechanics, as discussed by Sella and Hirsch (2005). The main difference between them is that in the neutral case, evolved proteins attain the minimum stability compatible with viability (which in this case is a parameter of the model), as discussed by Taverna and Goldstein (2002b), whereas in the non-neutral scenario stability increases with population size, and it also depends, in a non-trivial way, on the statistical properties of the mutation process (Mendez *et al.*, 2010). Therefore, neutral simulations depend on fewer parameters and they are more robust while non-neutral ones allow to explore more biological questions.

2.2.2 Simulation of the SCS model along an ancestral recombination graph

SCS models are site-dependent, so the simulation across an ARG is not straightforward,

since recombination events put together in the same sequence sites that have been evolving independently along different lineages. In order to evolve the protein as a whole across the ARG, we adapted the algorithm developed by Arenas and Posada (2010a) to evolve codons "broken" by recombination (see Figure 1). The protein evolution occurs from the ancestral to the descendant nodes (see, Yang, 2006). However, if the evolutionary process reaches a recombinant node (Figure 1, nodes in grey), a protein is assigned to such a node (Figure 1, step 3), but at this point the evolutionary process continues along another path (Figure 1, step 4) because its parental recombinant node remains empty (without an assigned protein). This is forced to occur because in the parental recombinant node there is no information about the protein, since the evolutionary process did not reach it yet. Later, the evolutionary process reaches the parental recombinant node (Figure 1, step 5) and, at this point, there are entire proteins assigned to both recombinant nodes. Therefore, now there is a combination of the material according to the recombination breakpoint. This combination results in a new protein (Figure 1, step 6) that continues the evolutionary process along its descendant branch.

3 Fit of the SCP models to real protein families

We studied 10 protein families in order to compare the performances of the integrated SCS models versus the empirical amino acid substitution models. Also, we estimated the temperature and entropic parameters that best reproduce the observed data.

We randomly selected 10 different protein families (Table 1) from the *Pfam* Database (http://pfam.sanger.ac.uk/) subject to two requirements: the *Pfam* seed alignment must possess at least 10 proteins and at least one representative structure included in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). For each protein family we downloaded the

seed alignment and its phylogenetic tree from the *Pfam* database, and chose a representative structure from the PDB. Amino acid positions not included in the protein structure were trimmed from the alignment. The empirical amino acid replacement model that fitted best each alignment was estimated with *ProtTest* (Abascal *et al.*, 2005) (see Table 1).

For each protein family, we simulated 200 realizations of the evolutionary process under the best-fitted empirical amino acid substitution model along the *Pfam* tree. We also performed 200 realizations under the neutral SCS model and the fitness SCS model using the representative PDB structure. For both neutral and fitness SCS models we explored 27 combinations of the thermodynamic parameters temperature (T=1.75, 1.50, 1.30), configurational entropy per unfolded residue (s_u=0.025, 0.05, 0.075) and configurational entropy per misfolded residue (s_c=0.025, 0.05, 0.075), performing a total of 11,000 simulations. For the fitness SCS model we assumed an effective population size N=100 (Cruzan, 2001; Oostermeijer *et al.*, 1994). Note that all substitution models simulated an overall similar number of substitution events (Table S2; supplementary material) because they were applied along the same branch lengths.

3.1 Analysis of amino acid distributions and inference of optimal thermodynamic conditions

We compared the simulated amino acid distributions with those observed in the original data sets measuring their Kullback-Leibler divergence at each site i:

$$d_{KL,i} = \sum_{a=1}^{20} P_i^{obs}(a)\left(\log\left(P_i^{obs}(a)\right) - \log\left(P_i^{sim}(a)\right)\right) \quad (6)$$

where a is any of the 20 amino-acids. We compute the weighted sum $D_{KL} = \sum_{i=1}^{L} w_i d_{KL,i}$ with

weights w_i proportional to the number of aligned residues (excluding gaps) in column *i* of the alignment and summing up to one. The smaller the quantity D_{KL}, the closer the observed and simulated distributions are.

Remarkably, we found that sequences simulated under the neutral SCS model are always closer to the observed distribution than sequences simulated under the empirical substitution model (see Figures 2 and S1; supplementary material). Varying the temperature or the configurational entropy parameters did not affect the divergence of the neutral SCS simulations, as it could be expected since the difference between the estimated ΔG and the neutral threshold ΔG_{thr} is independent of the entropic parameter and it depends only weakly on temperature.

On the other hand, the divergence of the fitness SCS model clearly depended on the particular thermodynamic parameters (see Figures 2 and S2; supplementary material). The average agreement between predicted and observed site-specific amino acid distributions in the fitness SCS model reached a minimum for the combination of entropic and temperature parameters (see Figure S3, minimum in the left plot), where its value was similar to the one produced by the neutral model that used as threshold the free energy of the sequence in the PDB (Figure S3, right plot). These optimal thermodynamic parameters were achieved for $T(s_u+s_C)$=0.16. Interestingly, for some protein families, the fitness SCS model with optimal parameters was significantly less divergent from the observed distribution than the neutral SCS model (see Figure 2). However, for some protein families the fitness SCS model with the 'worst' parameter values had a similar divergence from the observed sequences as the empirical substitution models. These findings indicate that the neutral SCS model is more robust, and suggest that it should be used by default, whereas

the fitness SCS model may be used for refinement with well calibrated thermodynamic parameters, as shown in Figure S3 (left). Comparisons between real sequence families, SCS simulations and simulations based on empirical amino acid substitution models are shown as sequence logos in Figure S6 (supplementary material).

3.2 Structural assessment of simulated proteins respect to the real proteins

We also assessed how much a simulated protein sequence fits a representative protein structure of its family by using homology modeling techniques (Marti-Renom *et al.*, 2000). For each protein family, 200 sequences simulated under the neutral SCS model and under the best-fit empirical site-independent substitution model, were modeled using the *Modeller s*oftware (Eswar *et al.*, 2006; Sali and Blundell, 1993). For each simulated sequence, 20 structural models were generated and they were assessed through their discrete optimized protein energy (DOPE) score (Shen and Sali, 2006), an effective energy function designed for selecting the best model built by *Modeller*. Note that this energetic score is independent from the one used in the SCS models. Then, we selected the sequence-structure pair with the lowest DOPE energy, whose sequence identity with the template is reported in Table S3 (supplementary material). We computed the DOPE energies for the experimentally known sequence-structure pair and for the best structural models of proteins simulated with the neutral SCS model and the best-fit empirical substitution model. Clearly, proteins simulated with the SCS model resulted in better sequence-structure pairs than proteins simulated with the empirical amino acid substitution model (see Figures 3 and S4; supplementary material). This result is not surprising, since we observed that the DOPE score was correlated with the contact energy of the native structure for proteins simulated under the SCS model. However, the two empirical energy functions were derived under different assumptions, therefore the DOPE score may be

regarded as another confirmation of the quality of our models. Of course it is expected that models based on substitution matrices, which do not take into account the structure, produce proteins that are less stable than proteins simulated under SCS conditions. Nevertheless the explicit proof of this expectation is a necessary test to assess the necessity of SCS approaches.

4 Software implementation

We implemented the algorithms described above in the program *ProteinEvolver*. The full list of capabilities of *ProteinEvolver* is shown in the Table S1. *ProteinEvolver* is written in C, it can be run in parallel using MPI and it is freely available under the GNU GPL license from http://code.google.com/p/proteinevolver/. The package includes executables, source code, a detailed documentation and several practical examples (including the files and settings to mimic the evolution of the real proteins described and analyzed in the previous section).

5 Discussion

During protein evolution, interactions within the protein structure lead to correlated evolution, since the rate at which a site experiences change is influenced by replacements at neighboring sites. To adequately model these correlations, we developed the simulation framework *ProteinEvolver* that integrates structure-based models of protein evolution and evolutionary histories that can be simulated under diverse evolutionary scenarios such as recombination (including hotspots and coldspots), migration and demographics. Importantly, our SCS models consider both the stability against unfolding and the stability against misfolding, which is difficult to estimate since it requires the use of a set of alternative conformations, and it is frequently neglected in simulations of protein evolution

despite the importance of negative design. Moreover, our approximation of the free energy of the misfolded state (Minning *et al.*, 2013) allow us to estimate the effect of each mutation on both unfolding and misfolding performing a number of computations that grows only linearly with the number of amino acids. As a consequence, these models can be applied along long phylogenetic histories.

The recently developed *CASS* tool (Grahnen and Liberles, 2012) can simulate protein sequence evolution accounting for selection to fold into a specific conformation. CASS is based on an all-atoms representation of the protein and adopts an atomistic force fields, which can make it more accurate than our contact-matrix based method (although it is known that contact matrices allow to reconstruct all atoms coordinates with precision), but limits its treatment of the misfolded ensemble, which is important to avoid bias towards hydrophobic sequences that are often unrealistically favored by energy functions, although additional considerations in the latter might avoid this effect. An important characteristic of CASS, not included in our program, is that it allows selecting for structures that bind a target molecule, therefore allowing investigating an important aspect of protein function (which is, nevertheless, intimately related with structural stability (e.g., Lukatsky *et al.*, 2007)).

Another important feature of our framework, that is not present in the CASS approach, is that it allows modeling evolutionary mechanisms other than point mutations, such as recombination, which has been shown to be a key element in protein engineering (Carbone and Arnold, 2007) and could affect structural constraints (Archer *et al.*, 2008; Simon-Loriere *et al.*, 2009). For example, Xu *et al.* (2005) have shown that recombination may influence structural divergence. In addition, we can simulate molecular evolution in population genetics scenarios including demographics, population structure and migration, which allows to address a wide range of biological problems by investigating how these evolutionary scenarios influence the properties of the evolved sequences by altering the

underlying genealogies and the properties of the substitution process, which in non-neutral fitness landscapes strongly depend on population size. But also the opposite, how including or ignoring structural considerations can affect population genetic inferences (i.e., inferences on recombination, demographics or migration).

We assessed the performance of the SCS models versus more traditional empirical replacement matrices. For all 10 protein families studied and tested combinations of temperature and configurational entropy parameters, the neutral SCS model always simulated sequences with amino acid frequencies closer to the observed ones than the best-fit empirical amino acid replacement model. On the other hand, the fitness SCS model may outperform the neutral model under optimal thermodynamic parameters, but it is very sensitive to the correct choice of parameters. Consequently, we recommend the use of the neutral SCS substitution model by default since it is more robust, while the fitness SCS model should be only used with the default thermodynamic parameters, in which case it may simulate more realistic proteins for some protein families. This difference stems from the fact that the non-neutral model optimizes protein stability when the population size is large, and it may bias the evolutionary process if the thermodynamic model is not reliable. On the other hand, we think that this dependence on parameters is reassuring, since it shows that the agreement between observed and simulated distributions that *ProteinEvolver* achieves is not a trivial result.

The benefits of using the SCS substitution models instead of the empirical substitution models were also observed by evaluating the adequacy between the simulated sequence and its best homology model with the DOPE energy. In particular, we found that the DOPE energy from proteins simulated by the SCS neutral model was always more negative (more stable three dimensional proteins) than proteins simulated under the empirical substitution

model, although, not surprisingly, less negative than the energy of the experimentally observed sequence-structure pair. These findings were expected because the empirical models consider independent sites and therefore they are unable to account for physical interactions that promote stability (e.g., Pollock *et al.*, 2012; Rodrigue *et al.*, 2005). Nevertheless, they constitute a minimal test that indicates the consistency of SCS models and confirms the limitation of empirical substitution matrices.

The consideration of structural information should result in more sensitive and more accurate representations of molecular evolution than those based on sequence data alone (Wilke, 2012). Our structure-based models could help for a more realistic benchmarking of methods trying to take into account site-dependency induced by protein structure (e.g., Grahnen *et al.*, 2011; Nasrallah *et al.*, 2011) and the important influences of the unfolded and misfolded configurations on the protein stability (Bastolla and Demetrius, 2005; Mendez *et al.*, 2010; Zheng *et al.*, 2013). At the population level, the framework may help, for example: (i) to evaluate the range of proteins that one may expect to observe in different populations (where these populations can change their sizes with time and can exchange migration), (ii) to validate analytical frameworks (for example, methods for the inference of ARGs, ancestral protein reconstruction, recombination breakpoints and recombination rates, from proteins or protein-coding genes while accounting for structural constrains) or even (iii) to infer evolutionary parameters of interest and carry out model choice in an Approximate Bayesian Computation (ABC) approach (Beaumont et al., 2002); for example, estimate recombination rates or select among different demographic and migration models from protein data while accounting for structural information. At the molecular level, the framework may help, for example, to study the influence of recombination events on the structure-based stability of the resulting proteins or to perform structurally constrained substitution model choice by using ABC.

Acknowledgments

We want to thank David Abia for constructive comments and useful suggestions made during this study.

Funding: This work was supported by the Spanish Government with the "Juan de la Cierva" fellowship JCI-2011-10452 to MA and grants BFU2011-24595 and BFU2012-40020 to UB. DP was financially supported by the European Research Council (ERC-2007-Stg 203161-PHYGENOM).

References

Abascal, F.*, et al.* (2005) ProtTest: selection of best-fit models of protein evolution, *Bioinformatics*, **21**, 2104-2105.
Anisimova, M. and Liberles, D.A. (2007) The quest for natural selection in the age of comparative genomics, *Heredity*, **99**, 567-579.
Anisimova, M.*, et al.* (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites, *Genetics*, **164**, 1229-1236.
Archer, J.*, et al.* (2008) Identifying the important HIV-1 recombination breakpoints, *PLoS Comput Biol*, **4**, e1000178.
Arenas, M. (2012) Simulation of Molecular Data under Diverse Evolutionary Scenarios, *PLoS Comput Biol*, **8**, e1002495.
Arenas, M. (2013) Computer programs and methodologies for the simulation of DNA sequence data with recombination, *Front Genet*, **4**, 9.
Arenas, M. and Posada, D. (2007) Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography, *BMC Bioinformatics*, **8**, 458.
Arenas, M. and Posada, D. (2010a) Coalescent simulation of intracodon recombination, *Genetics*, **184**, 429-437.
Arenas, M. and Posada, D. (2010b) The effect of recombination on the reconstruction of ancestral sequences, *Genetics*, **184**, 1133-1139.
Bastolla, U. and Demetrius, L. (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds, *Protein Eng Des Sel*, **18**, 405-415.
Bastolla, U.*, et al.* (2003) Statistical properties of neutral evolution, *J. Mol. Evol.*, **57 Suppl 1**, S103-119.
Bastolla, U.*, et al.* (2005a) Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles, *Gene*, **347**, 219-230.
Bastolla, U.*, et al.* (2005b) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins, *Proteins*, **58**, 22-30.
Bastolla, U.*, et al.* (2006) A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank, *BMC Evol. Biol.*, **6**, 43.

Bastolla, U., *et al.* (1999) Neutral evolution of model proteins: diffusion in sequence space and overdispersion, *J. Theor. Biol.*, **200**, 49-64.
Beaumont, M.A., *et al.* (2002) Approximate Bayesian computation in population genetics, *Genetics*, **162**, 2025-2035.
Berard, J. and Gueguen, L. (2012) Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context, *Syst. Biol.*, **61**, 510-521.
Berezovsky, I.N., *et al.* (2007) Positive and negative design in stability and thermal adaptation of natural proteins, *PLoS Comput Biol*, **3**, e52.
Carbone, M.N. and Arnold, F.H. (2007) Engineering by homologous recombination: exploring sequence and function within a conserved fold, *Curr. Opin. Struct. Biol.*, **17**, 454-459.
Cruzan, M.B. (2001) Population size and fragmentation thresholds for the maintenance of genetic diversity in the herbaceous endemic Scutellaria montana (Lamiaceae), *Evolution*, **55**, 1569-1580.
Derrida, B. (1981) Random Energy Model: An exactly solvable model of disordered systems, *Phys Rev B*, **24**, 2613-2626.
Eswar, N., *et al.* (2006) Comparative protein structure modeling using Modeller, *Curr Protoc Bioinformatics*, **Chapter 5**, Unit 5 6.
Ewens, W.J. (1979) *Mathematical Population Genetics*. Springer-Verlag, Berlin.
Goldstein, R.A. (2011a) The evolution and evolutionary consequences of marginal thermostability in proteins, *Proteins*, **79**, 1396-1407.
Goldstein, R.A. (2011b) The evolution and evolutionary consequences of protein marginal stability, *Proteins*, **In press**.
Grahnen, J.A. and Liberles, D.A. (2012) CASS: Protein sequence simulation with explicit genotype-phenotype mapping, *Trends in Evolutionary Biology*, **4**, 1.
Grahnen, J.A., *et al.* (2011) Biophysical and structural considerations for protein sequence evolution, *BMC Evol. Biol.*, **11**, 361.
Griffiths, R.C. and Marjoram, P. (1997) An ancestral recombination graph. In Donelly, P. and Tavaré, S. (eds), *Progress in population genetics and human evolution*. Springer-Verlag, Berlin, 257-270.
Guerois, R., *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *J. Mol. Biol.*, **320**, 369-387.
Hoban, S., *et al.* (2012) Computer simulations: tools for population and evolutionary genetics, *Nat. Rev. Genet.*, **13**, 110-122.
Hudson, R.R. (1983) Properties of a neutral allele model with intragenic recombination, *Theor. Popul. Biol.*, **23**, 183-201.
Hudson, R.R. (1998) Island models and the coalescent process, *Mol Ecol*, **7**, 413-418.
Javier Zea, D., *et al.* (2013) Protein conformational diversity correlates with evolutionary rate, *Mol. Biol. Evol.*, **30**, 1500-1503.
Juritz, E., *et al.* (2013) Protein conformational diversity modulates sequence divergence, *Mol. Biol. Evol.*, **30**, 79-87.
Kingman, J.F.C. (1982) The coalescent, *Stochastic Processes and their Applications*, **13**, 235-248.
Kosakovsky Pond, S.L., *et al.* (2008) Estimating selection pressures on HIV-1 using phylogenetic likelihood models, *Stat. Med.*, **27**, 4779 - 4789.
Krishna, M.M., *et al.* (2004) Protein misfolding: optional barriers, misfolded intermediates, and pathway heterogeneity, *J. Mol. Biol.*, **343**, 1095-1109.
Liberles, D.A., *et al.* (2012) The interface of protein structure, protein biophysics, and molecular evolution, *Protein Sci.*, **21**, 769-785.
Liò, P. and Goldman, N. (1998) Models of molecular evolution and phylogeny, *Genome Res.*, **8**, 1233-1244.

Lukatsky, D.B., *et al.* (2007) Structural similarity enhances interaction propensity of proteins, *J. Mol. Biol.*, **365**, 1596-1606.
Marti-Renom, M.A., *et al.* (2000) Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291-325.
Mendez, R., *et al.* (2010) Mutation bias favors protein folding stability in the evolution of small populations, *PLoS Comput Biol*, **6**, e1000767.
Minning, J., *et al.* (2013) Detecting selection for negative design in proteins through an improved model of the misfolded state, *Proteins*.
Nasrallah, C.A., *et al.* (2011) Quantifying the impact of dependent evolution among sites in phylogenetic inference, *Syst. Biol.*, **60**, 60-73.
Neuhauser, C. and Tavaré, S. (2001) The coalescent, *Encyclopedia of Genetics*, **I**, 392-397.
Noivirt-Brik, O., *et al.* (2009) Trade-off between positive and negative design of protein stability: from lattice models to real proteins, *PLoS Comput Biol*, **5**, e1000592.
Nordborg, M. (2000) Coalescent Theory, *Review.*
Oostermeijer, J.G.B., *et al.* (1994) Offspring fitness in relation to population size and genetic variation in the rare perennial plant species Gentiana pneumonanthe (Gentianaceae), *Oecologia*, **97**, 289-296.
Parisi, G. and Echave, J. (2001) Structural constraints and emergence of sequence patterns in protein evolution, *Mol. Biol. Evol.*, **18**, 750-756.
Pollock, D.D., *et al.* (2012) Amino acid coevolution induces an evolutionary Stokes shift, *Proc Natl Acad Sci U S A*, **109**, E1352-1359.
Posada, D. (2001) The effect of branch length variation on the selection of models of molecular evolution, *J. Mol. Evol.*, **52**, 434-444.
Posada, D., *et al.* (2002) Recombination in evolutionary genomics, *Annu. Rev. Genet.*, **36**, 75-97.
Rastogi, S., *et al.* (2006) Evaluation of models for the evolution of protein sequences and functions under structural constraint, *Biophys. Chem.*, **124**, 134-144.
Rodrigue, N., *et al.* (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution, *Gene*, **347**, 207-217.
Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.*, **234**, 779-815.
Schierup, M.H. and Hein, J. (2000) Recombination and the molecular clock, *Mol. Biol. Evol.*, **17**, 1578-1579.
Sella, G. and Hirsh, A.E. (2005) The application of statistical physics to evolutionary biology, *Proc Natl Acad Sci U S A*, **102**, 9541-9546.
Shen, M.Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures, *Protein Sci.*, **15**, 2507-2524.
Simon-Loriere, E., *et al.* (2009) Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus, *PLoS Pathog*, **5**, e1000418.
Slatkin, M. (1996) Gene genealogies within mutant allelic classes, *Genetics*, **143**, 579-587.
Taverna, D.M. and Goldstein, R.A. (2002a) Why are proteins marginally stable?, *Proteins*, **46**, 105-109.
Taverna, D.M. and Goldstein, R.A. (2002b) Why are proteins so robust to site mutations?, *J. Mol. Biol.*, **315**, 479-484.
Vendruscolo, M., *et al.* (1997) Recovery of protein structure from contact maps, *Fold Des*, **2**, 295-306.
Wakeley, J. (2008) *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, Colorado.
Wilke, C.O. (2012) Bringing molecules back into molecular evolution, *PLoS Comput Biol*, **8**, e1002572.
Wiuf, C. and Posada, D. (2003) A coalescent model of recombination hotspots, *Genetics*,

164, 407-417.
Xu, Y.O.*, et al.* (2005) Divergence, recombination and retention of functionality during protein evolution, *Hum Genomics*, **2**, 158-167.
Yang, Z. (2006) *Computational Molecular Evolution*. Oxford University Press.
Zheng, W.*, et al.* (2013) Frustration in the energy landscapes of multidomain protein misfolding, *Proc Natl Acad Sci U S A*, **110**, 1680-1685.

Figures

Figure 1. An example of protein evolution along the ARG. White and grey circles correspond to coalescence and recombination parental nodes, respectively. (1) Starting from the GMRCA, the protein is evolved along branches according to the SCS substitution model and the branch lengths. (3) The process encounters a recombinant node and because its parental node has not been assigned to a protein yet, the evolutionary process continues towards other direction (4). (5) Later, the process encounters the parental recombinant node, and because the other parental has already been assigned to a protein, (6) it combines the two proteins according to the recombination breakpoint.

Figure 2. Improvement of the Kullback-Leibler distance to the real protein alignments of the simulated alignments by the neutral and fitness SCS models with respect to the empirical amino acid substitution model. The "y" axis indicates decline of the distance of the neutral and fitness (best and worst conditions) SCS models respect to the distance of the empirical model. Note that the neutral SCS model was overall more robust than the fitness SCS model under different thermodynamic conditions (see Figures S1A and S1B). On the other hand, the fitness SCS model under the best conditions (see Figures S2A, S2B and S3, left plot) could improve the neutral model in half of protein families, however the worst conditions may lead to results without any improvement respect to the empirical model.

Figure 3. DOPE energy computed in the simulated proteins under the empirical and the neutral SCS substitution models and in the native protein, for the protein family "Phototactive yellow proteins". Note that the DOPE energy is unnormalized with respect to the protein size and therefore scores from different proteins cannot be compared directly.

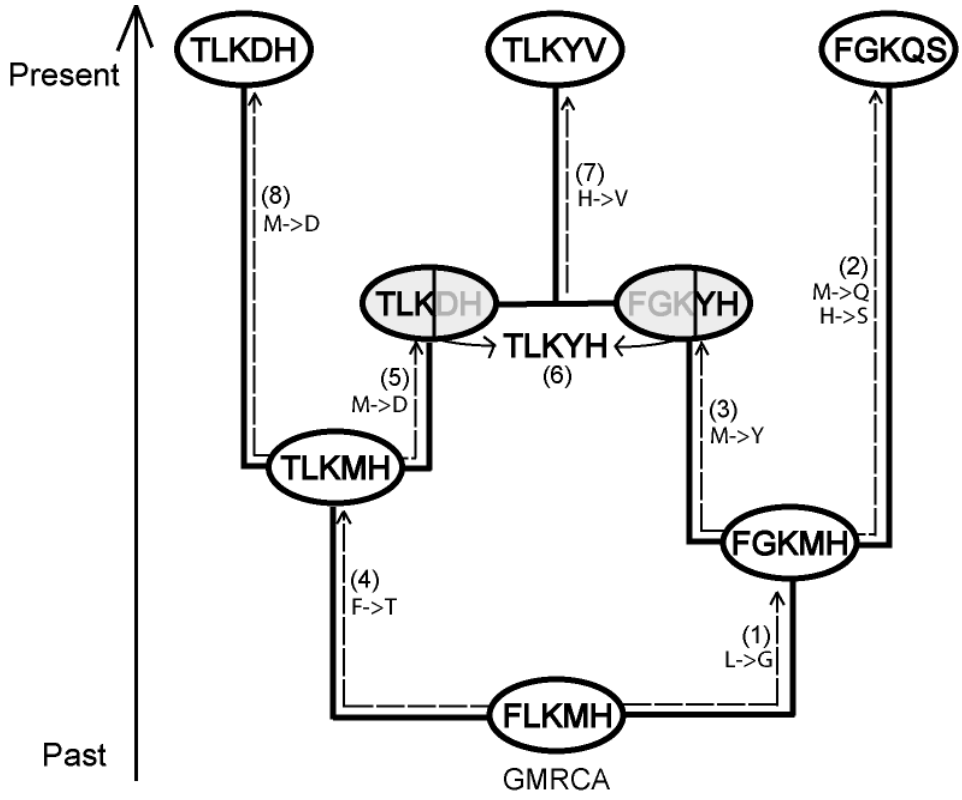
Present
TLKDH
TLKYV
FGKQS
(8)
M->D
(7)
H->V
(2)
M->Q
H->S
TLKDH
FGKYH
TLKYH
(6)
(5)
M->D
(3)
M->Y
TLKMH
FGKMH
(4)
F->T
(1)
L->G
FLKMH
GMRCA
Past

Figure 1.

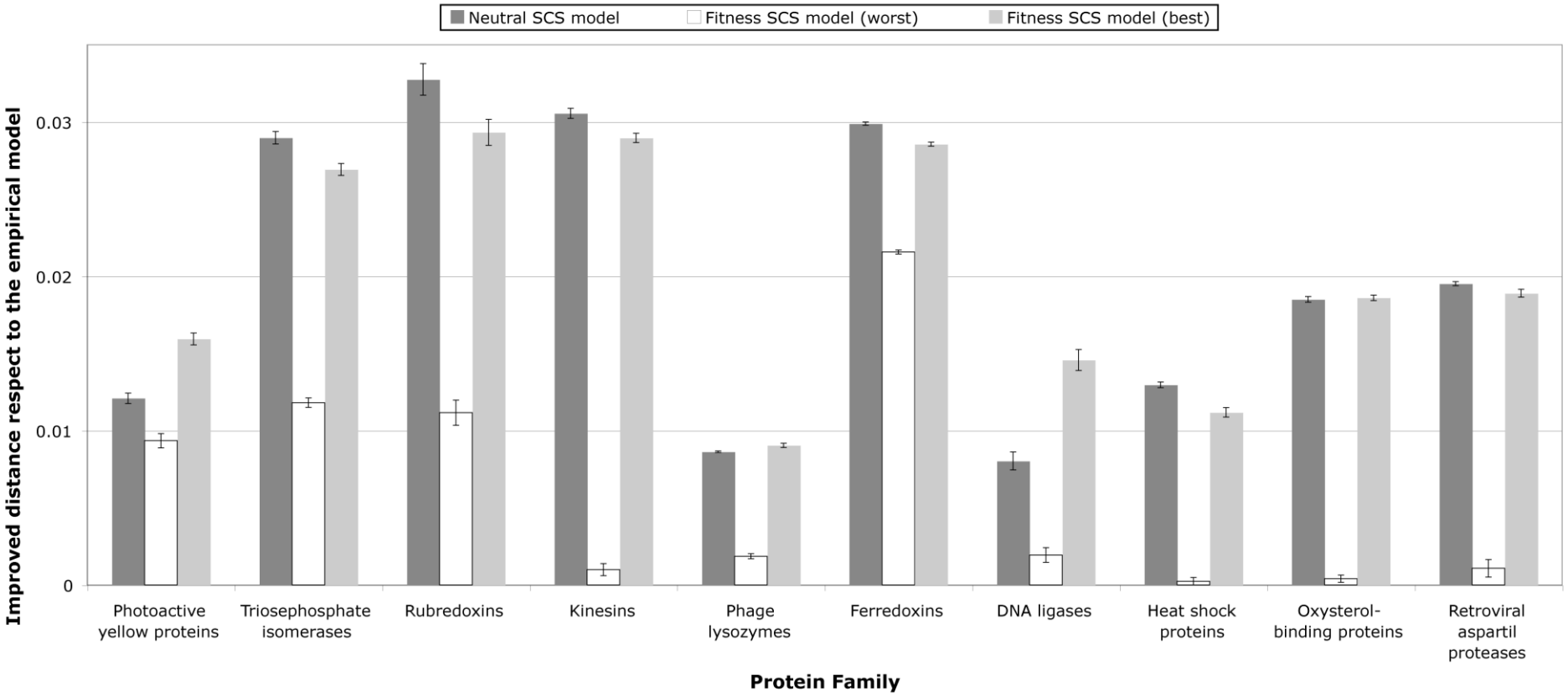
Neutral SCS model
Fitness SCS model (worst)
Fitness SCS model (best)
Improved distance respect to the empirical model
0.03
0.02
0.01
0
Photoactive yellow proteins
Triosephosphate isomerases
Rubredoxins
Kinesins
Phage lysozymes
Ferredoxins
DNA ligases
Heat shock proteins
Oxysterol-binding proteins
Retroviral aspartil proteases
Protein Family

Figure 2.

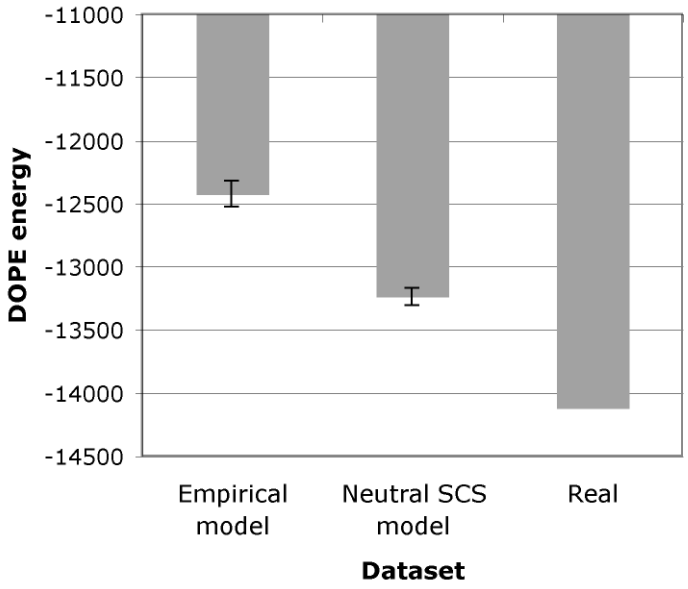
-11000
-11500
-12000
-12500
-13000
-13500
-14000
-14500
DOPE energy
Empirical model
Neutral SCS model
Real
Dataset

Figure 3.

Tables

Table 1. Protein families collected from the *Pfam* database. For each family, the table indicates the *Pfam* code, sample size, *UniProt* entry for a protein sequence with a PDB structure, the PDB code, number of amino acids and the empirical amino acid substitution model that better fits the dataset [+G indicates variable substitution rate across sites according to a gamma distribution, +I indicates a proportion of invariable sites and, +F indicates amino acid frequencies].

Entry	Protein family	Pfam code	Sample size	Uniprot entry	PDB code	Protein length	Best-fit amino acid model
1	Phototactive yellow proteins	PF00989	49	PYP_HALHA	2PHY	125	WAG +G +F
2	Triosephosphate isomerases	PF00121	56	TPIS_TRYBB	1TTI	243	RtREV +I +G +F
3	Rubredoxins	PF00301	43	RUBR2_PSEOL	1R0F	54	WAG +I +G
4	Kinesins	PF00225	87	KAR3_YEAST	3KAR	346	LG +I +G +F
5	Phage lysozymes	PF00959	18	LYS_BPT4	1OV5	164	Blosum62 +G +F
6	Ferredoxins	PF05996	62	PCYA_SYNY3	3NB8	248	WAG +I +G
7	DNA ligases	PF13298	136	B1L4V6_KORCO	3P4H	118	WAG +G
8	Heat shock proteins	PF00012	33	DNAK_ECOLI	2KHO	605	RtREV +G +F
9	Oxysterol-binding proteins	PF01237	153	KES1_YEAST	1ZHT	438	LG +I +G +F
10	Retroviral aspartil proteases	PF00077	50	POL_FIVPE	3OGQ	116	RtREV +I +G