Software for Systematics and Evolution

ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability

Miguel Arenas[1,2,3,4,*], Claudia C. Weber[6], David A. Liberles[5,6], and Ugo Bastolla[3]

[1]Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, Porto, Portugal.

[2]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal.

[3]Centre for Molecular Biology Severo Ochoa (CBMSO), Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain.

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

[5]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

[6]Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA.

Email addresses:

MA: miguelmmmab@gmail.com

CCW: claudia.weber@temple.edu

DAL: daliberles@temple.edu

UB: ubastolla@cbm.csic.es

Corresponding author:

Miguel Arenas

Instituto de Investigação e Inovação em Saúde (i3S)
University of Porto
Rua Alfredo Allen, 208
4200-135 Porto, Portugal
E-mail address: miguelmmmab@gmail.com
Phone: +351 220408800 Ext. 6153

Running head: ASR accounting for structural constraints

Keywords: ancestral sequence reconstruction, protein evolution, molecular adaptation, phylogenetics, folding stability, protein structure

Abstract

The computational reconstruction of ancestral proteins provides information on past biological events and has practical implications for biomedicine and biotechnology. Currently available tools for ancestral sequence reconstruction (ASR) are often based on empirical amino acid substitution models that assume that all sites evolve at the same rate and under the same process. However, this assumption is frequently violated because protein evolution is highly heterogeneous due to different selective constraints among sites. Here, we present *ProtASR*, a new evolutionary framework to infer ancestral protein sequences accounting for selection on protein stability. First, *ProtASR* generates site-specific substitution matrices through the structurally constrained mean-field substitution model (MF), which considers both unfolding and misfolding stability. We previously showed that MF models outperform empirical amino acid substitution models, as well as other structurally constrained substitution models, both in terms of likelihood and correctly inferring amino acid distributions across sites. In the second step, *ProtASR* adapts a well-established maximum-likelihood (ML) ASR procedure to infer ancestral proteins under MF models. A known bias of ML ASR methods is that they tend to overestimate the stability of ancestral proteins by under-estimating the frequency of deleterious mutations. We compared *ProtASR* under MF to two empirical substitution models (JTT and CAT), reconstructing the ancestral sequences of simulated proteins. *ProtASR* yields reconstructed proteins with less biased stabilities, which are significantly closer to those of the simulated proteins. Analysis of extant protein families suggests that folding stability evolves through time across protein families, potentially reflecting neutral fluctuation. Some families exhibit a more constant protein folding stability, while others are more variable. *ProtASR* is freely available from https://github.com/miguelarenas/protasr and includes detailed documentation and ready-to-use examples. It runs in seconds/minutes depending on protein length and alignment size.

INTRODUCTION

The reconstruction of ancestral genes is an intriguing and useful application of evolutionary biology (Chang and Donoghue 2000; Liberles 2007; Merkl and Sterner 2016). Inferred ancestral sequences provide knowledge about the evolution of life and the molecules that sustain it, allowing selection, functional change, or evolutionary paths to be studied. Ancestral sequence reconstruction (ASR) can also be applied to practical problems (Kodra et al. 2007). For example, ancestral sequences have been used to inform HIV vaccine development. Ideal sequences should maintain immunogenic properties while minimizing genetic distances to the descendant circulating target strains (Gao et al. 2003; Doria-Rose et al. 2005; Kothe et al. 2006), which may rely on the accuracy of ASR (Arenas and Posada 2010). Another example is the reconstruction of proteins from extinct organisms, such as enzymes with a higher thermodynamic stability than extant enzymes (Gaucher et al. 2008; Perez-Jimenez et al. 2011; Hobbs et al. 2012) that can be used for industrial processes (Thomson et al. 2005; Yamashiro et al. 2010; Alcalde 2015). In order to be useful for scientific inference as well as for practical applications, ASR methodologies must be unbiased and obtain ancestral sequences with realistic properties.

Most of the available software to perform ASR on proteins is based on a single empirical amino acid exchangeability matrix that is applied to all protein sites and does not consider protein folding stability (Kosakovsky Pond et al. 2005; Yang 2007; Ashkenazy et al. 2012). Further, independence between sites is commonly assumed in order to obtain the computationally tractable ML functions most currently available methods require. However, it is well established that considering structural constraints yields more realistic substitution models and evolutionary inferences (Govindarajan and Goldstein 1997; Bastolla et al. 1999; Parisi and

Echave 2001; Taverna and Goldstein 2002; DePristo et al. 2005; Bastolla et al. 2006; Bloom et al. 2006; Goldstein 2011; Grahnen et al. 2011; Liberles et al. 2012; Wilke 2012; Arenas et al. 2013; Huang et al. 2014; Arenas 2015; Arenas et al. 2015; Chi and Liberles 2016; Echave et al. 2016; Bastolla et al. 2017) since thermodynamic stability is an important source of selective constraint (intrinsically disordered proteins aside). Unfortunately, structurally constrained models of protein evolution are not yet well-established in the phylogenetic pipeline, mainly due to the complexity of incorporating site-dependence in ML functions. Of course, it would be more realistic to also incorporate selection on protein function in light of evidence from experimental studies that suggests relevant factors such as binding (Kachroo et al. 2015). However, this requires additional knowledge about the protein family, *ad-hoc* assumptions about the constraints on the functionally important sites, and how they may change under functional selection. Compared to structural constraints, it is challenging to formulate general rules about functional constraints beyond inter-molecular protein binding.

In order to capture structural constraints while retaining the computational simplicity of the independent sites models, we recently proposed a mean-field (MF) substitution model (Arenas et al. 2015) with constraints on the stability of the native state against both unfolding and misfolding (Minning et al. 2013). We have shown that accounting for stability against both unfolding and misfolding states prevents the generation of unrealistically high or low hydrophobicity (Arenas et al. 2015). The MF model is computed as the site-specific distribution with independent sites that is closest to a site-nonspecific background distribution (interpreted as arising from mutations alone), and that constrains the average stability of the native state. The Lagrange multiplier that imposes this constraint is interpreted as the strength of selection on folding stability. It is the only free parameter of the model, and is optimized by

ML. The MF model generates site-specific amino acid replacement matrices that can be incorporated into phylogenetic methods. Comparisons based on both the likelihood corrected through the Akaike Information Criterion (AIC) and amino acid distributions across sites, showed that MF models outperform empirical amino acid substitution models as well as other structurally constrained substitution models for all of the protein families analyzed (Arenas et al. 2015).

Here, we study the performance of MF for reconstructing ancestral proteins accounting for folding stability, a challenge that may be influenced by the MF modelling of selection on stability. We developed a user-friendly program called *ProtASR* to perform ASR under MF models. We applied *ProtASR* to sequences simulated under site-dependent models of protein evolution that consider structural constraints, and compared the reconstructed sequences to those obtained with site-homogeneous models. We found that proteins reconstructed with MF models are less biased towards higher stability and closer to the folding stability of the simulated proteins. We applied the new framework to reconstruct the history of the folding stability of Prokaryotic protein families analyzed in a previous study (Bastolla et al. 2004) and observed considerable variability in the evolution of thermodynamic properties through time.

NEW APPROACHES: PROTASR

The program *ProtASR* performs two main steps, the computation of the average and site-specific replacement matrices with a MF model and their incorporation into an ML ASR method that we adapted to operate with site-specific matrices.

(1) In the first step, using the MF model the program computes the site-specific amino acid frequencies that have minimal Kullback-Leibler divergence from background frequencies subject to constraint on the stability against unfolding and misfolding. The selection parameter that imposes this constraint and the background frequencies are fitted through ML, and site-specific substitution rates are obtained by applying a global exchangeability matrix (Arenas et al. 2015).

(2) These site-specific substitution matrices and the corresponding global matrix are incorporated into a modified version of the program *PAML* (Yang 2007), which allows both *marginal* and *joint* ML ASR. In the first step the global substitution matrix is applied to optimize the branch lengths for all sites. In the second step, ASR is performed for each site by considering the branch lengths obtained in the first step. To be able to meaningfully perform these computations, *PAML* was modified to circumvent the step that internally normalizes the rate matrix and sets the average rate to one.

The *ProtASR* user inputs a multiple alignment of protein sequences, a rooted phylogenetic tree, a PDB file with a protein structure representative of the alignment (see below) and a set of parameters to define the MF model. These include the environmental temperature, the configurational entropies per residue for the unfolded and misfolded states, the source of the background amino-acid frequencies (user-specified, derived from the protein structure or derived from the alignment) and the exchangeability matrix needed to compute the substitution rates, which may either correspond to an empirical substitution model or be internally computed from evolutionary parameters at the nucleotide level (e.g., nucleotide frequencies and transition/transversion rate ratio). Detailed information and recommendations about the input parameters are provided in the software documentation. Computation is efficient and times range from seconds to minutes depending on protein length and number of sequences (see Table 1).

TABLE 1. Protein families studied. For each protein family, the table indicates *Pfam* code, gene, *UniProt* entry for a protein sequence with a PDB structure, PDB code, protein length, alignment size (number of leaves), sequence identity and the time taken by *ProtASR* to perform the ASR under the MF model on an Intel® Core® i7 CPU 2.5GHz processor.

Entry	Protein Family	Gene	Pfam code	Uniprot code	PDB code	Protein length	Sample size	Seq Id (%)	Computing Time (s)
1	*D*-ala *D*-ala ligases	DDL	PF07478	DDLB_ECOLI	1IOV	399	42	39.7	67.3
2	Chaperone proteins dnaK	DNAK	PF00012	DNAK_ ECOLI	1DKZ	251	38	58.9	35.7
3	Triosephosphate isomerases	TPIS	PF00121	TPIS_ECOLI	1TRE	276	32	43.4	42.8
4	Tryptophan synthases α chain	TRPA	PF00290	TRPA_SALTY	1A50	276	25	47.4	38.5
5	Thioredoxins I	TRXB	PF00070	TRXB_ECOLI	1TDE	375	28	46.4	53.4
6	SH2 domain	SH2	PF00017		1D4T	104	10	69.8	9.3

ProtASR assumes that the input protein structure is representative of the proteins included in the alignment and therefore, protein sequences should fold into structures. This is a reasonable assumption since protein structures are typically conserved over the range of protein sequence divergence in a gene family (Illergard et al. 2009; Pascual-Garcia et al. 2009). To simplify computations and reduce potential artefacts from calculated structures that are not protein-like, *ProtASR* assumes perfect conservation of the protein structure through the evolutionary history of the analyzed protein family. Additionally, one sequence in the input alignment must correspond to the sequence of the input PDB structure (or alternatively, the input alignment and the sequence of the input PDB file must contain an equal number of sites that are homologous without gaps) to allow unambiguous alignment between the structure and sequences.

EVALUATING PROTASR

We have previously shown that MF models yield a higher likelihood and more realistic site-specific amino acid distributions than empirical substitution models and other structurally constrained models (Arenas et al. 2015). Here, in order to evaluate the application of MF models to ASR, we assessed the stabilities of ancestral proteins reconstructed under MF and empirical substitution models, and compared them to those of simulated ancestral proteins.

Evaluation with data simulated under the structurally constrained model of protein evolution adopted in ProteinEvolver

As a first benchmark we analyzed the following Prokaryotic protein families: DDL, DNAK, TPIS, TRPA and TRXB (Table 1). Each family consists of a putative group of homologs with extant sequences longer than 200 amino acids with members in many bacterial species (Bastolla et al. 2004), allowing well-supported phylogenies to be generated. The datasets were downloaded from the Pfam database, realigned with *MAFFT* (Katoh and Standley 2013) and ML phylogenetic trees were reconstructed under the JTT substitution model (Jones et al. 1992). The trees were rooted with an Eukaryotic protein (or an Eukaryotic group) as outgroup. Next, for each family we chose one representative protein with a known PDB structure as the root sequence and evolved it along the inferred phylogeny 50 times with *ProteinEvolver* (Arenas et al. 2013). *ProteinEvolver* employs a similar energy function with structural constraints as MF, but it is more realistic because it implements a model with site-dependent constraints, while MF assumes that sites evolve independently to allow its incorporation into likelihood functions. We ran *ProteinEvolver* under a site-dependent model with standard parameters suggested in Arenas *et* al. (2013). From each simulation we obtained sequences for all internal and tip nodes. We then used the multiple sequence alignment (MSA) of the tip nodes to perform ASR under the empirical JTT model and under the MF model with the exchangeability matrix determined by the same JTT rate matrix. Hence, the structural constraints captured in MF are the only difference between the two models. As an additional comparison, we performed ASR under the CAT model implemented in *PhyloBayes* (Lartillot et al. 2009), which estimates the exchange rates and amino-acid equilibrium frequency vectors from the data (Lartillot and Philippe 2004). Due to the computational cost of these calculations, we considered only the TPIS and TRPA protein families, which had fewer sequences to consider (technical details about ASR with *PhyloBayes* are described in Appendix I of the supplementary material).

Subsequently, we estimated the folding free energy of all inferred ancestral sequences by using the stability model implemented in *ProteinEvolver* (Minning et al. 2013), which considers the free energy difference between the native state and both the unfolded and misfolded states. In these computations, for each sequence of the MSA the native state is identified as the structure with the lowest contact free energy among a large number of structures available in the PDB for the studied protein family. Considering multiple structures is particularly important when analyzing real protein families in order to reduce the bias to assign a lower free energy to sequences closer to the sequence of the representative protein. The free energy of the misfolded state is estimated through a Random Energy Model (REM) based on the mean and the variance of the contact energy of generic compact contact matrices and on their estimated configurational entropy (Minning et al. 2013) and the free energy of the unfolded state is estimated through its configurational entropy. Finally, we calculated the bias (signed difference between average values) and the Root Mean Square Error (RMSE) of the folding free energies estimated for the simulated sequences and the inferred sequences derived from both MF and empirical models.

For all protein families, ancestral sequences generated through the MF model showed free energies significantly closer to those of the simulated sequences (that is, smaller RMSE) than ancestral sequences generated through the empirical model (Figs. 1, 2, S1 and S2, supplementary material). The improvement of MF models was heterogeneous with respect to the distance from the reconstructed node to the root (Figs. 1 and S1). This is consistent with the expectation that, due to the influence of the substitution model, error increases with evolutionary distance from extant sequences at tip nodes (e.g., Williams et al. 2006). The error is largest at the root. MF consistently significantly outperformed the empirical model in terms of reconstructing the stability at the root (Figs. 2 and S2; Wilcoxon signed-rank test for error p < 10e-5). Since the root is the sequence of the PDB structure, while other sequences are the result of simulations, this is an important test that assesses the stability of real protein sequences. In addition, MF also significantly improved the reconstruction of the stability of ancestral proteins compared to the CAT model (Fig. S3, supplementary material; Wilcoxon signed-rank test p < 5.9e-47).

In general, our reconstructed sequences were more stable than simulated or real sequences (Figs. S1 and S2), a bias also observed in previous analyses (Williams et al. 2006; Goldstein

2011). Importantly, the MF model reduces this bias when compared to sequences reconstructed with empirical models (Figs. S1 and S2). While this model explicitly considers the thermodynamic effects of a substitution, potentially generating more neutral behavior for destabilizing changes in an already stable protein, it still lacks the segregating deleterious changes that would be expected to be sampled in any sequence at the tips (or along the tree).

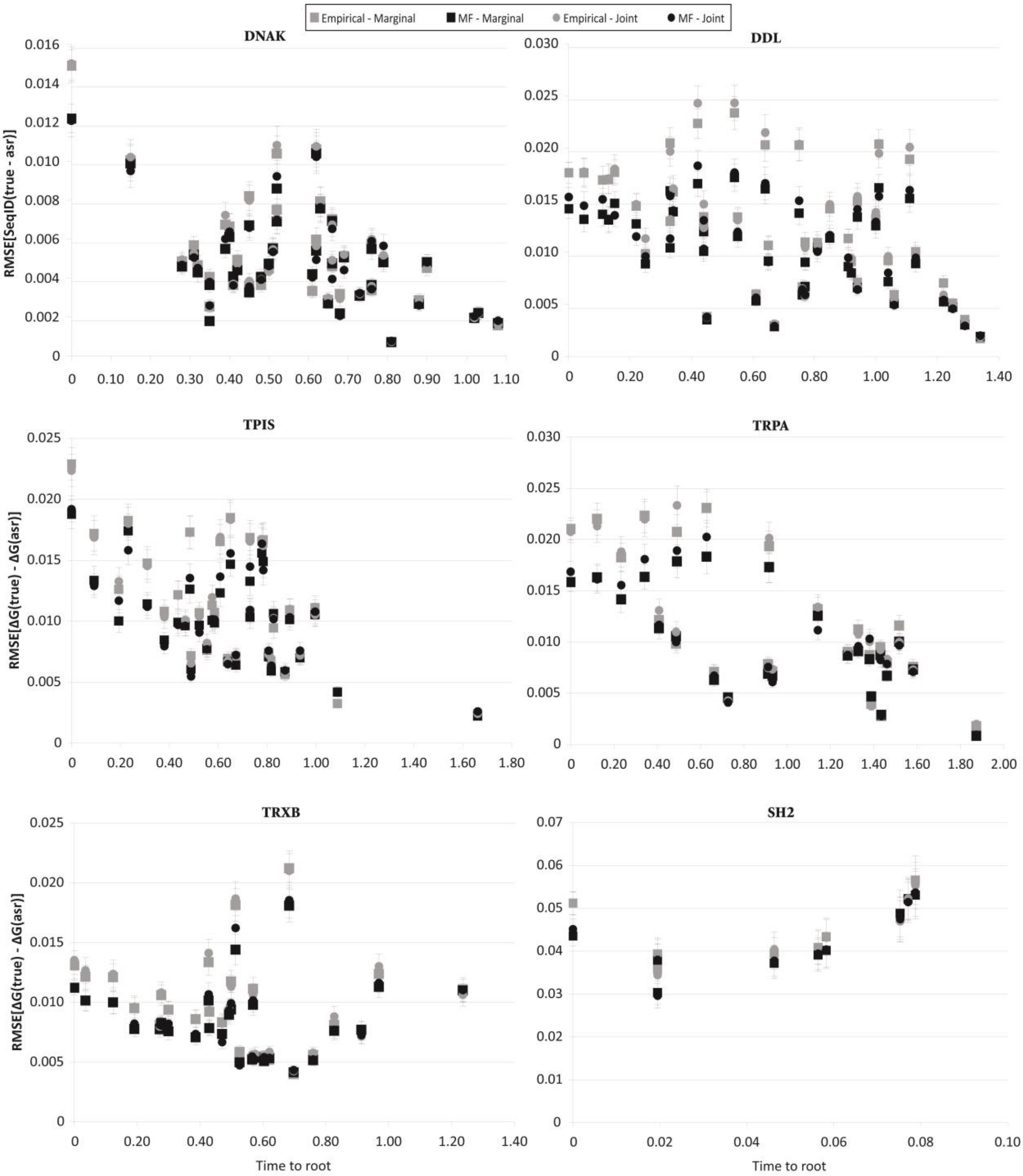

FIGURE 1. RMSE of the computed folding free energy between simulated and reconstructed ancestral sequences under MF and empirical substitution models. Each point represents a sequence, and the *x*-axis represents the evolutionary distance from the root. Both *joint* and *marginal* reconstructions are shown. Note that MF (black squares and circles) frequently generates ancestral proteins with energies closer (lower RMSE) to the simulated

proteins, with respect to the empirical model (grey squares and circles), although this effect is variable among nodes. As expected, the RMSE tends to increase at larger distance from the tip nodes. Note also the small differences between *joint* and *marginal* ASR, which are not significant. Error bars indicate standard error of the mean over 50 computer simulations.

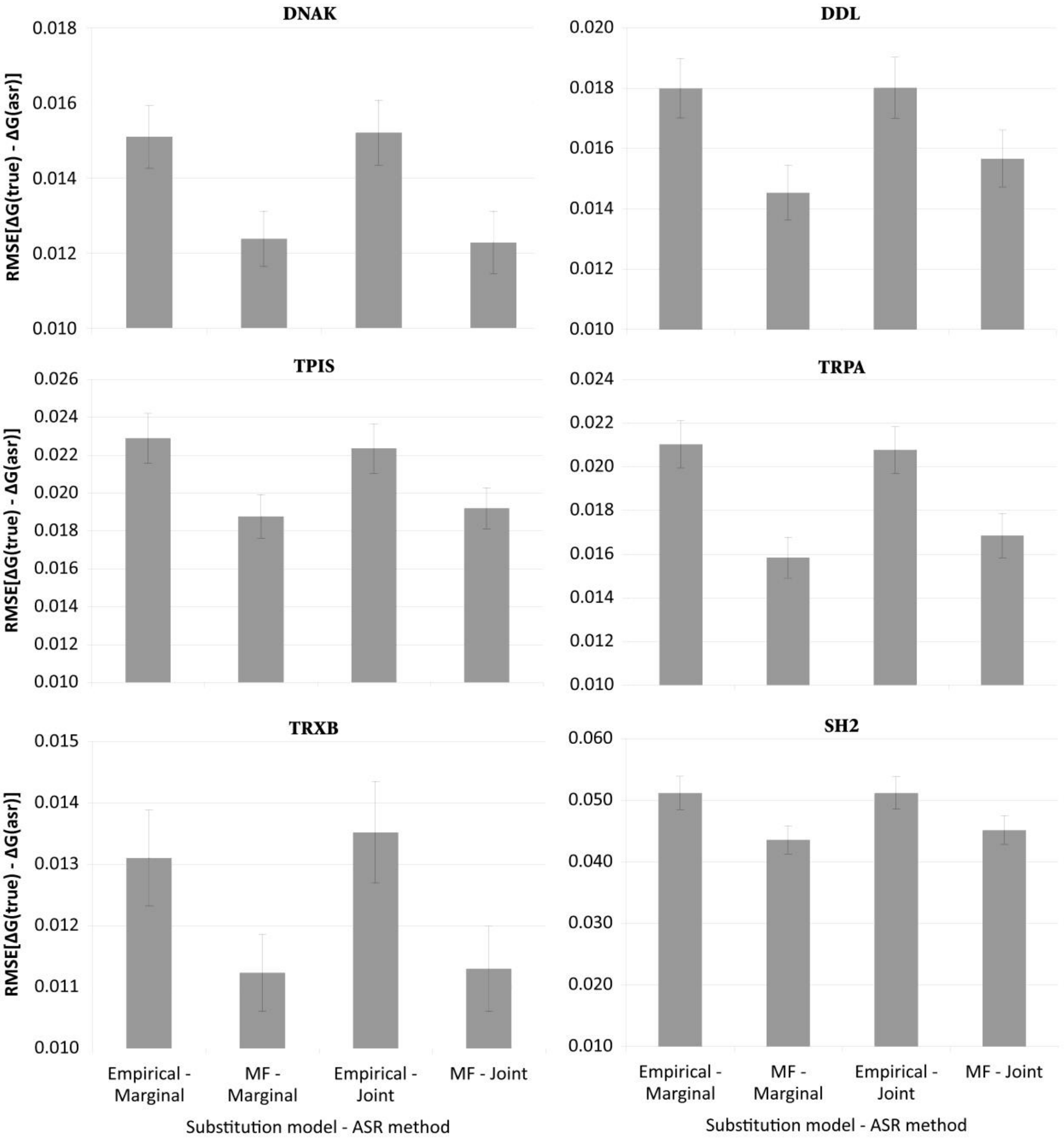

FIGURE 2. RMSE between the computed folding free energy of the extant and ancestral sequence at the root –sequence of the PDB– and the corresponding reconstructed ancestral sequence under MF and empirical substitution models. Both *joint* and *marginal* reconstructions are shown. Note that the MF model always generates ancestral proteins with energies closer to the extant protein (lower RMSE) compared to the empirical model. Error bars indicate standard error of the mean over 50 simulations.

We analyzed the behavior of both MF and empirical models under *marginal* and *joint* ASR reconstructions (Yang 1997). While the *joint* reconstruction estimates the most likely set of residues for all internal nodes (the global likelihood is calculated jointly considering all nodes at once) (Pupko et al. 2000), the *marginal* reconstruction obtains node by node estimates (the likelihood is calculated for each node and the global likelihood is obtained from all node-specific values) (Koshi and Goldstein 1996). We found similar results from both *joint* (RMSE median error for empirical: 0.0056; median error for MF: 0.005; Wilcoxon signed-rank test $p <$ 10e-14) and *marginal* reconstructions (median error for empirical: 0.0056; median error for MF: 0.0049; Wilcoxon signed-rank test $p <$ 10e-14). The *marginal* reconstruction estimates the free energies slightly more accurately (Figs. 1 and 2). The difference, assessed by computing the standard error of the mean over 50 simulations, is significant for the subtraction (Figs. S1 and S2; Wilcoxon signed-rank test p = 0.00022 for all comparisons) but not for the RMSE. The comparison between *joint* and *marginal* reconstructions did not depend on the underlying substitution model, either empirical or MF.

Simulated and inferred ancestral sequences showed that MF and empirical models generally yield similar sequence divergences (Figs. S4 and S5, supplementary material). Thus, the better performance of MF in reconstructing the folding stability of ancestral proteins is not due to higher identity between the reconstructed sequences. Nevertheless, the more realistic stability of the inferred ancestor represents a relevant improvement that addresses an important limitation of current ASR methods based on ML (see Williams et al. 2006).

Evaluation with data simulated under an additional structurally constrained substitution model of evolution

A caveat of the above analysis is that we estimated the stability of reconstructed proteins with a model similar to the one used to simulate evolution. To analyze whether this similarity explains the more realistic reconstructions, we also evaluated *ProtASR* through simulations under the structurally constrained substitution model utilized by Williams *et* al. (2006). Briefly, this model scores the difference in free energy between the native state and the denatured state, which consists of the unfolded state and misfolded states represented by 50 randomly generated decoy structures. The free energies are determined through a contact potential with interaction parameters given by Table VI in Miyazawa and Jernigan (1985), kT = 0.6 kcal/mol and number of alternative states $N = 10e^{54}$ (so that 3.4 conformations were available for each of the 104 amino acids). Individual nucleotides in the sequence were randomly mutated with a transition/transversion bias of two. Proposed mutations were stochastically fixed or rejected one at a time according to the Moran process (Moran 1958) with effective population size $Ne = 10e^{4}$ and fitness score f corresponding to the fraction of correctly folded protein, where $f=1/(1+exp(\Delta G/kT))$.

This model was applied to Human SAP protein (PDB: 1D4T), a member of the SH2 domain family (Table 1). The sequences were simulated along a randomly chosen tree with 10 terminal nodes after allowing the branch leading up to the root to burn into the model (that is, letting sequences evolve until the energy gap reached an asymptote with approximately similar density above and below the mean). Next, a neighbor-joining tree was inferred for each simulated alignment. As described above, we applied *ProtASR* to the simulated sequences at the tip nodes under both MF and the empirical model. Then, we estimated the folding free energies of the simulated and reconstructed ancestral proteins (following the procedure described in the previous section) and computed the RMSE and the bias between the simulated and estimated folding free energies.

ASR under MF generated ancestral sequences with energies closer to the energies of the simulated sequences compared to the empirical model (Figs. 1, 2, S1 and S2, SH2 at the bottom right; Wilcoxon signed-rank test for marginal error p < 10e-5). Again, the difference was more evident for the most ancestral node (Figs. 2 and S2), which displayed significant differences (assessed by comparing the standard error of the mean to simulations; Wilcoxon signed-rank test for marginal error p = 0.02673). However, the variation was smaller (a lower proportion of ancestral nodes present differences between models) than in the above benchmark where we applied a similar model of protein stability to simulate protein evolution and to compute the free energy of reconstructed and simulated proteins, suggesting that part (but not all) of the improvement in reconstructing ancestral stability may be explained by the similarity between the evolutionary process and the procedure to compute stability. *Marginal* and *joint* ASR again produced similar results (Figs. 1 and 2, bottom right). Divergence between the simulated and inferred ancestral sequences did not differ between MF and the empirical model (Figs. S4 and S5, bottom right), as was also seen for the simulations performed with *ProteinEvolver*.

PROTEIN FOLDING THERMODYNAMICS OF ANCESTRAL PROKARYOTIC PROTEINS

To illustrate how *ProtASR* can be applied to empirical data, we reconstructed the history of the protein folding thermodynamics of 5 extant Prokaryotic protein families (DDL, DNAK, TPIS, TRPA and TRXB; Table 1). These protein families allow investigating variations in thermodynamic properties of orthologous proteins that are likely to be due to the evolutionary process but not to changes of function (Bastolla et al. 2004). We inferred ancestral protein sequences for the aligned extant sequences with *ProtASR* under the MF model, using ML trees and *marginal* reconstruction. Using the stability model described in the previous section, we computed folding free energies for the inferred ancestral and extant sequences. Although we computed the folding free energy for all nodes, we recommend carefully interpreting internal nodes with low statistical support (bootstrap values < 0.7). Additionally, note that this is a gene tree and may differ from the species tree (Maddison 1997; Mallo et al. 2016), and therefore results should be interpreted at the protein/gene level rather than at the species level.

We found different levels of variation in free energy depending on the protein family, as well as the clades within a family (Figs. 3 and S6-S9, supplementary material). All studied protein families showed periods of increased, conserved and decreased folding stabilities through time (Fig. 4), consistent with a seascape model of protein evolution (Mustonen and Lassig 2009). The DDL enzyme family showed decreases in most lineages through time [e.g., remarkable in the species CAUCR (*Caulobacter crescentus*) and ZYMMO (*Zymomonas mobilis*)] (Figs. 3 and 4), a trend also found in TRPA (Figs. 4 and S8). DNAK, TPIS and TRXB had a similar number of branches with increased and decreased folding stabilities (Fig. 4). The chaperone DNAK and the Thioredoxin TRXB presented overall low variability in folding energies for all present and inferred sequences (Figs. 4, S6 and S9). Interestingly, chaperones exhibit signatures of strong selective pressure, in particular in endosymbiotic bacteria where they are highly expressed (Ishikawa 1984; Aksoy 1995; Warnecke and Rocha 2011), presumably to buffer against destabilizing changes that occur more frequently in small effective populations (Bastolla et al. 2004). We detected overall positive correlations between the free energy variation and the branch length (Figs. S10 and S11, supplementary material).

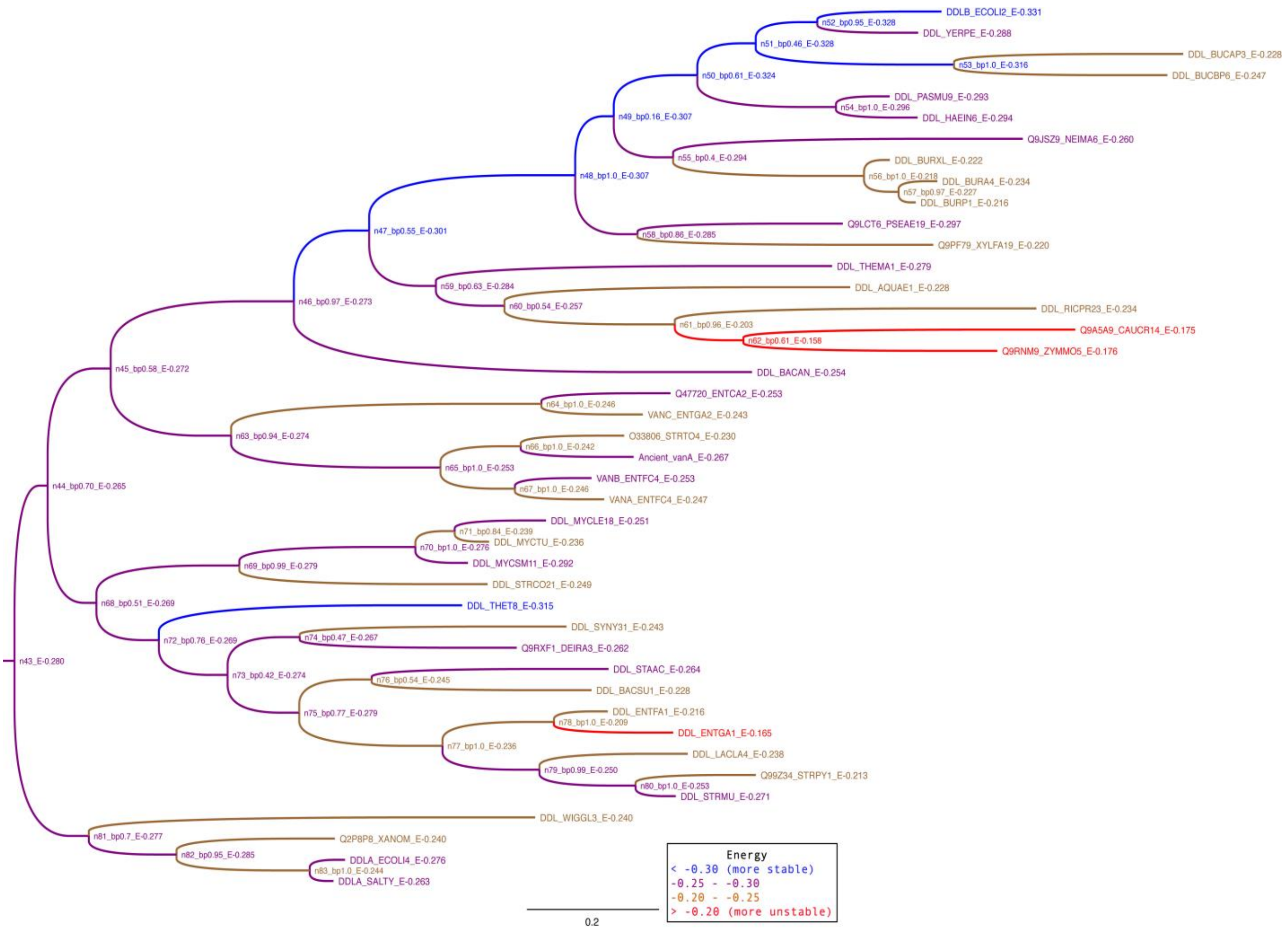

FIGURE 3. Folding free energy of the inferred ancestral proteins of the DDL protein family. The figure shows the ML phylogenetic tree (rooted to distinguish the paralogous genes *DdlA* and *DdlB*) with the following information for every node: Node number *n*, bootstrap *bp*

(only for internal nodes different to the root) and energy *E* of the corresponding sequence into the selected protein structure of the PDB.

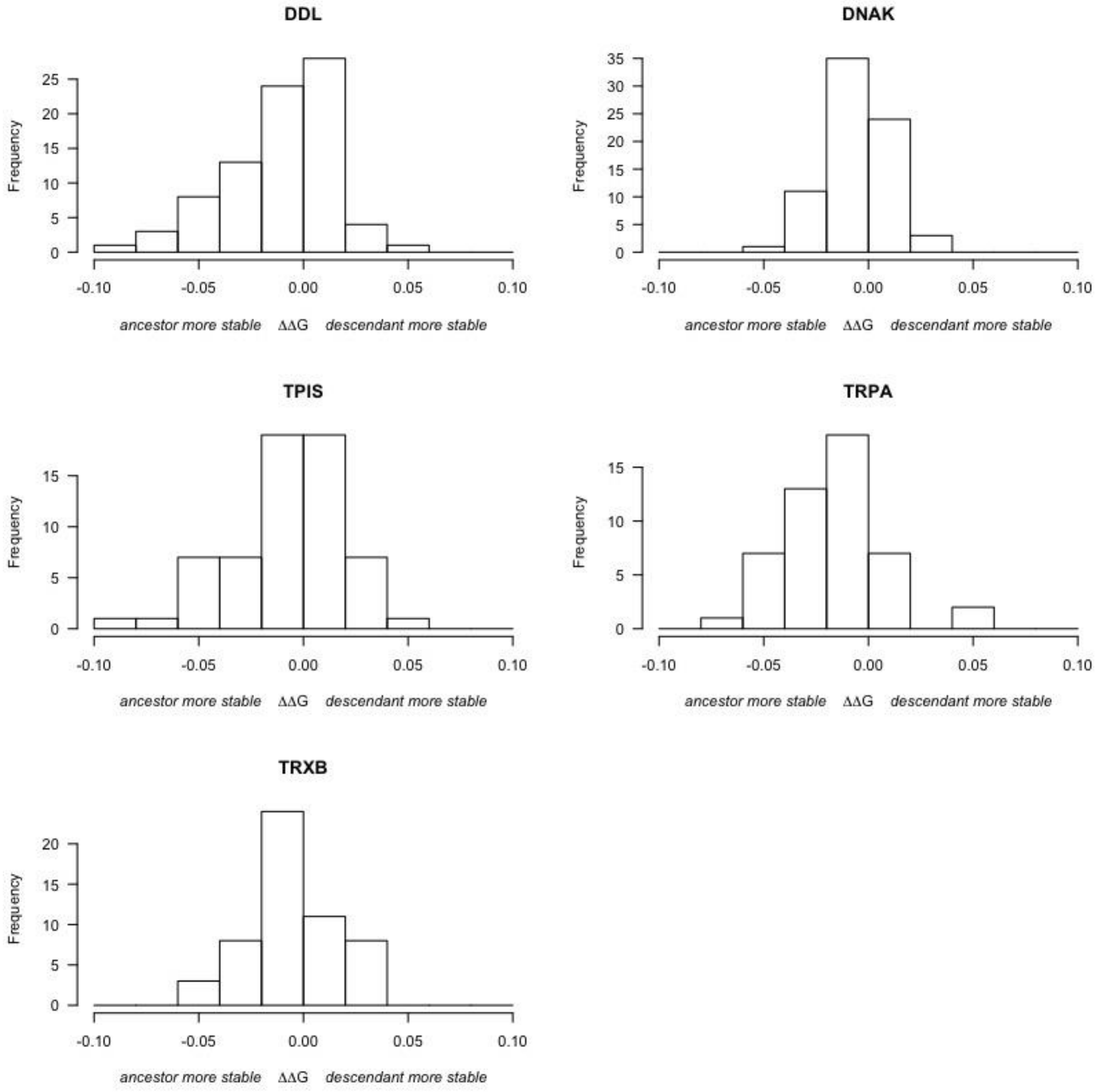

FIGURE 4. Histogram of folding free energy variation in branches ($\Delta G_{\mathbf{AncestralSequence}}$ - $\Delta G_{\mathbf{RecentSequence}}$) for the studied protein families. A negative free energy variation of a branch

indicates that the sequence of the ancestral node is more stable than the sequence of the descendant node. A positive value indicates the contrary.

DISCUSSION

MF models have previously been shown to more realistically represent the evolutionary process than empirical amino acid models and other structurally constrained models (Arenas et al. 2015), despite sharing the simplifying assumption of independently evolving sites. Here, we developed a new tool that applies MF to ASR of proteins. Our program *ProtASR* infers ancestral proteins while effectively accounting for stability constraints against both misfolding and unfolding, and it runs essentially in the same time as empirical models that do not consider structural constraints. We found that ancestral proteins reconstructed under MF have folding stabilities closer to those of simulated and extant proteins than proteins reconstructed through the empirical model or through a CAT model. It has been previously shown that ancestral sequences reconstructed with maximum likelihood methods tend to appear more stable than simulated or real sequences (Williams et al. 2006; Goldstein 2011). We found that this result also holds when applying MF as a substitution model, but that MF reduces the bias towards increased stability of reconstructed sequences. This finding is counterintuitive, since one might expect that the stability constraints considered in the MF model might have further increased the stability of reconstructed proteins, and it suggests that accounting for protein stability results in reconstructed ancestral proteins whose stability is more realistic, and not just stronger, than those obtained in the absence of structural constraints.

We advise users of *ProtASR* that care should be taken when specifying the input parameters, such as the temperature, the configurational entropies, or the exchangeability matrix used by

MF to compute the substitution rates. For first-time users we recommend using the default parameter values provided in the documentation and examples, since we have tested them on a variety of protein families (Arenas et al. 2013; Arenas et al. 2015 and the present work).

Our results suggest that *ProtASR* can be applied to estimate the history of protein stability in protein families, as we illustrate with five orthologous prokaryotic protein families. We find that protein stabilities vary through time in a complex manner, and ancestral proteins are not necessarily more stable than their descendants, contrasting with results obtained with simpler models (see Williams et al. 2006). Variations of protein stability along branches of the phylogenetic tree are consistent with a seascape model of protein evolution based on compensatory changes (Mustonen and Lassig 2009). More specifically, several lineage-specific biological processes may influence stability variations: (*i*) changes in effective population size that modulate natural selection (for instance passing from free living to intracellular lifestyles), (*ii*) changes in environmental temperature, which can affect the evolutionary process (at low temperature proteins evolve more neutrally, since the relationship between the free energy and the fraction of folded protein is more sigmoidal, and therefore smaller stabilities are sufficient to fold proteins (Serohijos and Shakhnovich 2014), (*iii*) changes in mutation rate and in mutation bias, which can also affect the protein stability that an evolving population can achieve (Mendez et al. 2010), or most interestingly, (*iv*) positive selection due to changes in protein function (e.g., Pascual-Garcia et al. 2009). Such effects, including discussions of how to model them, have recently been reviewed (Anisimova and Liberles 2012; Chi and Liberles 2016). Another advantage of the present framework is that it considers stability against both unfolding and misfolding, which may have evolutionary trade-offs (Mendez et al. 2010; Zheng et al. 2013). Overall, *ProtASR* is a useful tool in the phylogenetic toolbox, reflecting an advance over other methods currently available as software for the important problem of ancestral sequence reconstruction.

Supplementary Material

Supplementary Figures S1-S12, Appendix I and access to the studied data are available at Systematic Biology online (http://sysbio.oxfordjournals.org/).

Availability

ProtASR is written in C and Perl and it is freely available under the GPL license. Source code, executable files, a variety of ready-to-use examples and detailed documentation are available from https://github.com/miguelarenas/protasr. The program *DeltaGREM* to estimate the folding free energy against the unfolded and the misfolded state is available at https://ub.cbm.uam.es/software/Delta_GREM.php and accepts as input a list of protein structures and, optionally, a MSA or a list of mutations.

Funding

This work was supported by the Spanish Ministery of Economy and Competitivity through the grant BFU-40020 to UB. MA was initially supported by the Spanish Government with the “Juan de la Cierva” fellowship JCI-2011-10452 and the EMBO Short Term Fellowship ASTF 367-2013, and finally supported by the Portuguese Government through the FCT Starting Grant IF/00955/2014 and by the Spanish Government through the “Ramón y Cajal” fellowship RYC-2015-18241. Research at the CBMSO is facilitated by the Fundación Ramón Areces. CCW and DAL receive funding from US NSF DBI-1515704.

ACKNOWLEDGMENTS

We thank Tal Pupko for helpful advice, Stephanie Spielman for helpful discussions, Markus Porto for participating in prior stages of this work, and Jason Lai and Stephen Shank for technical assistance. We thank members of the Liberles research group, past and present, as well as members of Alfonso Valencia's research group for insightful comments. We thank Nicolas Lartillot for providing us technical recommendations about the program *PhyloBayes*. We also thank two anonymous reviewers for their constructive comments.

REFERENCES

Aksoy S. 1995. Molecular analysis of the endosymbionts of tsetse flies: 16S rDNA locus and over-expression of a chaperonin. Insect Mol. Biol. 4:23-29.

Alcalde M. 2015. Engineering the ligninolytic enzyme consortium. Trends Biotechnol. 33:155-162.

Anisimova M., Liberles D.A. 2012. Detecting and understanding natural selection. In: Cannarozzi G.M., Schneider A. editors. Codon Evolution. Oxford, Oxford University Press, p. 73-96.

Arenas M. 2015. Trends in substitution models of molecular evolution. Front. Genet. 6:319.

Arenas M., Dos Santos H.G., Posada D., Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics 29:3020-3028.

Arenas M., Posada D. 2010. Computational Design of Centralized HIV-1 Genes. Curr. HIV Res. 8:613-621.

Arenas M., Sanchez-Cobos A., Bastolla U. 2015. Maximum likelihood phylogenetic inference with selection on protein folding stability. Mol. Biol. Evol. 32:2195-2207.

Ashkenazy H., Penn O., Doron-Faigenboim A., Cohen O., Cannarozzi G., Zomer O., Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40:W580-584.

Bastolla U., Dehouck Y., Echave J. 2017. What evolution tells us about protein physics, and protein physics tells us about evolution. Curr. Opin. Struct. Biol. 42:59-66.

Bastolla U., Moya A., Viguera E., van Ham R.C. 2004. Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J. Mol. Biol. 343:1451-1466.

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

Bloom J.D., Labthavikul S.T., Otey C.R., Arnold F.H. 2006. Protein stability promotes evolvability. Proc. Natl. Acad. Sci. U S A 103:5869-5874.

Chang B.S., Donoghue M.J. 2000. Recreating ancestral proteins. Trends Ecol. Evol. 15:109-114.

Chi P.B., Liberles D.A. 2016. Selection on protein structure, interaction, and sequence. Protein Sci. 25:1168-1178.

DePristo M.A., Weinreich D.M., Hartl D.L. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 6:678-687.

Doria-Rose N.A., Learn G.H., Rodrigo A.G., Nickle D.C., Li F., Mahalanabis M., Hensel M.T., McLaughlin S., Edmonson P.F., Montefiori D., Barnett S.W., Haigwood N.L., Mullins J.I. 2005. Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. J. Virol. 79:11214-11224.

Echave J., Spielman S.J., Wilke C.O. 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17:109-121.

Gao F., Bhattacharya T., Gaschen B., Taylor J., Moore J.P., Novitsky V., Yusim K., Lang D., Foley B., Beddows S., Alam M., Haynes B., Hahn B.H., Korber B. 2003. Consensus and ancestral state HIV vaccines. Science 299:1515-1518.

Gaucher E.A., Govindarajan S., Ganesh O.K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704-707.

Goldstein R.A. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:1396-1407.

Govindarajan S., Goldstein R.A. 1997. Evolution of model proteins on a foldability landscape. Proteins 29:461-466.

Grahnen J.A., Nandakumar P., Kubelka J., Liberles D.A. 2011. Biophysical and structural considerations for protein sequence evolution. BMC Evol. Biol. 11:361.

Hobbs J.K., Shepherd C., Saul D.J., Demetras N.J., Haaning S., Monk C.R., Daniel R.M., Arcus V.L. 2012. On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. Mol. Biol. Evol. 29:825-835.

Huang T.T., del Valle Marcos M.L., Hwang J.K., Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14:78.

Illergard K., Ardell D.H., Elofsson A. 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77:499-508.

Ishikawa H. 1984. Characterization of the protein species synthetized *in vivo* and *in vitro* by an aphid endosymbiont. Insect Biochem. 14:417-425.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275-282.

Kachroo A.H., Laurent J.M., Yellman C.M., Meyer A.G., Wilke C.O., Marcotte E.M. 2015. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348:921-925.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772-780.

Kodra J.T., Skovgaard M., Madsen D., Liberles D.A. 2007. Linking sequence to function in drug design with ancestral sequence reconstruction. In: Liberles D.A. editor. Ancestral Sequence Reconstruction, Oxford University Press, p. 34-39.

Kosakovsky Pond S.L., Frost S.D., Muse S.V. 2005. HYPHY: Hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Koshi J.M., Goldstein R.A. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. 42:313-320.

Kothe D.L., Li Y., Decker J.M., Bibollet-Ruche F., Zammit K.P., Salazar M.G., Chen Y., Weng Z., Weaver E.A., Gao F., Haynes B.F., Shaw G.M., Korber B.T., Hahn B.H. 2006. Ancestral and consensus envelope immunogens for HIV-1 subtype C. Virology 352:438-449.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095-1109.

Liberles D.A. 2007. Ancestral Sequence Reconstruction. Oxford University Press.

Liberles D.A., Teichmann S.A., Bahar I., Bastolla U., Bloom J., Bornberg-Bauer E., Colwell L.J., de Koning A.P., Dokholyan N.V., Echave J., Elofsson A., Gerloff D.L., Goldstein R.A., Grahnen J.A., Holder M.T., Lakner C., Lartillot N., Lovell S.C., Naylor G., Perica T., Pollock

D.D., Pupko T., Regan L., Roger A., Rubinstein N., Shakhnovich E., Sjolander K., Sunyaev S., Teufel A.I., Thorne J.L., Thornton J.W., Weinreich D.M., Whelan S. 2012. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21:769-785.

Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Mallo D., Sánchez-Cobos A., Arenas M. 2016. Diverse Considerations for Successful Phylogenetic Tree Reconstruction: Impacts from Model Misspecification, Recombination, Homoplasy, and Pattern Recognition. In: Elloumi M., Iliopoulos C., Wang J., Zomaya A. editors. Pattern Recognition in Computational Molecular Biology, John Wiley & Sons, Inc, p. 439-456.

Mendez R., Fritsche M., Porto M., Bastolla U. 2010. Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comput. Biol. 6:e1000767.

Merkl R., Sterner R. 2016. Ancestral protein reconstruction: techniques and applications. Biol. Chem. 397:1-21.

Minning J., Porto M., Bastolla U. 2013. Detecting selection for negative design in proteins through an improved model of the misfolded state. Proteins 81:1102-1112.

Miyazawa S., Jernigan R.L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534-552.

Moran P.A.P. 1958. Random processes in genetics. Proc. Camb. Philos. Soc. 54:60-71.

Mustonen V., Lassig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25:111-119.

Parisi G., Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18:750-756.

Pascual-Garcia A., Abia D., Mendez R., Nido G.S., Bastolla U. 2009. Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78:181-196.

Perez-Jimenez R., Ingles-Prieto A., Zhao Z.M., Sanchez-Romero I., Alegre-Cebollada J., Kosuri P., Garcia-Manyes S., Kappock T.J., Tanokura M., Holmgren A., Sanchez-Ruiz J.M., Gaucher E.A., Fernandez J.M. 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat. Struct. Mol. Biol. 18:592-596.

Pupko T., Pe'er I., Shamir R., Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17:890-896.

Serohijos A.W., Shakhnovich E.I. 2014. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr. Opin. Struct. Biol. 26:84-91.

Taverna D.M., Goldstein R.A. 2002. Why are proteins marginally stable? Proteins 46:105-109.

Thomson J.M., Gaucher E.A., Burgan M.F., De Kee D.W., Li T., Aris J.P., Benner S.A. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nat. Genet. 37:630-635.

Warnecke T., Rocha E.P. 2011. Function-specific accelerations in rates of sequence evolution suggest predictable epistatic responses to reduced effective population size. Mol. Biol. Evol. 28:2339-2349.

Wilke C.O. 2012. Bringing molecules back into molecular evolution. PLoS Comput. Biol. 8:e1002572.

Williams P.D., Pollock D.D., Blackburne B.P., Goldstein R.A. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput. Biol. 2:e69.

Yamashiro K., Yokobori S., Koikeda S., Yamagishi A. 2010. Improvement of Bacillus circulans beta-amylase activity attained using the ancestral mutation method. Protein Eng. Des. Sel. 23:519-528.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosciences 13:555-556.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Zheng W., Schafer N.P., Wolynes P.G. 2013. Frustration in the energy landscapes of multidomain protein misfolding. Proc. Natl. Acad. Sci. U S A 110:1680-1685.