

This article has been published in *Systematic Biology* by Oxford University Press

**Accepted manuscript.**

Miguel Arenas, Claudia C. Weber, David A. Liberles, Ugo Bastolla, ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability, *Systematic Biology*, Volume 66, Issue 6, November 2017, Pages 1054–1064, <https://doi.org/10.1093/sysbio/syw121>

General rights:

© The Author(s) 2017. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

2

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

5

6 Miguel Arenas<sup>1,2,3,4,\*</sup>, Claudia C. Weber<sup>6</sup>, David A. Liberles<sup>5,6</sup>, and Ugo Bastolla<sup>3</sup>

7

8

9 <sup>1</sup>Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, Porto, Portugal.

10 <sup>2</sup>Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP),  
11 Porto, Portugal.

12 <sup>3</sup>Centre for Molecular Biology Severo Ochoa (CBMSO), Consejo Superior de Investigaciones  
13 Científicas (CSIC), Madrid, Spain.

14 <sup>4</sup>Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain.

15 <sup>5</sup>Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

16 <sup>6</sup>Department of Biology and Center for Computational Genetics and Genomics, Temple  
17 University, Philadelphia, PA 19122, USA.

18

19

20 **Email addresses:**

21 MA: [miguellmmab@gmail.com](mailto:miguellmmab@gmail.com)

22 CCW: [claudia.weber@temple.edu](mailto:claudia.weber@temple.edu)

23 DAL: [daliberles@temple.edu](mailto:daliberles@temple.edu)

24 UB: [ubastolla@cbm.csic.es](mailto:ubastolla@cbm.csic.es)

25

26

27 **Corresponding author:**

28 *Miguel Arenas*

29 *Instituto de Investigação e Inovação em Saúde (i3S)*

30 *University of Porto*

31 *Rua Alfredo Allen, 208*

32 *4200-135 Porto, Portugal*

33 *E-mail address: [miguelmmmab@gmail.com](mailto:miguelmmmab@gmail.com)*

34 *Phone: +351 220408800 Ext. 6153*

35

36

37 **Running head:** ASR accounting for structural constraints

38 **Keywords:** ancestral sequence reconstruction, protein evolution, molecular adaptation,

39 phylogenetics, folding stability, protein structure

## 40 ABSTRACT

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

## 65 INTRODUCTION

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

81

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

89 Echave 2001; Taverna and Goldstein 2002; DePristo et al. 2005; Bastolla et al. 2006; Bloom et  
90 al. 2006; Goldstein 2011; Grahnen et al. 2011; Liberles et al. 2012; Wilke 2012; Arenas et al.  
91 2013; Huang et al. 2014; Arenas 2015; Arenas et al. 2015; Chi and Liberles 2016; Echave et al.  
92 2016; Bastolla et al. 2017) since thermodynamic stability is an important source of selective  
93 constraint (intrinsically disordered proteins aside). Unfortunately, structurally constrained  
94 models of protein evolution are not yet well-established in the phylogenetic pipeline, mainly  
95 due to the complexity of incorporating site-dependence in ML functions. Of course, it would be  
96 more realistic to also incorporate selection on protein function in light of evidence from  
97 experimental studies that suggests relevant factors such as binding (Kachroo et al. 2015).  
98 However, this requires additional knowledge about the protein family, *ad-hoc* assumptions  
99 about the constraints on the functionally important sites, and how they may change under  
100 functional selection. Compared to structural constraints, it is challenging to formulate general  
101 rules about functional constraints beyond inter-molecular protein binding.

102

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

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

119

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

130

### 131 **NEW APPROACHES: PROTASR**

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

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

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

149  
150 The *ProtASR* user inputs a multiple alignment of protein sequences, a rooted phylogenetic tree,  
151 a PDB file with a protein structure representative of the alignment (see below) and a set of  
152 parameters to define the MF model. These include the environmental temperature, the  
153 configurational entropies per residue for the unfolded and misfolded states, the source of the  
154 background amino-acid frequencies (user-specified, derived from the protein structure or  
155 derived from the alignment) and the exchangeability matrix needed to compute the substitution  
156 rates, which may either correspond to an empirical substitution model or be internally  
157 computed from evolutionary parameters at the nucleotide level (e.g., nucleotide frequencies and  
158 transition/transversion rate ratio). Detailed information and recommendations about the input



159 parameters are provided in the software documentation. Computation is efficient and times  
 160 range from seconds to minutes depending on protein length and number of sequences (see  
 161 Table 1).

162

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

167

Entry	Protein Family	Gene	Pfam code	Uniprot code	PDB code	Protein length	Sample size	Seq Id (%)	Computing Time (s)
1	<i>D</i> -ala <i>D</i> -ala ligases	DDL	PF07478	DDLB_E COLI	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 $\alpha$ chain	TRPA	PF00290	TRPA_ S ALTY	1A50	276	25	47.4	38.5
5	Thioredoxins I	TRXB	PF00070	TRXB_ E COLI	1TDE	375	28	46.4	53.4
6	SH2 domain	SH2	PF00017		1D4T	104	10	69.8	9.3

168

169

170 *ProtASR* assumes that the input protein structure is representative of the proteins included in  
 171 the alignment and therefore, protein sequences should fold into structures. This is a reasonable  
 172 assumption since protein structures are typically conserved over the range of protein sequence  
 173 divergence in a gene family (Illergard et al. 2009; Pascual-Garcia et al. 2009). To simplify

174 computations and reduce potential artefacts from calculated structures that are not protein-like,  
175 *ProtASR* assumes perfect conservation of the protein structure through the evolutionary history  
176 of the analyzed protein family. Additionally, one sequence in the input alignment must  
177 correspond to the sequence of the input PDB structure (or alternatively, the input alignment and  
178 the sequence of the input PDB file must contain an equal number of sites that are homologous  
179 without gaps) to allow unambiguous alignment between the structure and sequences.

180

### 181 ***EVALUATING PROTASR***

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

187

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

190 As a first benchmark we analyzed the following Prokaryotic protein families: DDL, DNAK,  
191 TPIS, TRPA and TRXB (Table 1). Each family consists of a putative group of homologs with  
192 extant sequences longer than 200 amino acids with members in many bacterial species  
193 (Bastolla et al. 2004), allowing well-supported phylogenies to be generated. The datasets were  
194 downloaded from the Pfam database, realigned with *MAFFT* (Katoh and Standley 2013) and  
195 ML phylogenetic trees were reconstructed under the JTT substitution model (Jones et al. 1992).  
196 The trees were rooted with an Eukaryotic protein (or an Eukaryotic group) as outgroup. Next,

197 for each family we chose one representative protein with a known PDB structure as the root  
198 sequence and evolved it along the inferred phylogeny 50 times with *ProteinEvolver* (Arenas et  
199 al. 2013). *ProteinEvolver* employs a similar energy function with structural constraints as MF,  
200 but it is more realistic because it implements a model with site-dependent constraints, while  
201 MF assumes that sites evolve independently to allow its incorporation into likelihood functions.  
202 We ran *ProteinEvolver* under a site-dependent model with standard parameters suggested in  
203 Arenas *et al.* (2013). From each simulation we obtained sequences for all internal and tip nodes.  
204 We then used the multiple sequence alignment (MSA) of the tip nodes to perform ASR under  
205 the empirical JTT model and under the MF model with the exchangeability matrix determined  
206 by the same JTT rate matrix. Hence, the structural constraints captured in MF are the only  
207 difference between the two models. As an additional comparison, we performed ASR under the  
208 CAT model implemented in *PhyloBayes* (Lartillot *et al.* 2009), which estimates the exchange  
209 rates and amino-acid equilibrium frequency vectors from the data (Lartillot and Philippe 2004).  
210 Due to the computational cost of these calculations, we considered only the TPIS and TRPA  
211 protein families, which had fewer sequences to consider (technical details about ASR with  
212 *PhyloBayes* are described in Appendix I of the supplementary material).

213 Subsequently, we estimated the folding free energy of all inferred ancestral sequences by using  
214 the stability model implemented in *ProteinEvolver* (Minning *et al.* 2013), which considers the  
215 free energy difference between the native state and both the unfolded and misfolded states. In  
216 these computations, for each sequence of the MSA the native state is identified as the structure  
217 with the lowest contact free energy among a large number of structures available in the PDB  
218 for the studied protein family. Considering multiple structures is particularly important when  
219 analyzing real protein families in order to reduce the bias to assign a lower free energy to  
220 sequences closer to the sequence of the representative protein. The free energy of the misfolded

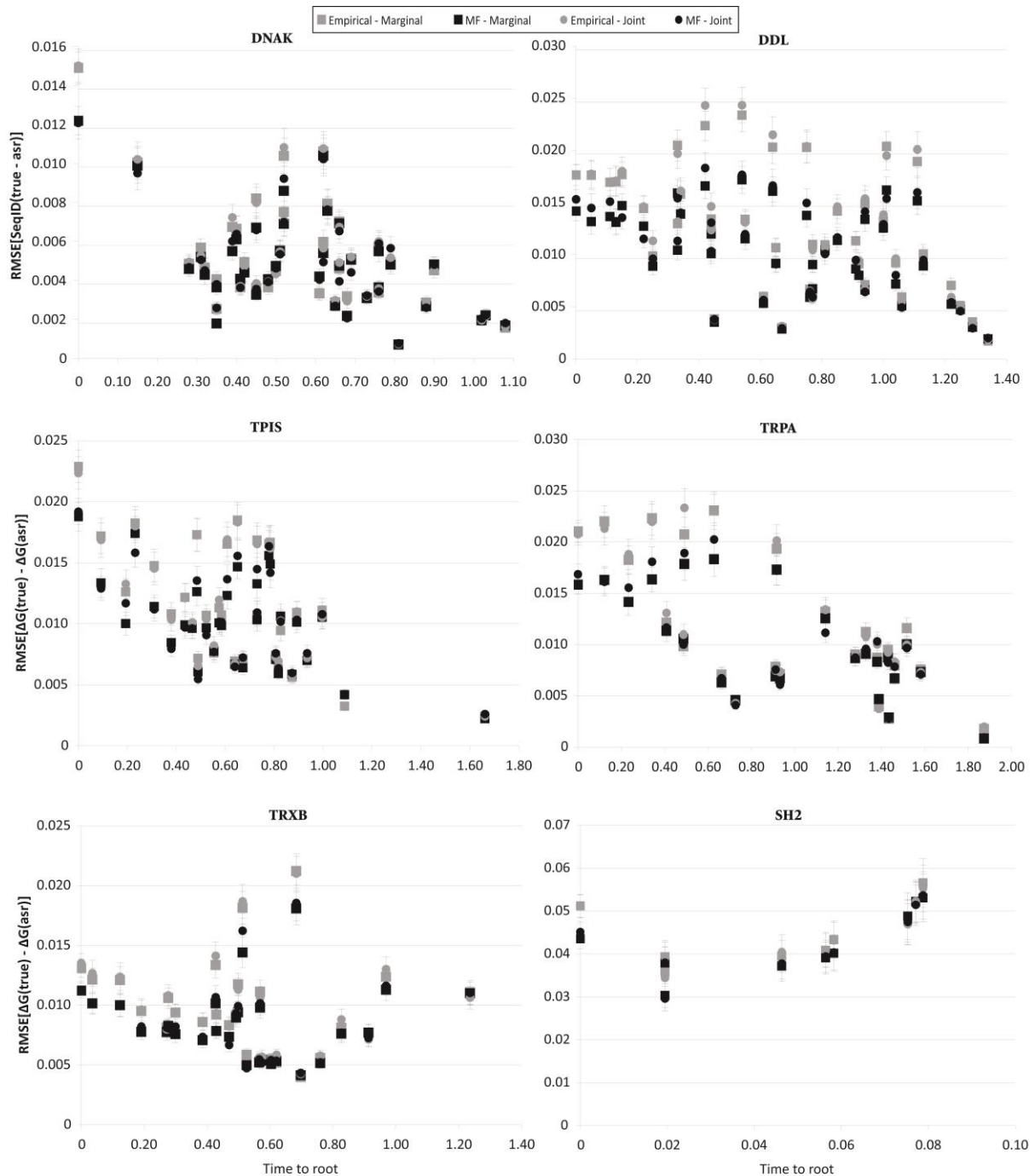
221 state is estimated through a Random Energy Model (REM) based on the mean and the variance  
222 of the contact energy of generic compact contact matrices and on their estimated  
223 configurational entropy (Minning et al. 2013) and the free energy of the unfolded state is  
224 estimated through its configurational entropy. Finally, we calculated the bias (signed difference  
225 between average values) and the Root Mean Square Error (RMSE) of the folding free energies  
226 estimated for the simulated sequences and the inferred sequences derived from both MF and  
227 empirical models.

228

229 For all protein families, ancestral sequences generated through the MF model showed free  
230 energies significantly closer to those of the simulated sequences (that is, smaller RMSE) than  
231 ancestral sequences generated through the empirical model (Figs. 1, 2, S1 and S2,  
232 supplementary material). The improvement of MF models was heterogeneous with respect to  
233 the distance from the reconstructed node to the root (Figs. 1 and S1). This is consistent with  
234 the expectation that, due to the influence of the substitution model, error increases with  
235 evolutionary distance from extant sequences at tip nodes (e.g., Williams et al. 2006). The error  
236 is largest at the root. MF consistently significantly outperformed the empirical model in terms  
237 of reconstructing the stability at the root (Figs. 2 and S2; Wilcoxon signed-rank test for error  $p$   
238  $< 10e-5$ ). Since the root is the sequence of the PDB structure, while other sequences are the  
239 result of simulations, this is an important test that assesses the stability of real protein  
240 sequences. In addition, MF also significantly improved the reconstruction of the stability of  
241 ancestral proteins compared to the CAT model (Fig. S3, supplementary material; Wilcoxon  
242 signed-rank test  $p < 5.9e-47$ ).

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

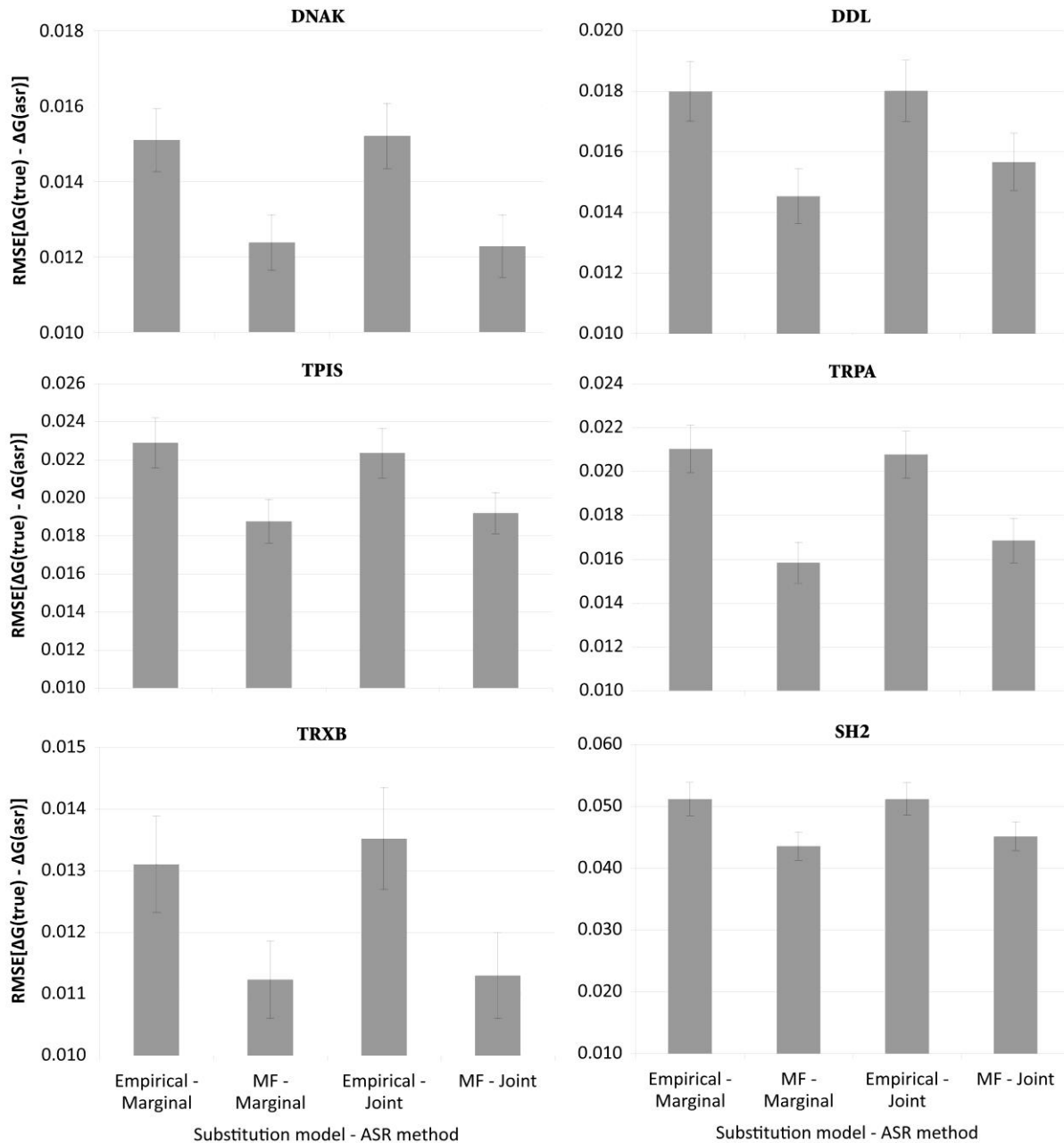
245 2011). Importantly, the MF model reduces this bias when compared to sequences reconstructed  
246 with empirical models (Figs. S1 and S2). While this model explicitly considers the  
247 thermodynamic effects of a substitution, potentially generating more neutral behavior for  
248 destabilizing changes in an already stable protein, it still lacks the segregating deleterious  
249 changes that would be expected to be sampled in any sequence at the tips (or along the tree).  
250



251

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

257 proteins, with respect to the empirical model (grey squares and circles), although this effect is  
 258 variable among nodes. As expected, the RMSE tends to increase at larger distance from the tip  
 259 nodes. Note also the small differences between *joint* and *marginal* ASR, which are not  
 260 significant. Error bars indicate standard error of the mean over 50 computer simulations.  
 261



262

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

269  
270 We analyzed the behavior of both MF and empirical models under *marginal* and *joint* ASR  
271 reconstructions (Yang 1997). While the *joint* reconstruction estimates the most likely set of  
272 residues for all internal nodes (the global likelihood is calculated jointly considering all nodes  
273 at once) (Pupko et al. 2000), the *marginal* reconstruction obtains node by node estimates (the  
274 likelihood is calculated for each node and the global likelihood is obtained from all node-  
275 specific values) (Koshi and Goldstein 1996). We found similar results from both *joint* (RMSE  
276 median error for empirical: 0.0056; median error for MF: 0.005; Wilcoxon signed-rank test  $p <$   
277  $10e-14$ ) and *marginal* reconstructions (median error for empirical: 0.0056; median error for MF:  
278 0.0049; Wilcoxon signed-rank test  $p < 10e-14$ ). The *marginal* reconstruction estimates the free  
279 energies slightly more accurately (Figs. 1 and 2). The difference, assessed by computing the  
280 standard error of the mean over 50 simulations, is significant for the subtraction (Figs. S1 and  
281 S2; Wilcoxon signed-rank test  $p = 0.00022$  for all comparisons) but not for the RMSE. The  
282 comparison between *joint* and *marginal* reconstructions did not depend on the underlying  
283 substitution model, either empirical or MF.

284 Simulated and inferred ancestral sequences showed that MF and empirical models generally  
285 yield similar sequence divergences (Figs. S4 and S5, supplementary material). Thus, the better  
286 performance of MF in reconstructing the folding stability of ancestral proteins is not due to



287 higher identity between the reconstructed sequences. Nevertheless, the more realistic stability  
288 of the inferred ancestor represents a relevant improvement that addresses an important  
289 limitation of current ASR methods based on ML (see Williams et al. 2006).

290

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

293 A caveat of the above analysis is that we estimated the stability of reconstructed proteins with a  
294 model similar to the one used to simulate evolution. To analyze whether this similarity explains  
295 the more realistic reconstructions, we also evaluated *ProtASR* through simulations under the  
296 structurally constrained substitution model utilized by Williams *et al.* (2006). Briefly, this  
297 model scores the difference in free energy between the native state and the denatured state,  
298 which consists of the unfolded state and misfolded states represented by 50 randomly generated  
299 decoy structures. The free energies are determined through a contact potential with interaction  
300 parameters given by Table VI in Miyazawa and Jernigan (1985),  $kT = 0.6$  kcal/mol and number  
301 of alternative states  $N = 10e^{54}$  (so that 3.4 conformations were available for each of the 104  
302 amino acids). Individual nucleotides in the sequence were randomly mutated with a  
303 transition/transversion bias of two. Proposed mutations were stochastically fixed or rejected  
304 one at a time according to the Moran process (Moran 1958) with effective population size  $N_e =$   
305  $10e^4$  and fitness score  $f$  corresponding to the fraction of correctly folded protein, where  
306  $f = 1 / (1 + \exp(\Delta G / kT))$ .

307 This model was applied to Human SAP protein (PDB: 1D4T), a member of the SH2 domain  
308 family (Table 1). The sequences were simulated along a randomly chosen tree with 10 terminal  
309 nodes after allowing the branch leading up to the root to burn into the model (that is, letting  
310 sequences evolve until the energy gap reached an asymptote with approximately similar density

311 above and below the mean). Next, a neighbor-joining tree was inferred for each simulated  
312 alignment. As described above, we applied *ProtASR* to the simulated sequences at the tip nodes  
313 under both MF and the empirical model. Then, we estimated the folding free energies of the  
314 simulated and reconstructed ancestral proteins (following the procedure described in the  
315 previous section) and computed the RMSE and the bias between the simulated and estimated  
316 folding free energies.

317  
318 ASR under MF generated ancestral sequences with energies closer to the energies of the  
319 simulated sequences compared to the empirical model (Figs. 1, 2, S1 and S2, SH2 at the bottom  
320 right; Wilcoxon signed-rank test for marginal error  $p < 10e-5$ ). Again, the difference was more  
321 evident for the most ancestral node (Figs. 2 and S2), which displayed significant differences  
322 (assessed by comparing the standard error of the mean to simulations; Wilcoxon signed-rank  
323 test for marginal error  $p = 0.02673$ ). However, the variation was smaller (a lower proportion of  
324 ancestral nodes present differences between models) than in the above benchmark where we  
325 applied a similar model of protein stability to simulate protein evolution and to compute the  
326 free energy of reconstructed and simulated proteins, suggesting that part (but not all) of the  
327 improvement in reconstructing ancestral stability may be explained by the similarity between  
328 the evolutionary process and the procedure to compute stability. *Marginal* and *joint* ASR again  
329 produced similar results (Figs. 1 and 2, bottom right). Divergence between the simulated and  
330 inferred ancestral sequences did not differ between MF and the empirical model (Figs. S4 and  
331 S5, bottom right), as was also seen for the simulations performed with *ProteinEvolver*.

332

### 333 **PROTEIN FOLDING THERMODYNAMICS OF ANCESTRAL PROKARYOTIC PROTEINS**

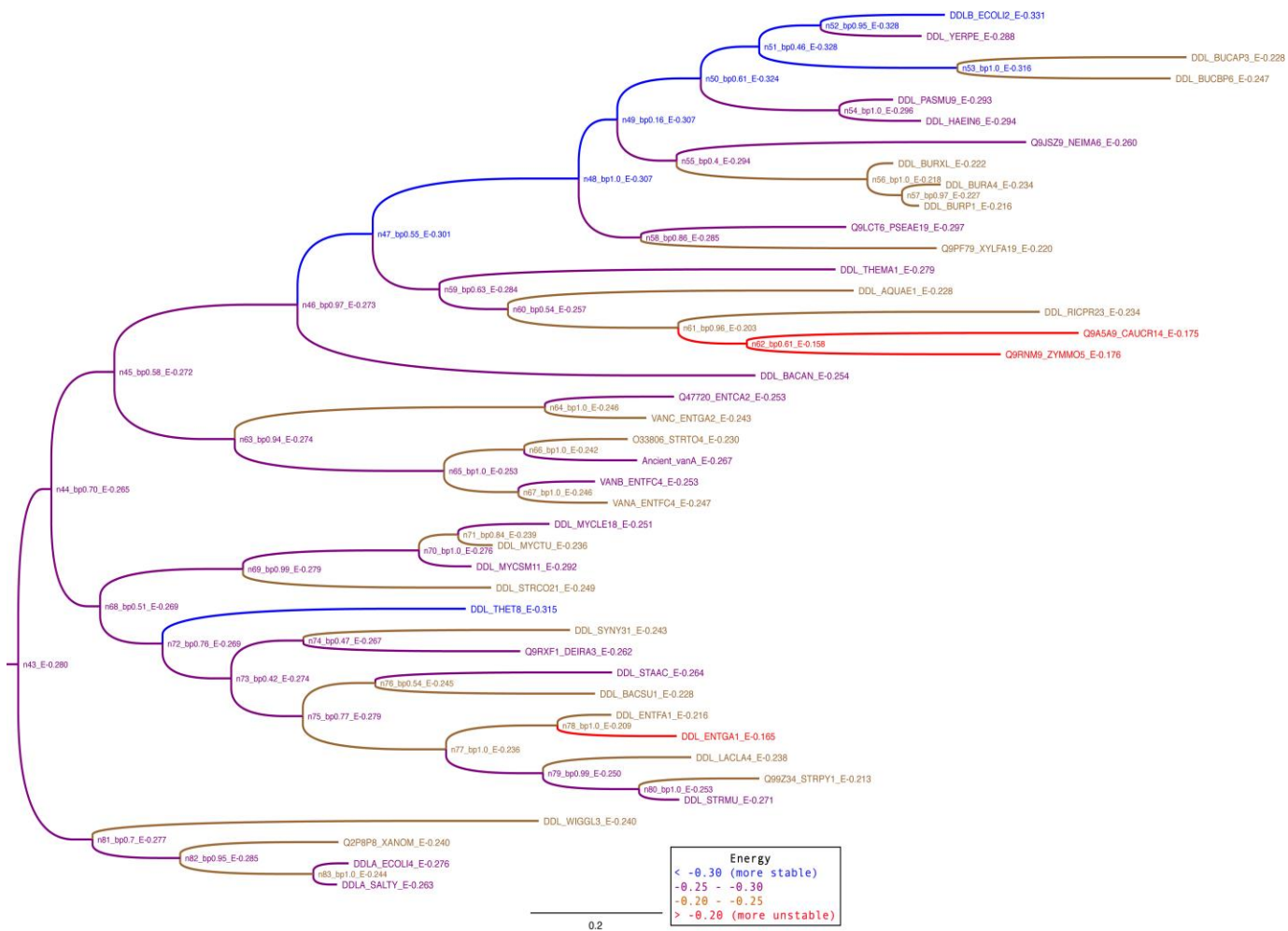
334 To illustrate how *ProtASR* can be applied to empirical data, we reconstructed the history of the  
335 protein folding thermodynamics of 5 extant Prokaryotic protein families (DDL, DNAK, TPIS,  
336 TRPA and TRXB; Table 1). These protein families allow investigating variations in  
337 thermodynamic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

346

347 We found different levels of variation in free energy depending on the protein family, as well  
348 as the clades within a family (Figs. 3 and S6-S9, supplementary material). All studied protein  
349 families showed periods of increased, conserved and decreased folding stabilities through time  
350 (Fig. 4), consistent with a seascape model of protein evolution (Mustonen and Lassig 2009).  
351 The DDL enzyme family showed decreases in most lineages through time [e.g., remarkable in  
352 the species CAUCR (*Caulobacter crescentus*) and ZYMMO (*Zymomonas mobilis*)] (Figs. 3  
353 and 4), a trend also found in TRPA (Figs. 4 and S8). DNAK, TPIS and TRXB had a similar  
354 number of branches with increased and decreased folding stabilities (Fig. 4). The chaperone  
355 DNAK and the Thioredoxin TRXB presented overall low variability in folding energies for all  
356 present and inferred sequences (Figs. 4, S6 and S9). Interestingly, chaperones exhibit signatures

357 of strong selective pressure, in particular in endosymbiotic bacteria where they are highly  
 358 expressed (Ishikawa 1984; Aksoy 1995; Warnecke and Rocha 2011), presumably to buffer  
 359 against destabilizing changes that occur more frequently in small effective populations  
 360 (Bastolla et al. 2004). We detected overall positive correlations between the free energy  
 361 variation and the branch length (Figs. S10 and S11, supplementary material).

362

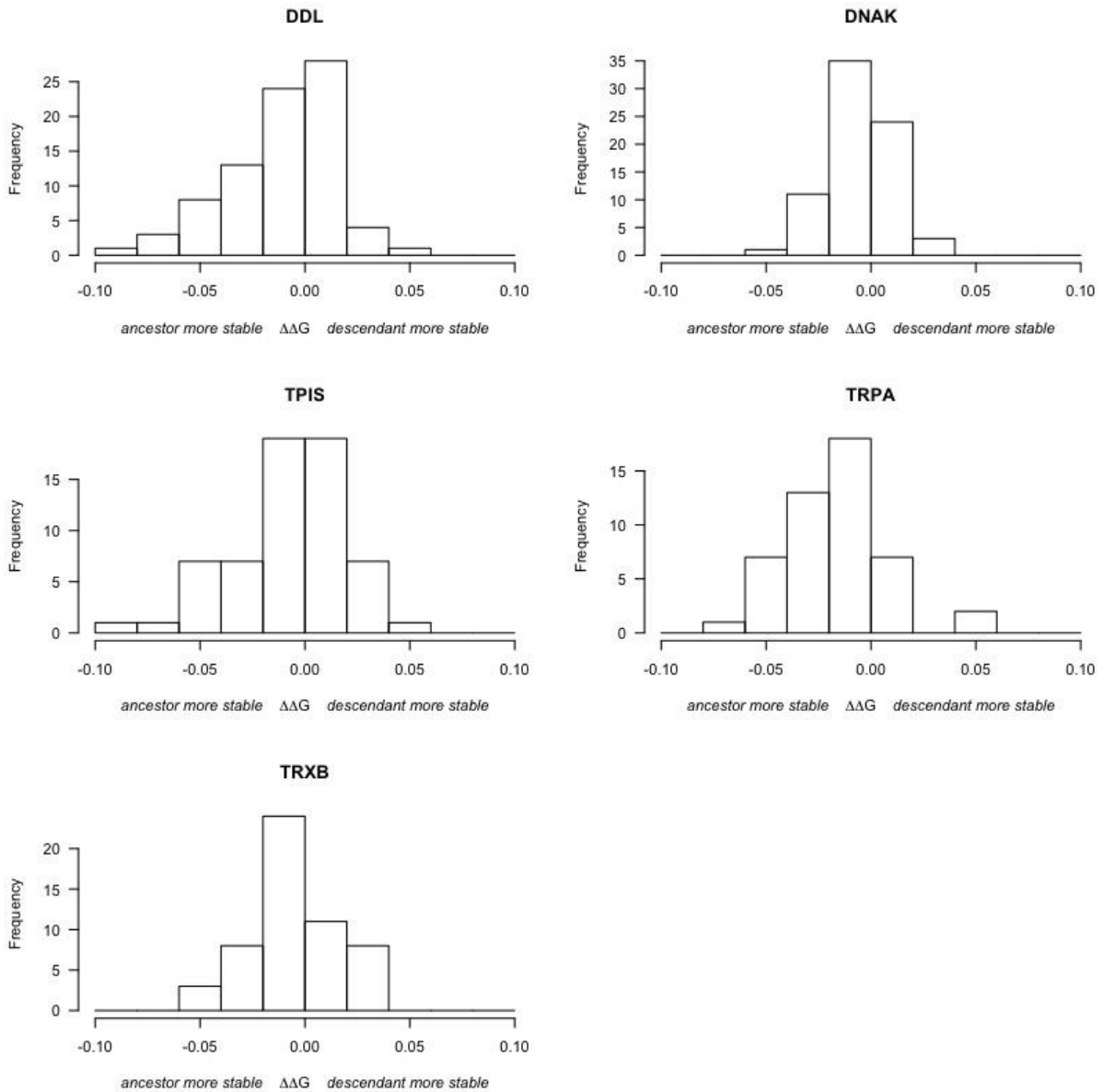


363

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

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

369



370

371 **FIGURE 4. Histogram of folding free energy variation in branches ( $\Delta G_{\text{AncestralSequence}}$  -**

372  **$\Delta G_{\text{RecentSequence}}$ ) for the studied protein families. A negative free energy variation of a branch**

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

375

## 376 **DISCUSSION**

377 MF models have previously been shown to more realistically represent the evolutionary process  
378 than empirical amino acid models and other structurally constrained models (Arenas et al.  
379 2015), despite sharing the simplifying assumption of independently evolving sites. Here, we  
380 developed a new tool that applies MF to ASR of proteins. Our program *ProtASR* infers  
381 ancestral proteins while effectively accounting for stability constraints against both misfolding  
382 and unfolding, and it runs essentially in the same time as empirical models that do not consider  
383 structural constraints. We found that ancestral proteins reconstructed under MF have folding  
384 stabilities closer to those of simulated and extant proteins than proteins reconstructed through  
385 the empirical model or through a CAT model. It has been previously shown that ancestral  
386 sequences reconstructed with maximum likelihood methods tend to appear more stable than  
387 simulated or real sequences (Williams et al. 2006; Goldstein 2011). We found that this result  
388 also holds when applying MF as a substitution model, but that MF reduces the bias towards  
389 increased stability of reconstructed sequences. This finding is counterintuitive, since one might  
390 expect that the stability constraints considered in the MF model might have further increased  
391 the stability of reconstructed proteins, and it suggests that accounting for protein stability  
392 results in reconstructed ancestral proteins whose stability is more realistic, and not just stronger,  
393 than those obtained in the absence of structural constraints.

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

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

399 Our results suggest that *ProtASR* can be applied to estimate the history of protein stability in  
400 protein families, as we illustrate with five orthologous prokaryotic protein families. We find  
401 that protein stabilities vary through time in a complex manner, and ancestral proteins are not  
402 necessarily more stable than their descendants, contrasting with results obtained with simpler  
403 models (see Williams et al. 2006). Variations of protein stability along branches of the  
404 phylogenetic tree are consistent with a seascape model of protein evolution based on  
405 compensatory changes (Mustonen and Lassig 2009). More specifically, several lineage-specific  
406 biological processes may influence stability variations: (i) changes in effective population size  
407 that modulate natural selection (for instance passing from free living to intracellular lifestyles),  
408 (ii) changes in environmental temperature, which can affect the evolutionary process (at low  
409 temperature proteins evolve more neutrally, since the relationship between the free energy and  
410 the fraction of folded protein is more sigmoidal, and therefore smaller stabilities are sufficient  
411 to fold proteins (Serohijos and Shakhnovich 2014), (iii) changes in mutation rate and in  
412 mutation bias, which can also affect the protein stability that an evolving population can  
413 achieve (Mendez et al. 2010), or most interestingly, (iv) positive selection due to changes in  
414 protein function (e.g., Pascual-Garcia et al. 2009). Such effects, including discussions of how to  
415 model them, have recently been reviewed (Anisimova and Liberles 2012; Chi and Liberles  
416 2016). Another advantage of the present framework is that it considers stability against both  
417 unfolding and misfolding, which may have evolutionary trade-offs (Mendez et al. 2010; Zheng  
418 et al. 2013). Overall, *ProtASR* is a useful tool in the phylogenetic toolbox, reflecting an

419 advance over other methods currently available as software for the important problem of  
420 ancestral sequence reconstruction.

421

## 422 **SUPPLEMENTARY MATERIAL**

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

425

## 426 **AVAILABILITY**

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

433

## 434 **FUNDING**

435 This work was supported by the Spanish Ministry of Economy and Competitiveness through the  
436 grant BFU-40020 to UB. MA was initially supported by the Spanish Government with the  
437 “Juan de la Cierva” fellowship JCI-2011-10452 and the EMBO Short Term Fellowship ASTF  
438 367-2013, and finally supported by the Portuguese Government through the FCT Starting Grant  
439 IF/00955/2014 and by the Spanish Government through the “Ramón y Cajal” fellowship RYC-  
440 2015-18241. Research at the CBMSO is facilitated by the Fundación Ramón Areces. CCW and  
441 DAL receive funding from US NSF DBI-1515704.



442

443 **ACKNOWLEDGMENTS**

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

450

451 **REFERENCES**

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

454 Alcalde M. 2015. Engineering the ligninolytic enzyme consortium. *Trends Biotechnol.* 33:155-  
455 162.

456 Anisimova M., Liberles D.A. 2012. Detecting and understanding natural selection. In:  
457 Cannarozzi G.M., Schneider A. editors. *Codon Evolution*. Oxford, Oxford University Press, p.  
458 73-96.

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

460 Arenas M., Dos Santos H.G., Posada D., Bastolla U. 2013. Protein evolution along  
461 phylogenetic histories under structurally constrained substitution models. *Bioinformatics*  
462 29:3020-3028.

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

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

467 Ashkenazy H., Penn O., Doron-Faigenboim A., Cohen O., Cannarozzi G., Zomer O., Pupko T.  
468 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic  
469 Acids Res.* 40:W580-584.

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

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

474 Bastolla U., Porto M., Roman H.E., Vendruscolo M. 2006. A protein evolution model with  
475 independent sites that reproduces site-specific amino acid distributions from the Protein Data  
476 Bank. *BMC Evol. Biol.* 6:43.

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

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

481 Chang B.S., Donoghue M.J. 2000. Recreating ancestral proteins. *Trends Ecol. Evol.* 15:109-  
482 114.

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

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

487 Doria-Rose N.A., Learn G.H., Rodrigo A.G., Nickle D.C., Li F., Mahalanabis M., Hensel M.T.,  
488 McLaughlin S., Edmonson P.F., Montefiori D., Barnett S.W., Haigwood N.L., Mullins J.I.  
489 2005. Human immunodeficiency virus type 1 subtype B ancestral envelope protein is

490 functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating  
491 subtype B envelope. *J. Virol.* 79:11214-11224.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

540 D.D., Pupko T., Regan L., Roger A., Rubinstein N., Shakhnovich E., Sjolander K., Sunyaev S.,  
541 Teufel A.I., Thorne J.L., Thornton J.W., Weinreich D.M., Whelan S. 2012. The interface of  
542 protein structure, protein biophysics, and molecular evolution. *Protein Sci.* 21:769-785.  
543 Maddison W. 1997. Gene trees in species trees. *Syst. Biol.* 46:523-536.  
544 Mallo D., Sánchez-Cobos A., Arenas M. 2016. Diverse Considerations for Successful  
545 Phylogenetic Tree Reconstruction: Impacts from Model Misspecification, Recombination,  
546 Homoplasy, and Pattern Recognition. In: Elloumi M., Iliopoulos C., Wang J., Zomaya A.  
547 editors. *Pattern Recognition in Computational Molecular Biology*, John Wiley & Sons, Inc, p.  
548 439-456.  
549 Mendez R., Fritsche M., Porto M., Bastolla U. 2010. Mutation bias favors protein folding  
550 stability in the evolution of small populations. *PLoS Comput. Biol.* 6:e1000767.  
551 Merkl R., Sterner R. 2016. Ancestral protein reconstruction: techniques and applications. *Biol.*  
552 *Chem.* 397:1-21.  
553 Minning J., Porto M., Bastolla U. 2013. Detecting selection for negative design in proteins  
554 through an improved model of the misfolded state. *Proteins* 81:1102-1112.  
555 Miyazawa S., Jernigan R.L. 1985. Estimation of effective interresidue contact energies from  
556 protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534-552.  
557 Moran P.A.P. 1958. Random processes in genetics. *Proc. Camb. Philos. Soc.* 54:60-71.  
558 Mustonen V., Lassig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics  
559 of selection and adaptation. *Trends Genet.* 25:111-119.  
560 Parisi G., Echave J. 2001. Structural constraints and emergence of sequence patterns in protein  
561 evolution. *Mol. Biol. Evol.* 18:750-756.  
562 Pascual-Garcia A., Abia D., Mendez R., Nido G.S., Bastolla U. 2009. Quantifying the  
563 evolutionary divergence of protein structures: the role of function change and function  
564 conservation. *Proteins* 78:181-196.

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

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

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

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

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

577 Warnecke T., Rocha E.P. 2011. Function-specific accelerations in rates of sequence evolution  
578 suggest predictable epistatic responses to reduced effective population size. *Mol. Biol. Evol.*  
579 28:2339-2349.

580 Wilke C.O. 2012. Bringing molecules back into molecular evolution. *PLoS Comput. Biol.*  
581 8:e1002572.

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

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

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

589 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*  
590 24:1586-1591.

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

593