*Research article*

**ProtModel: estimation of the best-fitting substitution model by approximate Bayesian computation**

David Ferreiro1,2, Catarina Branco1,2, Ugo Bastolla and Miguel Arenas1,2,3\*

1CINBIO, University of Vigo, 36310 Vigo, Spain.

2Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain.

3Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain.

**Email addresses:**

DF: [david.ferreiro.garcia@uvigo.es](mailto:david.ferreiro.garcia@uvigo.es)

CB: caraujo@uvigo.es

UB:

MA: [marenas@uvigo.es](mailto:marenas@uvigo.es)

**\* Corresponding author:**

*Miguel Arenas*

*Department of Biochemistry, Genetics and Immunology*

*University of Vigo*

*36310 Vigo, Spain*

*Email:* [marenas@uvigo.es](mailto:marenas@uvigo.es)

### *Phone: +34 986 130047*

**Running head:** Estimation of the best-fitting substitution model

**Keywords:** Structural substitution models, Empirical substitution models, Approximate Bayesian computation

**Abstract**

**Motivation:** Since the first substitution model was proposed, the development of new models and some frameworks to predict the best-fitting allowed its consolidation as an essential step in evolutionary analysis. However, these frameworks can only work considering empirical substitution models, which have been largely overcame by structurally constrained substitution (SCS) models in phylogenetic inferences due to their mathematical complexity.

**Results:** In order to facilitate the implementation of SCS models we have developed a computational framework, called *ProtModel*, to estimate the best-fitting substitution model using the approximate Bayesian computation (ABC) approach. The framework requires a protein multiple sequences alignment (MSA), a template structure and some mandatory parameters to perform the analysis and obtain an accurate estimation of the best-fitting substitution model. We applied our method to some diverse protein families, including monkeypox, coronavirus and HIV virus and conserved domains across species.

**Availability and implementation:** *ProtModel* is freely available from <https://github.com/DavidFerreiro/ProtModel>, including a version for clusters and a graphical user interface to facilitate its use. We also include an extensive documentation and some examples ready to run.

**Introduction**

Substitution models of protein evolution are routinely used to study evolutionary processes of amino acid data (Arenas 2015; Thorne 2000). They attempt to predict the rate of change of the amino acids across the entire protein sequence. However, applying an inappropriate amino acid substitution model may bias the evolutionary analyses. To prevent this, selection of the best-fitting substitution model became a crucial initial step in every evolutionary study (Posada & Crandall 2001; Lemmon & Moriarty 2004; Arbiza et al. 2011; Del Amparo & Arenas 2022).

From 1966 to 1978 Dayhoff and co-workers presented the first amino acid substitution model, using protein alignments with at least 85% identity to estimate the amino acids replacement rates (Dayhoff et al. 1978). These models produced from large protein datasets are known as classic or empirical amino acid substitution models and they consist on a 20 × 20 exchangeability matrix with the relative rates of change and the 20 amino acid frequencies (Arenas 2015). Since then some alternative empirical models were proposed using more specific or larger datasets (e.g., Blosum62 (Henikoff & Henikoff 1992), HIVw (Nickle et al. 2007), JTT (Jones et al. 1992) , LG (Le & Gascuel 2008), RtRev (Dimmic et al. 2002), VT (Müller & Vingron 2000), WAG (Whelan & Goldman 2001)) and considering new variables (e.g., the proportion of invariable sites (+I) (Shoemaker & Fitch 1989) and the rate of variation across sites (+G) (Yang 1994)) which increased the efficiency of the models. Conveniently, those models remained mathematically simple which allowed us to implement maximum likelihood (ML) approaches in evolutionary frameworks for phylogenetic studies (such as tree reconstructions (e.g., Kozlov et al. 2019), ancestral sequence reconstructions, (e.g., Yang 2007; Arenas et al. 2017; Arenas & Bastolla 2019) or estimation of the best-fitting substitution model (e.g., Darriba et al. 2011, 2020; Keane et al. 2006)). Nevertheless, even though empirical models can model heterogenous evolution (Zoller & Schneider 2013), they still assume that all protein sites mutate following the same evolutionary patterns (site-independence evolution) which may be unrealistic. Indeed, residues from protein alpha-helices or from protein surface exhibit different substitution rate patterns than those located in other protein regions (e.g., beta-sheet or protein core) (Perron et al. 2019). Moreover, the protein structure induces specific interactions between amino acids at different sites (Shakhnovich et al. 1996), hence the replacement of a given amino acid affects the 3D interactions among nearby residues (especially for replacements involving residues with distinct steric properties). Conveniently, some techniques enable us to study protein folding and so increase our understanding about how it evolved (e.g., homology modelling, docking, molecular dynamics simulations, among others). Particularly, solvent accessibility or constrains in the protein folding are considered to generate structural constrain substitution (SCS) models of protein evolution. SCS models were successfully included in evolutionary frameworks that yielded accurate inferences of protein evolution in comparison with empirical models (Arenas & Bastolla 2019; Arenas et al. 2017, 2015, 2013; García-Portugués et al. 2018; Challis & Schmidler 2012; Golden et al. 2017; Herman et al. 2014; Perron et al. 2019; Norn et al. 2021). Some SCS models aim to estimate the stability of a protein in its native state, assuming that each mutation do not change the native residues contacts (Bastolla & Arenas 2019; Arenas et al. 2015). They use pairwise contact interactions (amino acids co-evolution) to estimate protein free energy (stability) (Bastolla et al. 2000), requiring probabilistic relations between different amino acids and therefore greatly increasing the mathematical complexity of the estimation. Consequently, evolutionary frameworks which use ML functions (e.g., phylogenetic tree reconstruction~~, ancestral sequence reconstruction~~ or estimation of the best-fitting substitution model) cannot incorporate SCS models. Conveniently, the approximate Bayesian computation (ABC) approach (Beaumont 2010; Beaumont et al. 2002; Csilléry et al. 2012; Sunnåker et al. 2013) is an alternative to ML and enable us to work with SCS models. Our group had previously developed an ABC framework to estimate different evolutionary parameters (e.g., recombination or substitution rate) which outperformed other methods (Arenas 2021). Furthermore, ABC approach was successfully used to model estimation (e.g., Bemmels et al. 2016; Arenas et al. 2020; Leuenberger & Wegmann 2010) and demonstrated to be a trustworthy method for evolutionary analyses.

Here, we present a new methodology implemented in a user-friendly computational framework called *ProtModel* based on the ABC approach. *ProtModel* can estimate the best-fitting substitution model from protein multiple sequences alignment (MSA) under ABC. It uses an adapted version of the simulator *ProteinEvolver* (Arenas et al. 2013) to perform the computer simulations of the evolutionary history of the protein MSA (Kingman 1982) followed by protein evolution under the considered substitution models (empirical or SCS). Regarding SCS models, *ProtModel* incorporates Neutral and Fitness mean-field substitution models (Arenas et al. 2015). They compute site specific amino acid frequencies considering independent site evolution combined with contact-based models of protein folding (Arenas et al. 2015). Concerning empirical models, many alternatives are available: Blosum62 (Henikoff & Henikoff 1992), CpRev (Adachi et al. 2000), Dayhoff (Dayhoff et al. 1978), DayhoffDCMUT (Kosiol & Goldman 2005), HIVb (Nickle et al. 2007), HIVw (Nickle et al. 2007), JTT (Jones et al. 1992), JonesDCMUT (Kosiol & Goldman, 2005), LG (Le & Gascuel 2008), Mtart (Abascal et al. 2006), Mtmam (Yang et al. 1998), Mtrev24 (Adachi & Hasegawa), RtRev (Dimmic et al. 2002), VT (Müller & Vingron 2000), WAG (Whelan & Goldman 2001) and UserEAAM models (Arenas 2015)). It works with several fixed and nuisance parameters to provide more realistic simulations, and seven summary statistics (SS) designed to extract the evolutionary information from the protein sequences. Finally, *ProtModel* coupled the simulations with an ABC method to estimate the best-fitting substitution model. To prove the utility of the framework, we applied the method to different protein families of different organisms.

**System and methods**

*ProtModel* framework

The evolutionary framework *ProtModel* estimates the best-fitting substitution model for a given protein MSA through four main steps: (1) specification of input information including the prior distribution of the substitution rate, (2) simulation of protein sequences evolution upon different substitution models, (3) computation of SS and (4) estimation of the best-fitting substitution model using ABC approach (Figure 1). Details about these steps are provided below.

1. Before implement the ABC approach the user must take some decisions related with the evolutionary and demographic parameters of the simulations (e.g., the population size or prior distribution of the substitution rate), and some nuisance parameters (e.g., amino acid frequencies, heterogeneity in the substitution rate across sites and proportion of invariable sites) that can influence the accuracy of the simulations (Table S1, Supplementary Data). The applied prior distributions should be wide to increase the probability of include parameter values similar to the true value (Beaumont 2010). Next, the user needs to specify the substitution models of protein evolution that will be evaluated (Table S1). We strongly recommend to include Fitness and Neutral SCS models, and the empirical substitution model that best fits with the data. The later can be identified using frameworks such as *ProtTest, ModelTest* or *ModelGenerator* (Darriba et al. 2011, 2020; Keane et al. 2006). Finally, regarding the ABC estimation the user has to take some decisions, such as the tolerance (retained simulations that show SS closer to the observed SS), the ABC method (i.e., rejection, multiple linear regression or neural networks based [*rejection, mnlogistic* or *neuralnet,* respectively]) (Beaumont et al. 2002) or the number of iterations (the number of simulations that are considered as pseudo-observed data) (Table S1). Check framework documentation to more information for parameters specification.
2. The simulations of protein MSA are performed with a version of *ProteinEvolver* simulator (Arenas et al. 2013) adapted to ABC (Arenas 2021). It consists in two main steps: (*i*) firstly, it simulates the evolutionary history of the protein MSA with the coalescent method (Kingman 1982) and, (*ii*) then it simulates the evolution of the protein sequence upon the previously simulated evolutionary histories under the specified substitution model of protein evolution. Coalescent method allows rapid simulations (Arenas, 2012), but SCS models simulations require a lot of time. Conveniently, the simulations can run in parallel on multicore computers or clusters.
3. *ProtModel* uses 7 SS summarizing the evolutionary information from the observed protein MSA and from the protein sequences simulated under every substitution model (Table S2, Supplementary Data). In general, they either explore the stability of the individual sequences of the MSA or the physicochemical properties of the replacements of the amino acids. To explore the stability, we include the mean and the standard deviation of the Gibbs free energy difference between folded and unfolded states (ΔG kcal/mol) calculated with *DeltaGREM* software (Arenas et al. 2017) (Table S2). Regarding the amino acid replacements and their physicochemical properties, we incorporate the number of segregating sites and the mean, standard deviation, skewness and kurtosis of Grantham physicochemical distance (Grantham 1974) between protein sites (Table S2). It is a measure that assesses the difference between replaced amino acids depending on their composition, polarity and molecular volume properties.
4. Finally, *ProtModel* estimates the posterior probability of all the substitution models analysed with the ABC R library (Csilléry et al. 2012). Before the estimation, some informative outputs are created: (*i*) A histogram with the substitution rate per site considered in every simulation drawn from the prior distribution. Together, another plot represents the corresponding theta (*θ*) values from the prior distribution used in every simulation. (*ii*) Boxplots of the calculated SS values for each considered substitution model with a blue line representing the values calculated for the observed dataset. (*iii*) A plot representing the confusion matrix (i.e., a specific table that allows the visualization of the performance of the distinction between models). These plots can help the user to better understand the observed data, including help the user to predict the results. Next, the estimation of the posterior probability can be made using three different methods (*rejection*, *mnlogistic* and *neuralnet*)) (Beaumont 2010; Beaumont et al. 2002; Blum & François 2010; Csilléry et al. 2012). The confusion matrix, the estimated best-fitting substitution model and the goodness of fit of the real data with the specified substitution models will be written in the “*Results\_text.txt*” file. Additionally, to easily understand the selected best-fitting substitution model we also provide: (*iv*) A histogram of every investigated substitution model representing the median of the distance between accepted summary statistics and observed ones. The p-value is also computed to test the fit of every substitution model. (*v*) A histogram of the values of each SS from the retained simulations for every substitution model evaluated. (*vi*) A scatter plot of the values of the SS and the corresponding values of *θ* used for the simulation. (*vii*) A plot of the two first principal components of a principal component analysis (PCA) of the SS values of the considered substitution models.

Overall, *ProtModel* allows to perform ABC analyses in a simple way, only taking some decisions that can affect the model selection. It consists of a pipeline written in Python, C and R that can run either on a local computer (on the command line or using a graphical user interface (GUI)) or on a cluster. It is freely available from <https://github.com/DavidFerreiro/ProtModel>. Together with the framework, some documentation and illustrative examples with input and output file are included.

*ProtModel* validation

Model selection using ABC method was already tested in population genetics (e.g., (Bemmels et al. 2016; Arenas et al. 2020; Leuenberger & Wegmann 2010; Branco et al. 2022), but it was never used to compare amino acid substitution models. Here, we performed an evaluation of *ProtModel* ABC method, testing its effectiveness under different values of ABC-related parameters: (*i*) number of simulations (10,000, 50,000 and 100,000), (*ii*) tolerance (0.005, 0.01, and 0.05) and (*iii*) under three different methods (*rejection*, *mnlogistic* and *neuralnet*). We performed the analysis considering three substitution models, specifically the empirical model Dayhoff (Dayhoff et al. 1978), and the SCS models Fitness and Neutral (Arenas et al. 2015), using an observed dataset of 27 sequences and the 1TDE structure of the thioredoxin reductase protein family (Waksman et al. 1994). For every possible combination (3 models × 3 sets of simulations × 3 methods = 27 combinations), we evaluated the power of *ProtModel* to distinguish between the three substitution models with a cross-validation based on 100 pseudo-observed simulations (Csilléry et al. 2012). Overall, we found that the number of simulations did not influence the results, contrarily to the tolerance value. Specifically, increasing the tolerance value decreased the accuracy of the estimation. Regarding the *mnlogistic* and *neuralnet* ABC methods, both failed under the considered number of simulations and with any value of tolerance.

Next, we simulated 100 multiple protein MSA under each substitution model considered (Dayhoff, Fitness and Neutral), which we considered as test datasets. Then, the framework was used to estimate the best-fitting substitution model of the 300 test datasets (100 per model) considering 10,000, 50,000 and 100,000 simulations parameterized under a *θ* uniform prior distribution (0, 500), which encompass values that are commonly observed in real data (e.g., Carvajal-Rodriguez et al., 2006; Monteiro et al., 2021; Perez-Losada et al., 2011; Perez-Losada et al., 2009; Stumpf and McVean, 2003). As in previous studies (Arenas et al. 2020; Branco et al. 2022), we made a cross-validation of the ABC rejection method based on 100 pseudo-observed simulations with a tolerance of 0,5%, 1% and 5% (Csilléry et al. 2012) using the three ABC methods. Again, considering the *mnlogistic* or *neuralnet* methods some of the 300 estimations failed. Note that when the estimation using these methods fails the *rejection* method is used. *ProtModel* validation showed that the best-fitting substitution model estimation are generally accurate under the *rejection* method (Figure 2), being able to distinguish between empirical and SCS models, as well as between SCS models. The estimations effectiveness increases with decreasing the tolerance, but they are almost invariable with the number of simulations. Indeed, 10,000 simulations of each model were enough to obtain accurate results (Figure 2).

*Illustrative examples of application to different protein families*

We applied *ProtModel* to 8 datasets of different protein families (Table 1). We select some interesting proteins related with the viral pathogenic of current pandemics (specifically, HIV-1 PR, HIV-1 GAG, influenza NS1, SARS-CoV endopeptidase C30 and SARS-CoV 2'-O-methyltransferase) as well as the tumour necrosis factor receptor (TNF) of monkeypox virus (that recently widely spread to humans). We also applied our framework to two evolutionary conserved protein domains, the Calcium-binding EGF and the intracellular signalling Toll-Interleukin (Table 1). All these proteins have different amino acid length (from 99 to 299), and we used a different number of sequences for every MSA (from 10 to 50) to represent a wide range of proteins MSA. The majority of the protein sequences were downloaded from the PFAM database (Mistry et al. 2021) representing the whole protein family (lower identity), except the monkeypox TNF and the HIV-1 PR and GAG that were downloaded from the PopSet NCBI database (Sayers et al. 2022), representing a population (higher identity). Then, for every dataset the sequences were re-aligned using *MUSCLE* (Edgar 2004) and the wild type (WT) sequence was created using the most common amino acid in every position using a python script (*FindWT.py* provided to the user). Taking advance of the WT sequences, we searched in the *SWISS-MODEL* (Arnold et al. 2006) database the corresponding templates, the structure that best represents the WT sequences and, therefore, the protein MSA, and we downloaded them from the Protein Data Bank (Berman 2000) (Table 1). Next, we included the templates of the proteins sequences in their corresponding MSA and they were again re-aligned using MUSCLE (Edgar 2004). To perform protein evolution simulations under SCS models and for calculate the protein free energy we need the sequences of the input alignment to have the same length as the template sequence. Hence, we removed every position of the MSA which did not match with the corresponding template sequence. Finally, we removed the template sequences and obtained the final input protein MSA. These steps after the template download can be done using the *Align.py* provided to the user, requiring only that MUSCLE (Edgar 2004) is installed in the computer.

We analyzed the best-fitting substitution model of protein evolution between the best-fitting empirical substitution models according *ProtTest* framework and the two *mean-field* SCS models (Neutral and Fitness) (Table 1). We ran 10,000 simulations per model using different prior distributions for the substitution rate per site depending on the analysis (Table 1). Choosing a prior distribution too wide or too narrow can affect the results (Beaumont 2010). In particular, the SS related with the amino acids replacements will have values far from the observed SS values. The first example of the monkeypox TNF (Table 1) illustrates the importance of choosing appropriate prior distributions. Here, narrowing the prior distribution changed the estimated best-fitting substitution model and the goodness of fit (Table S3, Supplementary Data). Regarding the SS, we generally recommend the user to calculated all of them, however the DGREM\_sd (which measures the standard deviation of the sequences folding stability of each MSA) could be problematic. In some cases, the stability of the simulated sequences could be very different among them, resulting in high standard deviation and therefore far from the observed SS. Consequently, this may decreased the power of ABC to distinguish among the models, and cause it to lose robustness, affecting the estimation of the best-fitting substitution model. We exemplify this weakness with the SARS-CoV endopeptidase C30 (Table 1), where using or not the the DGREM\_sd (SS ID 2) dramatically affects the estimation and the goodness of fit (Table S3). Finally, we also tested of *mnlogistic* and *neuralnet* methods using TNF family. In this case, we were able to run the framework without any error and we obtain the same best-fitting substitution model with the three methods. For all these analyses we used a tolerance value of 0.5% or 10%, and 100 pseudo-observed simulations for cross-validation. We consider that we obtained good predictions about the best-fitting substitution models for the different protein families (Table 1) supported by the p-values of the goodness of fit (Table S3).

Users can run *ProtModel* using the command line, the GUI or the cluster version. Simulations under a SCS model and the calculation of the free energy may take several hours, thus we recommend choosing the lowest number of simulations. However, simulation time also depends on the length and the number of the protein sequences. Concerning the substitution models, the computer time does not depend on which empirical or SCS model of protein evolution is used. All of the empirical models have the same dimension (20 ×20) (Arenas, 2015; Yang, 2006) and the two SCS models will both have as many matrices as number of sites along the sequences. Still, of course the execution time will increase with the number of substitution models. Both command line and GUI version can use more than one computer core, running the simulations in a parallelized way but sharing computer memory. Thus, even though using more than one core may reduce the computer time, it will not follow a linear function (Figure S1; Supplementary Data). Indeed, the execution time will mainly depend on the computer activity while the program is running. By contrast, cluster version can work with a large number of cores without sharing computer memory, but the computer time and the number of processors also does not follow a linear function (Figure S1) since a new process only starts after the cores have finished their process. Finally, the time of ABC estimation depends on the number of simulations per model.

**Discussion**

The first SCS model was proposed in 2001 by Parisi & Echave (2001). Since then, updates and models were proposed to overcome the limitations of the first model (Robinson 2003; Rodrigue et al. 2009; Bonnard et al. 2009), yet their implementation in evolutionary frameworks remains a difficult task due to the involved complex mathematics. Particularly, it is unfeasible to apply the likelihood function in an evolutionary framework to estimate the best-fitting substitution model considering these models. Even so, we consider that there is a need for such a framework. Aside from the importance of applying an accurate substitution model in an evolutionary analysis to obtain realistic results (Del Amparo & Arenas 2022), to consider SCS models increases the veracity of evolutionary studies and their outcome (e.g., Arenas & Bastolla 2019; Arenas et al. 2017, 2015, 2013; García-Portugués et al. 2018; Challis & Schmidler 2012; Golden et al. 2017; Herman et al. 2014; Perron et al. 2019; Norn et al. 2021). In that concern, we implement in our framework an ABC procedure as a trustworthy alternative to select between evolutionary models using simulations and the derived SS (e.g., Bemmels et al. 2016; Arenas et al. 2020; Leuenberger & Wegmann 2010; Branco et al. 2022). Indeed, to estimate the best-fitting substitution model with ABC we need simulations, which we performed under empirical and SCS models using *ProteinEvolver* simulator (Arenas et al. 2013). The usefulness of this simulator was previously validated in estimations of recombination and substitution rate for protein MSA (Arenas 2021). However, unlike classic best-substitution model estimators which only require the MSA, to run *ProtModel* the user has to provide some information (Table S1). Combining some mandatory details (e.g., the substitution models to analyze, the SS to consider or the substitution rate) and a wide range of optional parameters the user can produce realistic estimations (Table S1). To simplify the required input information, we recommend to carefully read the manual of the framework and to modify the examples input files or to use GUI. Nonsense or not accurate input information may bias the coalescent evolutionary history of the simulations (e.g. provide unrealistic substitution rates or not consider enough simulations), affecting the framework efficiency. For some parameters, such as the substitution rate per site, it could be difficult to obtain discrete accurate values, in which case *ProtModel* includes the substitution rate prior distribution and some nuisance parameters. Concerning the prior distribution, we recommend using uniform distributions within biologically reasonable values when the user does not have previous knowledge. As in any other evolutionary framework, we recommend cautious dealing with the results and to check the outputs files for the estimations. Indeed, we observed in the first example of the monkeypox TNF factor (Table 1) that choosing an inappropriate prior distribution produced a wide deviation on the best-fitting substitution model estimation (Table 1; Table S3). We also recommend attention on the ABC settings and to repeat the analyses with multiple ABC methods, tolerance or number of iterations. Indeed, in some cases it can be necessary to repeat the analysis considering a higher number of simulations or different SS to obtain reliable results. In the provided example of the SARS-CoV endopeptidase C30 (Table 1) we observe that considering the protein stability standard deviation SS (DGREM\_sd) affects drastically the estimation (Table 1; TableS3).

We are aware that ABC analysis require high computational efforts and sometimes multiple analysis, thus we develop a cluster version which decreases drastically (more than 10 times) the execution time (Figure S1). Still, even when the user cannot access to a cluster, we strongly believe that the accuracy of the estimations and the improvements that SCS models yield on the evolutionary analysis justify the computing time. For the illustrative examples we selected some interesting protein families. First, we analyzed the monkeypox TNF receptor which is involve in tumorigenesis and viral replication inhibition (Al Rumaih et al. 2020). Recent studies found that this receptor is able to maintain its structure despite mutation (Benvenuto et al. 2022) (preprint, maybe herbet), so we expected that a SCS model would be the best-fitting substitution model for the data. Indeed, we found that the empirical substitution model HIVw is quite far from the real data (Table 1), supporting the use of SCS models in TNF evolutionary analysis. In addition, we explored some interesting therapeutic targets proteins (SARS-CoV endopeptidase C30, SARS-CoV 2'-O-methyltransferase, influenza NS1, HIV-1 GAG and HIV-1 PR). Generally, we found that the best-fitting substitution model estimation corresponds with a SCS model. The function of all these proteins is to interact with some other proteins or substrate, so it is likely that they have to maintain their structure to properly functionate. HIV protease is an extraordinary example of accumulate mutations maintaining its structure (Wu et al. 2003). Looking closer at the HIV gag polyprotein, we see that both SCS models fit with the data. This should not be considered an estimation failure since we were able to exclude the empirical model that resulted from the *ProtTest* analysis. Fitness and Neutral are similar models with few distinctions (e.g., considering or ignoring the population size, respectively), so we believe that both models could produce accurate results in evolutionary analysis. Finally, we analyzed two conserved domains, calcium-binding EGF and the intracellular signalling Toll-Interleukin, widely distributed in different organisms proteins (Yáñez et al. 2012; Bayless & Nishimura 2020). Thus, we expect them to have a very conserved protein structure and so their best-fitting substitution model would be a SCS model. However, we only found that for the Toll-Interleukin receptor, where the WAG model is very far from the real data. Regarding the EFG domain, we obtained that the best-fitting substitution model was the Blosum62 empirical model. Notice that the fact that SCS models have overcome empirical models in some evolutionary studies does not mean that a SCS model is always the best-fitting substitution model. *ProtModel* needs a template structure to represent the whole MSA, so alignments with low sequences identity may not be properly represented by the template structure, in which case the SS obtained from simulations under SCS models may be far away from the real data. Similarly, although a wide range of different protein structures are available, some MSA or proteins do not present a reasonably good template so, again simulations under SCS models may be far from the real data.

We conclude that *ProtModel* is a trustworthy tool to estimate the best-fitting substitution model of protein evolution. Although carrying out a *ProtModel* analysis will require some extra effort comparing with traditional frameworks, we believe that the accurate estimations considering SCS models justify its use.

**Data availability**

**Supplementary data**

Supplementary data is available at the journal online.

**Acknowledgments**

This work was supported by the Spanish Ministry of Economy and Competitiveness through the Grants [RYC-2015-18241] and by the Xunta de Galicia through the Grants [ED431F 2018/08] and [ED481A-2020/192].

**Conflicts of Interest**

None declared.

**References**

Abascal F, Posada D, Zardoya R. 2006. MtArt: A New Model of Amino Acid Replacement for Arthropoda. Mol. Biol. Evol. 24:1–5. doi: 10.1093/molbev/msl136.

Adachi J, Hasegawa M. Programs for Molecular Phylogenetics Based on Maximum Likelihood. 150.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid Genome Phylogeny and a Model of Amino Acid Substitution for Proteins Encoded by Chloroplast DNA. J. Mol. Evol. 50:348–358. doi: 10.1007/s002399910038.

Al Rumaih Z et al. 2020. Poxvirus-encoded TNF receptor homolog dampens inflammation and protects from uncontrolled lung pathology during respiratory infection. Proc. Natl. Acad. Sci. 117:26885–26894. doi: 10.1073/pnas.2004688117.

Arbiza L, Patricio M, Dopazo H, Posada D. 2011. Genome-Wide Heterogeneity of Nucleotide Substitution Model Fit. Genome Biol. Evol. 3:896–908. doi: 10.1093/gbe/evr080.

Arenas M. 2021. ProteinEvolverABC: coestimation of recombination and substitution rates in protein sequences by approximate Bayesian computation Schwartz, R, editor. Bioinformatics. 38:58–64. doi: 10.1093/bioinformatics/btab617.

Arenas M et al. 2020. The Early Peopling of the Philippines based on mtDNA. Sci. Rep. 10:4901. doi: 10.1038/s41598-020-61793-7.

Arenas M. 2015. Trends in substitution models of molecular evolution. Front. Genet. 6. doi: 10.3389/fgene.2015.00319.

Arenas M, Bastolla U. 2019. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. Methods Ecol. Evol. 11:248–257. doi: 10.1111/2041-210X.13341.

Arenas M, Dos Santos HG, Posada D, Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics. 29:3020–3028. doi: 10.1093/bioinformatics/btt530.

Arenas M, Sánchez-Cobos A, Bastolla U. 2015. Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. Mol. Biol. Evol. 32:2195–2207. doi: 10.1093/molbev/msv085.

Arenas M, Weber CC, Liberles DA, Bastolla U. 2017. ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. Syst. Biol. 66:1054–1064. doi: 10.1093/sysbio/syw121.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 22:195–201. doi: 10.1093/bioinformatics/bti770.

Bastolla U, Arenas M. 2019. The Influence of Protein Stability on Sequence Evolution: Applications to Phylogenetic Inference. In: Computational Methods in Protein Evolution. Sikosek, T, editor. Methods in Molecular Biology Vol. 1851 Springer New York: New York, NY pp. 215–231. doi: 10.1007/978-1-4939-8736-8\_11.

Bastolla U, Vendruscolo M, Knapp E-W. 2000. A statistical mechanical method to optimize energy functions for protein folding. Proc. Natl. Acad. Sci. 97:3977–3981. doi: 10.1073/pnas.97.8.3977.

Bayless AM, Nishimura MT. 2020. Enzymatic Functions for Toll/Interleukin-1 Receptor Domain Proteins in the Plant Immune System. Front. Genet. 11:539. doi: 10.3389/fgene.2020.00539.

Beaumont MA. 2010. Approximate Bayesian Computation in Evolution and Ecology. Annu. Rev. Ecol. Evol. Syst. 41:379–406. doi: 10.1146/annurev-ecolsys-102209-144621.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian Computation in Population Genetics. Genetics. 162:2025–2035. doi: 10.1093/genetics/162.4.2025.

Bemmels JB, Title PO, Ortego J, Knowles LL. 2016. Tests of species‐specific models reveal the importance of drought in postglacial range shifts of a Mediterranean‐climate tree: insights from integrative distributional, demographic and coalescent modelling and ABC model selection. Mol. Ecol. 25:4889–4906. doi: 10.1111/mec.13804.

Benvenuto D et al. 2022. *The evolution of Monkeypox virus: a genetic and structural analysis reveals mutations in proteins involved in host-pathogen interaction*. Bioinformatics doi: 10.1101/2022.06.22.497195.

Berman HM. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235–242. doi: 10.1093/nar/28.1.235.

Blum MGB, François O. 2010. Non-linear regression models for Approximate Bayesian Computation. Stat. Comput. 20:63–73. doi: 10.1007/s11222-009-9116-0.

Bonnard C, Kleinman CL, Rodrigue N, Lartillot N. 2009. Fast optimization of statistical potentials for structurally constrained phylogenetic models. BMC Evol. Biol. 9:227. doi: 10.1186/1471-2148-9-227.

Branco C, Kanellou M, González-Martín A, Arenas M. 2022. Consequences of the Last Glacial Period on the Genetic Diversity of Southeast Asians. Genes. 13:384. doi: 10.3390/genes13020384.

Challis CJ, Schmidler SC. 2012. A Stochastic Evolutionary Model for Protein Structure Alignment and Phylogeny. Mol. Biol. Evol. 29:3575–3587. doi: 10.1093/molbev/mss167.

Csilléry K, François O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC): *R package: abc*. Methods Ecol. Evol. 3:475–479. doi: 10.1111/j.2041-210X.2011.00179.x.

Darriba D et al. 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models Crandall, K, editor. Mol. Biol. Evol. 37:291–294. doi: 10.1093/molbev/msz189.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 27:1164–1165. doi: 10.1093/bioinformatics/btr088.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure.Vol. 5 National Biomedical Research Foundation: Washington DC pp. 345–352.

Del Amparo R, Arenas M. 2022. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction Chang, B, editor. Mol. Biol. Evol. 39:msac144. doi: 10.1093/molbev/msac144.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny. J. Mol. Evol. 55:65–73. doi: 10.1007/s00239-001-2304-y.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. doi: 10.1093/nar/gkh340.

García-Portugués E et al. 2018. *Toroidal diffusions and protein structure evolution*. doi: 10.1201/9781315228570.

Golden M et al. 2017. A Generative Angular Model of Protein Structure Evolution. Mol. Biol. Evol. 34:2085–2100. doi: 10.1093/molbev/msx137.

Grantham R. 1974. Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 185:862–864. doi: 10.1126/science.185.4154.862.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89:10915–10919. doi: 10.1073/pnas.89.22.10915.

Herman JL, Challis CJ, Novák Á, Hein J, Schmidler SC. 2014. Simultaneous Bayesian Estimation of Alignment and Phylogeny under a Joint Model of Protein Sequence and Structure. Mol. Biol. Evol. 31:2251–2266. doi: 10.1093/molbev/msu184.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 8:275–282. doi: 10.1093/bioinformatics/8.3.275.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6:29. doi: 10.1186/1471-2148-6-29.

Kingman JFC. 1982. The coalescent. Stoch. Process. Their Appl. 13:235–248. doi: 10.1016/0304-4149(82)90011-4.

Kosiol C, Goldman N. 2005. Different Versions of the Dayhoff Rate Matrix. Mol. Biol. Evol. 22:193–199. doi: 10.1093/molbev/msi005.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference Wren, J, editor. Bioinformatics. 35:4453–4455. doi: 10.1093/bioinformatics/btz305.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol. Biol. Evol. 25:1307–1320. doi: 10.1093/molbev/msn067.

Lemmon AR, Moriarty EC. 2004. The Importance of Proper Model Assumption in Bayesian Phylogenetics Sullivan, J, editor. Syst. Biol. 53:265–277. doi: 10.1080/10635150490423520.

Leuenberger C, Wegmann D. 2010. Bayesian Computation and Model Selection Without Likelihoods. Genetics. 184:243–252. doi: 10.1534/genetics.109.109058.

Mistry J et al. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49:D412–D419. doi: 10.1093/nar/gkaa913.

Müller T, Vingron M. 2000. Modeling Amino Acid Replacement. J. Comput. Biol. 7:761–776. doi: 10.1089/10665270050514918.

Nickle DC et al. 2007. HIV-Specific Probabilistic Models of Protein Evolution Pybus, O, editor. PLoS ONE. 2:e503. doi: 10.1371/journal.pone.0000503.

Norn C, André I, Theobald DL. 2021. A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices. Protein Sci. 30:2057–2068. doi: 10.1002/pro.4155.

Parisi G, Echave J. 2001. Structural Constraints and Emergence of Sequence Patterns in Protein Evolution. Mol. Biol. Evol. 18:750–756. doi: 10.1093/oxfordjournals.molbev.a003857.

Perron U, Kozlov AM, Stamatakis A, Goldman N, Moal IH. 2019. Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States Pupko, T, editor. Mol. Biol. Evol. 36:2086–2103. doi: 10.1093/molbev/msz122.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. 98:13757–13762. doi: 10.1073/pnas.241370698.

Robinson DM. 2003. Protein Evolution with Dependence Among Codons Due to Tertiary Structure. Mol. Biol. Evol. 20:1692–1704. doi: 10.1093/molbev/msg184.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational Methods for Evaluating Phylogenetic Models of Coding Sequence Evolution with Dependence between Codons. Mol. Biol. Evol. 26:1663–1676. doi: 10.1093/molbev/msp078.

Sayers EW et al. 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res. 50:D20–D26. doi: 10.1093/nar/gkab1112.

Shakhnovich E, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. Nature. 379:96–98. doi: 10.1038/379096a0.

Shoemaker JS, Fitch WM. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289. doi: 10.1093/oxfordjournals.molbev.a040550.

Sunnåker M et al. 2013. Approximate Bayesian Computation Wodak, S, editor. PLoS Comput. Biol. 9:e1002803. doi: 10.1371/journal.pcbi.1002803.

Thorne JL. 2000. Models of protein sequence evolution and their applications. Curr. Opin. Genet. Dev. 10:602–605. doi: 10.1016/S0959-437X(00)00142-8.

Waksman G, Krishna TS, Williams CH, Kuriyan J. 1994. Crystal structure of Escherichia coli thioredoxin reductase refined at 2 A resolution. Implications for a large conformational change during catalysis. J. Mol. Biol. 236:800–816.

Whelan S, Goldman N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol. Biol. Evol. 18:691–699. doi: 10.1093/oxfordjournals.molbev.a003851.

Wu TD et al. 2003. Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatments. J. Virol. 77:4836–4847. doi: 10.1128/JVI.77.8.4836-4847.2003.

Yáñez M, Gil-Longo J, Campos-Toimil M. 2012. Calcium Binding Proteins. In: Calcium Signaling. Islam, MdS, editor. Advances in Experimental Medicine and Biology Vol. 740 Springer Netherlands: Dordrecht pp. 461–482. doi: 10.1007/978-94-007-2888-2\_19.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314. doi: 10.1007/BF00160154.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591. doi: 10.1093/molbev/msm088.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15:1600–1611. doi: 10.1093/oxfordjournals.molbev.a025888.

Zoller S, Schneider A. 2013. Improving Phylogenetic Inference with a Semiempirical Amino Acid Substitution Model. Mol. Biol. Evol. 30:469–479. doi: 10.1093/molbev/mss229.

**Tables**

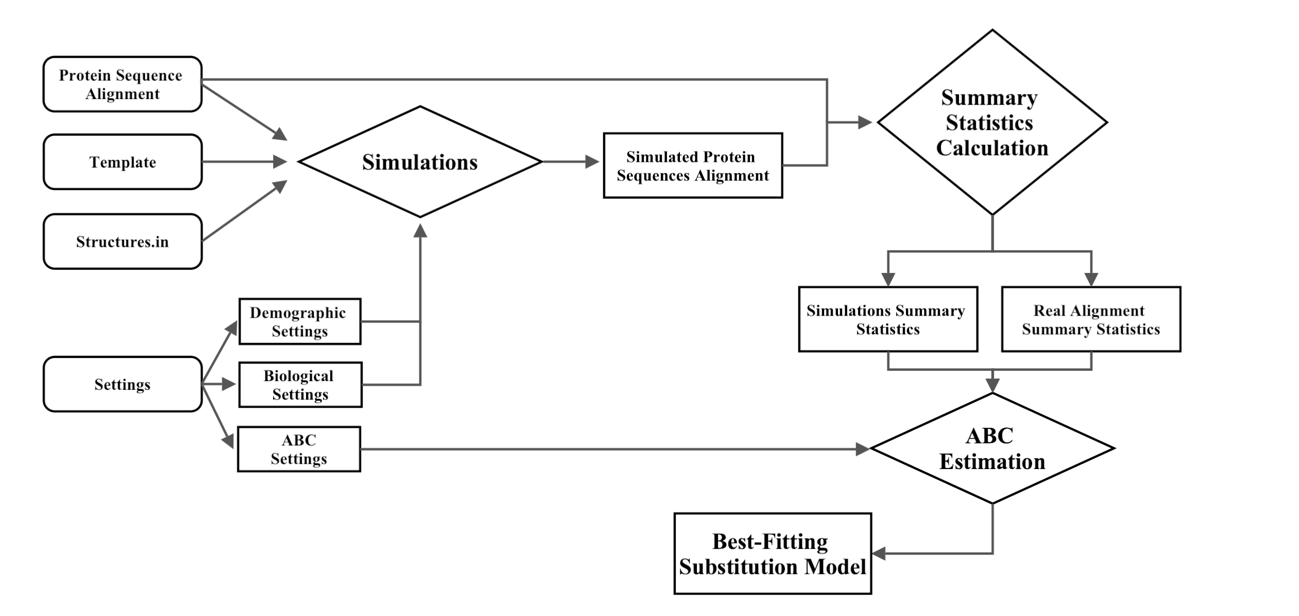
**Table 1. Best-fitting substitution model probabilities estimated with *ProtModel* for the studied protein families.** The table shows the results and some important parameters for the best-fitting substitution model estimation. The Summary Satistics ID is identified in the Table S2, Supplementary Data.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Protein family description* | *Number and length of sequences* | *Template* | *Substitution rate prior* | *Theta prior* | *Summary Statistics ID* | *Nº Simu x model* | *Emp Subs Model* | *ABC Method* | *ABC estimation* | |
| Monkeypox tumour necrosis receptor | 10 sequences  160 amino acid | 3on9 | Uniform  0 7.8125e-04 | Uniform  0 500 | 1 2 3 4 5 6 7 | 1,000 | HIVw | rejection | **Fitness** HIVw Neutral  **0.4333** 0.1667 0.4 | |
| Monkeypox tumour necrosis receptor | 10 sequences  160 amino acid | 3on9 | Uniform  0 1.8750e-04 | Uniform  0 120 | 1 2 3 4 5 6 7 | 1 000 | HIVw | rejection | Fitness HIVw **Neutral**  0.2333 0.1 **0.6667** | |
| Monkeypox tumour necrosis receptor | 10 sequences  160 amino acid | 3on9 | Uniform  0 1.8750e-04 | Uniform  0 120 | 1 2 3 4 5 6 7 | 1 000 | HIVw | mnlogistic | | Fitness HIVw **Neutral**  0 0 **1** |
| Monkeypox tumour necrosis receptor | 10 sequences  160 amino acid | 3on9 | Uniform  0 1.8750e-04 | Uniform  0 120 | 1 2 3 4 5 6 7 | 1 000 | HIVw | neuralnet | | Fitness HIVw **Neutral**  0.0333 0.0051 **0.9616** |
| HIV protease (PR) | 50 sequences,  99 amino acid | 1tcx | Uniform 0 2.5253e-04 | Uniform  0 100 | 1 2 3 4 5 6 7 | 10 000 | JTT | rejection | **Fitness** JTT Neutral  **0.8533** 0.0667 0.08 | |
| HIV gag polyprotein | 27 sequences 288 amino acid | 1l6n | Uniform 0 1.041667e-04 | Uniform 0 120 | 1 2 3 4 5 6 7 | 10 000 | JTT | rejection | Fitness HIVb **Neutral**  0.4333 0.1067 **0.4600** | |
| Influenza NS1 | 25 sequences  202 amino acid | 4OPH | Uniform 0 2.4752e-04 | Uniform  0 200 | 1 2 3 4 5 6 7 | 10 000 | JTT | rejection | Fitness JTT **Neutral**  0.0067 0.22 **0.7733** | |
| Coronavirus endopeptidase C30 | 30 sequences  299 amino acid | 1LVO | Uniform 0 4.180602e-04 | Uniform 0 500 | 1 2 3 4 5 6 7 | 10 000 | LG | rejection | Fitness **LG** Neutral  0 **1** 0 | |
| Coronavirus endopeptidase C30 | 30 sequences  299 amino acid | 1LVO | Uniform 0 4.180602e-04 | Uniform 0 500 | 1 3 4 5 6 7 | 10 000 | LG | rejection | **Fitness** LG Neutral  **0.88** 0.0133 0.1067 | |
| Coronavirus 2'-O-methyltransferase | 28 sequences  298 amino acid | 7c2i | Uniform 0 4.194631e-04 | Uniform 0 500 | 1 2 3 4 5 6 7 | 10 000 | LG | rejection | **Fitness** LG Neutral  **0.6733** 0.1267 0.2 | |
| Calcium-binding EGF domain | 18 sequences  263 amino acid | 6pog | Uniform 0 2.118644e-04 | Uniform 0 100 | 1 2 3 4 5 6 7 | 10 000 | Blosum62 | rejection | **Blosum62** Fitness Neutral  **0.9933** 0 0.0067 | |
| Toll-Interleukin receptor domain | 23 sequences  171 amino acid | 5ku7 | Uniform 0 7.309942e-04 | Uniform 0 500 | 1 3 4 5 6 7 | 10 000 | WAG | rejection | **Fitness** Neutral WAG  **0.9933** 0.0067 0 | |

**Figure captions**

**Figure 1. Pipeline of the *ProtModel* framework.** The settings input file that includes the demographic, biological and the ABC settings is read. Then, considering the *structures.in* file, the template structure and the protein MSA, the protein sequences are simulated under the specified substitution models. Next, for each simulated MSA and for the real protein MSA the SS are calculated. Finally, based on the ABC settings specified by the user the best-fitting substitution model is estimated comparing the observed and the calculated SS values.

**Figure 2. Evaluation of the estimation of the best-fitting substitution model**. The estimation was performed under the rejection method and shows the probability of selecting the true substitution model within the 100 simulations per model for every ABC tolerance (0.005, 0.01 and 0.05) and number of simulations (10 000, 50 000 and 100 000). Error bars indicate 95% confidence intervals from the mean.



*Figure 1*



*Figure 2*