*Methods article*

**Selection among site-dependent structurally constrained substitution models of protein evolution by approximate Bayesian computation**

David Ferreiro1,2, Catarina Branco1,2, Ugo Bastolla3 and Miguel Arenas1,2\*

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

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

3Centre for Molecular Biology Severo Ochoa (CSIC-UAM), 28049 Madrid, Spain.

**Email addresses:**

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

CB: caraujo@uvigo.es

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

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:** Selection among site-dependent substitution models

**Keywords:** protein evolution, structurally constrained substitution models, empirical substitution models, approximate Bayesian computation, protein phylogenetics

**Abstract**

The selection among substitution models of molecular evolution constitutes an essential step in phylogenetic analyses. At the protein level, evolutionary analyses are traditionally based on empirical substitution models but these models make unrealistic assumptions and were surpassed by the structurally constrained substitution (SCS) models. The SCS models often consider site-dependent evolution, a process that provides realism but complicates the implementation of these substitution models into likelihood functions like those traditionally used for substitution model selection. Here we present an alternative method to perform selection among site-dependent SCS and empirical substitution models of protein evolution based on the approximate Bayesian computation (ABC) approach and its implementation into a computational framework called *ProteinModelerABC*. The framework implements ABC with and without regression adjustments and includes diverse empirical and site-dependent SCS models of protein evolution. Using extensive simulated data, we found that the method produces accurate selection between empirical and SCS models. As illustrative examples we applied the framework to analyse a variety of protein families observing that all of them fit better with SCS models.*ProteinModelerABC* is freely available from <https://github.com/DavidFerreiro/ProteinModelerABC> and includes a version for running the program in parallel on computer clusters and a graphical user interface. It is also distributed with a detailed documentation and ready-to-use examples.

**Introduction**

Substitution model selection is a traditional step of the phylogenetics pipeline because the accuracy of phylogenetic tree and ancestral sequence reconstructions, among other inferences, can be affected by the applied substitution model (Yang et al. 1994; Zhang and Nei 1997; Zhang 1999; Minin et al. 2003; Lemmon and Moriarty 2004; Ripplinger and Sullivan 2010; Arenas and Bastolla 2019; Del Amparo and Arenas 2022). The need of selecting among substitution models is essentially based on the common observation of diverse evolutionary processes occurring in nature at the molecular level, where genomic regions (Arbiza et al. 2011; Pandey and Braun 2020) or even amino acid sites of proteins (Pupko et al. 2002; Robinson et al. 2003; Echave et al. 2016; Jiménez-Santos et al. 2018; Neverov et al. 2021) evolve under different selection pressures and thus better fit with different substitution models. As a consequence, multiple substitution models of molecular evolution are frequently used in the field (see for a review Arenas 2015a).

At the protein level, two main types of substitution models were developed so far. First, the empirical substitution models, which consist of a 20 × 20 matrix with the relative rates of change among amino acids (exchangeability matrix) and the 20 amino acid frequencies at the equilibrium (Thorne 2000; Yang 2006; Arenas 2015a). These models are obtained from large empirical datasets such as nuclear (Jones et al. 1992; Whelan and Goldman 2001), chloroplast (Adachi et al. 2000), mitochondrial (Yang et al. 1998; Abascal et al. 2007) and virus (Nickle et al. 2007; Dang et al. 2010; Del Amparo and Arenas 2022) proteins, among others. Empirical substitution models of molecular evolution assume site-independent evolution so that all the protein sites are modeled with the same exchangeability matrix and amino acid frequencies, which allow a straightforward implementation of these models into likelihood functions (where the likelihood is site-specific and site-independent) (Yang 2006; Puller et al. 2020) and, in extension, in phylogenetic methods based on maximum likelihood (ML) (e.g., Darriba et al., 2011; Kozlov et al., 2019; Tamura et al., 2021). However, several studies showed that empirical substitution models produce proteins with unrealistic amino acid distributions and folding stability (Keane et al. 2006; Bordner and Mittelmann 2014; Arenas et al. 2015; Arenas and Bastolla 2019). Second, the structurally constrained substitution (SCS) models, which directly consider selection on the protein structure and usually on the protein folding stability (see for a review Liberles et al., 2012). Some SCS models account for site-dependent evolution and produce proteins with amino acid distributions and folding stability more realistic than those derived from empirical substitution models (Arenas et al. 2013). Notice that residues at the protein core often exhibit substitution patterns different from those located in other regions of the protein (e.g., surface) due to selection on the folding stability and activity (Jiménez-Santos et al. 2018; Echave 2019; Perron et al. 2019). Indeed, physicochemical interactions between amino acids at different sites of a protein are often observed (Shakhnovich et al. 1996), promoting coevolution among sites (Starr and Thornton 2016; Neverov et al. 2021; Chaurasia and Dutheil 2022) and suggesting that site-dependent substitution models of evolution should be preferably considered over those that ignore coevolution. However, site-dependent models of evolution cannot be incorporated into likelihood functions due to the consideration of the site-dependence evolutionary process, note that current phylogenetic likelihood functions calculate site-independent likelihoods (Yang 2006). Consequently, these models cannot be used in likelihood-based methods and thus cannot be evaluated respect to other models with the traditional substitution model selection based on likelihood ratio tests (LRTs), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), among others (Sullivan and Joyce 2005; Luo et al. 2010; Darriba et al. 2011; Darriba et al. 2020). As a consequence, there is a need of likelihood-free methods to perform selection among substitution models of evolution that can include those which consider site-dependent evolution.

As an alternative, the approximate Bayesian computation (ABC) approach is traditionally used to perform model selection in population genetics and ecology without the need of a likelihood function (Beaumont et al. 2002; Beaumont 2010). This approach is based on extensive computer simulations with parameters sampled from prior distributions, summary statistics (SS) that extract the information from the query and simulated data and a statistical adjustment (i.e., rejection or multiple linear regression, among others) to obtain the posterior distribution of each evaluated model (Csilléry et al. 2012). Despite ABC does not require likelihood analyses, it can provide estimates with similar (sometimes higher) accuracy compared to those obtained with likelihood-based methods (Lopes et al. 2014). Some previous studies demonstrated that ABC can be used to study molecular evolution (Wilson et al. 2009; Lopes et al. 2014; Arenas 2015b; M. Arenas et al. 2015; Moshe et al. 2022). For example, at the protein level, we previously applied ABC to estimate substitution and recombination rates with acceptable accuracy (Arenas 2022). Key factors for adapting ABC to the evolutionary analysis of protein sequences are the simulation of protein data under substitution models of evolution operating along evolutionary histories (i.e., phylogenetic trees) and the design of informative SS to extract evolutionary information from this genetic marker. Concerning the simulation of protein evolution upon evolutionary histories, it was implemented in diverse evolutionary frameworks (see the reviews Arenas, 2012; Hoban et al., 2012) and some of them include SCS models (Grahnen and Liberles 2012; Arenas et al. 2013), although we believe that efforts should be made in implementing additional SCS models into computer simulators. Concerning the SS to extract evolutionary information from protein sequences, Arenas (2022) found that several statistics (i.e., mean, standard deviation, skewness and kurtosis) of heterozygosity and pairwise sequence identity are informative for ABC-based analyses based on simulations of protein evolution under empirical substitution models. However, SCS models produce evolutionary signatures in sequences that could only be detected by evaluating the fitting of the protein sequence with a respective protein structure. Conveniently, there are statistics that can be used for this purpose such as the hydrophobicity (Jiménez-Santos et al. 2018), entropy (Goldstein and Pollock 2017), contact interactions (Franzosa and Xia 2009), solvent accessibility (Yeh et al. 2014), and, in general, the protein folding stability These statistics could allow the application of ABC to study patterns of protein evolution with selection on the protein structure.

Here, we present the application of ABC to perform selection among substitution models of protein evolution that can consider site-dependent evolution, thus providing an alternative strategy to evaluate substitution models that, due to their complexity, cannot be implemented into likelihood functions. We implemented this method into a user-friendly computational framework called *ProteinModelerABC*. This framework predicts the best-fitting substitution model among a set of site-dependent SCS models or a set of site-dependent SCS and empirical models, for a multiple alignment of protein sequences under ABC based on protein evolution simulated upon coalescent evolutionary histories or user-specified phylogenetic trees. It includes optional evolutionary parameters (including fixed and nuisance parameters, the latter involving user-specified prior distributions) that can be optionally used to provide a more realistic modeling, several SS designed to extract evolutionary information about protein evolution at both sequence and structure levels and, substitution model selection with ABC under rejection and multiple linear regression approaches. We evaluated the framework with data simulated under site-dependent SCS and empirical models and obtained an acceptable accuracy. As illustrative practical examples, we evaluated the fitting of site-dependent SCS and empirical substitution models with some protein families from diverse organisms of general interest.

**System and methods**

The ABC method identifies the best-fitting substitution model for a given multiple alignment of protein sequences through four main methodological steps (Figure 1) that include the reading of the input information [the query multiple sequence alignment (MSA), the evaluated substitution models and the desired ABC estimation method], the simulation of protein sequences evolution under the evaluated substitution models and along a simulated or specified evolutionary history, the calculation of informative SS and the substitution model selection with an ABC method. Details about these steps are provided below.

1. Specifying the input information. The ABC approach presents diverse advantages respect to other analytical approaches, for example it allows the analysis of complex models of evolution that cannot be incorporated into likelihood functions (see Introduction). However, it also requires that users make certain decisions, such as the number of computer simulations or the fraction of computer simulations retained for the estimation (tolerance), that can affect the estimations. A list with all the parameters implemented in the framework is presented in Table S1 (Supplementary Information). A variety of input parameters are optional but can be useful to provide a more realistic modeling of the evolutionary scenario. For example, the user can optionally specify diverse population genetics parameters used to simulate coalescent evolutionary histories (i.e., population growth rate and migration rate) or the empirical substitution models can optionally include variable substitution rate among sites (Yang et al. 1994) and proportion of invariable sites (Shoemaker and Fitch 1989). Despite the aim of *ProteinModelerABC* is evaluating site-dependent SCS models [note that other well-established frameworks are already available for selecting the best-fitting substitution model among a set of empirical substitution models (i.e., Keane et al. 2006; Darriba et al. 2011; Kalyaanamoorthy et al. 2017; Darriba et al. 2020)], it implements a variety of empirical substitution models that allow comparisons between site-dependent SCS and empirical models. The empirical substitution models implemented in *ProteinModelerABC* are *Blosum62* (Henikoff and Henikoff 1992), *CpRev* (Adachi et al. 2000), *Dayhoff* (Dayhoff et al. 1978), *DayhoffDCMUT* (Kosiol and Goldman 2005), *HIVb* (Nickle et al. 2007), *HIVw* (Nickle et al. 2007), *JTT* (Jones et al. 1992), *JonesDCMUT* (Kosiol & Goldman, 2005), *LG* (Le and Gascuel 2008), *Mtart* (Abascal et al. 2007), *Mtmam* (Yang et al. 1998), *Mtrev24* (Adachi and Hasegawa 1996), *RtRev* (Dimmic et al. 2002), *VT* (Müller and Vingron 2000), *WAG* (Whelan and Goldman 2001) and also the user can specify an exchangeability matrix and amino acid frequencies not implemented in the framework. Concerning site-dependent SCS models, the framework implements two site-dependent SCS models, named “*Fitness*” and “*Neutral*” (Arenas et al. 2013), that consider configurational entropies of the misfolding and unfolding states (Minning et al. 2013), number of contacts and hydrophobicity and including negative design (correction of predictions considering other protein structures present in the PDB), among others aspects to model protein evolution accounting for structural constraints (Arenas et al. 2013). The *Fitness* and *Neutral* SCS models are similar, their difference is that only the former model considers the effective population size in the Moran process that evaluates the probability of accepting mutation events (see the review Sella and Hirsh 2005). Despite the *Fitness* model produced amino acid distributions more similar to the real observations than those obtained with the *Neutral* model for some protein families, the *Neutral* model was in general more robust for the analysis of data with different nature (Arenas et al. 2013). Both site-dependent SCS models produced proteins with more realistic amino acid distributions and folding stabilities than those obtained with empirical models (Arenas et al. 2013). As input information, these SCS models require the specification of several thermodynamic parameters and a PDB protein structure representative of the query multiple alignment of protein sequences. Conveniently, the framework is distributed with a detailed documentation that includes recommendations about the specification of the input parameters.
2. Computer simulations. The computer simulations of protein data are performed in *ProteinModelerABC* with a recent version of the simulator *ProteinEvolver* (Arenas et al. 2013) adapted to ABC (Arenas 2022). The simulations include two main steps: (*i*) Simulation of the evolutionary history with the coalescent (Kingman 1982) under diverse evolutionary scenarios (Table S1) or, alternatively, specification of a phylogenetic tree and, (*ii*) Simulation of protein sequences evolution upon the previous evolutionary history with the substitution models of protein evolution that are evaluated. As expected, simulations of protein evolution under an empirical substitution model are rapid (seconds) and simulations under site-dependent SCS model are slower (from seconds to minutes depending on the protein length and sample size) due to the consideration of structural constraints. However, conveniently *ProteinModelerABC* can run the simulations in parallel on a multicore machine.
3. SS. We designed 7 SS (details below and in Table S2, Supplementary Information) that provided sufficient evolutionary information to distinguish between the implemented SCS models and between SCS and empirical models (details shown in the following section). In general, these SS consider the protein folding stability, molecular diversity and physicochemical properties of the amino acids involved in the replacements. Concerning the protein folding stability, it includes the mean and standard deviation of the free energy of the study proteins predicted with the *DeltaGREM* framework (Minning et al. 2013; Arenas et al. 2015). As a measure of molecular diversity we considered the number of segregating sites, following previous ABC-based studies (Lopes et al. 2014; M. Arenas et al. 2015; Arenas 2022). We also considered the change of physicochemical properties among amino acids by the mean, standard deviation, skewness and kurtosis of the Grantham distances (Grantham 1974) for protein site.
4. Substitution model selection with ABC. The presented framework calculates the posterior probability of every studied substitution model with the query protein dataset using the *abc* R library (Csilléry et al. 2012). The substitution model selection can be performed with the rejection, weighted multiple linear regression and neural networks methods (Beaumont et al. 2002; Beaumont 2010; Blum and François 2010; Csilléry et al. 2012). Under any ABC estimation method, the posterior probability of every substitution model with the query data, the confusion matrix (summary of accuracy of predictions under every studied substitution model) and the goodness of fit of the real data with the studied substitution models are provided. In addition, the framework provides the SS values of the retained simulations under every studied substitution model and the distributions of the distance between the retained and observed SS under every studied substitution model.

Altogether, *ProteinModelerABC* provides the selection among substitution models of protein evolution including complex models that cannot be implemented in likelihood functions through ABC. The framework is written in Python, C and R, and optionally can run in parallel on local or cluster computers. Interestingly, the program includes a graphical user interface (GUI) that can be useful for users that are not familiar with the command line. *ProteinModelerABC* is freely available from <https://github.com/DavidFerreiro/ProteinModelerABC> and it is distributed with a detailed documentation and illustrative practical examples.

**Results**

***ProteinModelerABC* validation**

The use of ABC for selecting among evolutionary scenarios is well-established in population genetics and ecology (e.g., Branco et al., 2022; Leuenberger & Wegmann, 2010; Sousa et al., 2012). However, ABC was not extended to substitution model selection despite it can provide a proper alternative to evaluate complex substitution models that cannot be analyzed with a likelihood functions. In this section, we evaluated the accuracy of *ProteinModelerABC* to perform selection among empirical and SCS models under different scenarios: (*i*) Number of simulations (10,000, 50,000 and 100,000), (*ii*) tolerance (0.005, 0.01, and 0.05) and (*iii*) ABC estimation method: rejection, weighted multiple linear regression and neural networks. We performed the evaluations using data simulated under the *Dayhoff* empirical substitution model (which is a widely used empirical substitution model), the *Fitness* site-dependent SCS model and the *Neutral* site-dependent SCS model (Arenas et al. 2013). The simulations were inspired in the thioredoxin protein family (27 sequences and 316 amino acids, showing a sequence identity of 0.44; PFAM code PF00070) and a representative protein structure (PDB code 1TDE) (Waksman et al. 1994) selected by *SWISS-MODEL* (Arnold et al. 2006) from the alignment consensus sequence. Next, we simulated the protein sequence alignments upon coalescent evolutionary histories (Kingman 1982), based on a population size of 1,000 individuals and a population substitution rate sampled from a uniform prior distribution between 0 and 500 that include values commonly observed nature (e.g. Arenas, 2022; Carvajal-Rodriguez, 2006; Lopes et al., 2014). For every scenario (3 substitution models × 3 different number of simulations per model × 3 ABC tolerance levels × 3 ABC estimation methods = 81 scenarios), we evaluated the power of *ProteinModelerABC* to distinguish between the three substitution models with a cross-validation based on 100 simulations (Csilléry et al. 2012). We found that the framework distinguishes between the studied substitution models with few error regardless of the applied number of simulations, the tolerance level and the ABC estimation method used (Table S3, Supplementary Information).

Once we demonstrated that the method can distinguish between SCS and empirical models, we evaluated the accuracy of the framework in identifying the best-fitting substitution model. To do so, we used 100 multiple alignments of protein sequences simulated under each studied substitution model (Dayhoff, *Fitness* SCS and *Neutral* SCS models), hereafter named as pseudo-observed data. These analyses were also performed considering ABC based on 10,000, 50,000 and 100,000 simulations, tolerance levels of 0.005, 0.01, and 0.05 and the cited three ABC estimation methods. Again, we found that the accuracy of the substitution model selection is not affected by the number of simulations (thus 10,000 simulations are sufficient to distinguish between the studied models) while the optimal tolerance varies among the studied ABC methods (Figure S1, Supplementary Information). The rejection method showed a high robustness (all the pseudo-observed data produced acceptable results) and accuracy for predicting the true model, although the accuracy slightly decreased when increasing the tolerance (Figure 2), a pattern not observed for the substitution model selection with the weighted multiple linear regression and neural networksmethods (Figure S1). Indeed, analyses based on the latter methods could not converge if the tolerance is small (not enough retained simulations to allow the prediction), thus those methods were more sensible to the tolerance (at least, a tolerance of 0.05 was required to obtain a sufficient number of retained simulations for the estimation) (Figure S1). Next, we explored the influence of every type of substitution model (i.e., empirical or site-dependent SCS model) as true model on the accuracy of the substitution model selection. In general, we found an acceptable distinction between the models using the rejection method (Figure 2). Concerning the weighted multiple linear regression and neural networksmethods, again they required a high tolerance to obtain acceptable predictions especially when the true model was the empirical substitution model (Figure S1). Altogether, these findings recommend using the rejection method because it was less sensible to the tolerance level. We did not observe effects from the number of simulations on the accuracy of the predictions (Figure S1; compare the three plots).

***Illustrative examples of substitution model selection in diverse protein families***

We used *ProteinModelerABC* to identify the best-fitting substitution model, among the best-fitting empirical substitution model previously selected with *ProtTest3* (Darriba et al. 2011)and the site-dependent SCS models implemented in *ProteinModelerABC*, in 10 protein families (Table 1). These protein families belong to viruses related with human diseases such as HIV-1 PR, HIV-1 gag, influenza NS1, SARS-CoV-2 endopeptidase C30 and 2'-O-methyltransferase, Ebola nucleoprotein and the tumour necrosis factor (TNF) of monkeypox (Mpox) virus. We also explored the highly conserved intracellular signalling Toll-Interleukin protein domain, the squalene epoxidase and the mitochondria membrane translocase, all of them randomly selected but presenting a protein structure. We obtained the protein datasets from the PFAM (Mistry et al. 2021) and PROSITE (Sigrist et al. 2012) databases, presenting different sequence length (from 99 to 450 amino acids), sample size (from 8 to 128 sequences) and sequence identity (Table 1). Next, for every dataset, we aligned the sequences with *MUSCLE* (Edgar 2004) and obtained a consensus sequence that we used to identify a representative protein structure with *SWISS-MODEL* (Arnold et al. 2006) (Table 1). The simulation of protein evolution under site-dependent SCS models and the prediction of protein folding stability (free energy) require homology between the representative protein structure and the sequences of the dataset and thus, sites of the dataset without homology with the protein structure were excluded (the available SCS models cannot deal with indels in the protein structure). Next, *ProteinModelerABC* ran 10,000 simulations under each studied substitution model and under a prior distribution for the substitution rate that produces simulated data with a distribution of sequence identity that includes the sequence identity of the real data (Table 1). Indeed, following the previous section, we performed the prediction with the rejection estimation method and under a tolerance of 0.005. The goodness of fit analysis showed that, in general, the SS of the real data fall within the SS of the retained simulated data especially for the best-fitting substitution model (an illustrative example is shown in Figure S2, Supplementary Information). For all the studied datasets, we found that the *Fitness* or *Neutral* site-dependent SCS models fitted better with the real data than the empirical substitution model selected with *ProtTest3* (Table 1).

The user can run *ProteinModelerABC* on local computers using the command line or with the provided graphical user interface (GUI) or on cluster computers. The computer simulations under SCS models can take several hours and depend on the sample size (see Table 1). Conveniently, the simulations can run in parallel (in both command line, GUI or cluster versions) on a multicore machine, reducing computer time (Figure S3, Supplementary Information). Concerning the substitution models, as expected the required computer time is higher (around 30 times) for simulations under the site-dependent SCS models than for simulations under the empirical substitution model.

**Discussion**

Currently available methods for substitution model selection are based on the likelihood that quantifies the fitting between the substitution model and the query data. This likelihood is commonly calculated per site, assuming site-independent evolution and, to our knowledge, none current likelihood function allows the analysis of substitution models that consider site-dependent evolution [which are increasing in popularity because they can produce proteins with more realistic folding stabilities and distribution of amino acid frequencies than traditional models (i.e., Arenas et al., 2013; Larson et al., 2020; Robinson et al., 2003; Rodrigue et al., 2005; Yu & Thorne, 2006)]. As a consequence, likelihood-free methodologies for substitution model selection can be relevant. Here, we present the application of the ABC approach, which does not require a likelihood function, to perform selection among substitution models that can consider site-dependent evolution. We extended our previous ABC studies oriented to estimate parameters of molecular evolution (Arenas et al. 2015; Arenas 2022) by adapting simulations and SS to build an ABC method designed to perform selection among substitution models that consider site-dependent evolution with constraints on the protein structure. We found that the evaluation of the ABC method for selecting among site-dependent SCS models and empirical models was successful. We noted that 10,000 simulations are sufficient to distinguish between the studied models, which are much less simulations than those required to estimate parameters (at least 50,000 simulations in our previous studies). We implemented the method into a freely available framework named *ProteinModelerABC* that implements a variety of evolutionary parameters related with the evolutionary history (either simulated with the coalescent under diverse population genetics processes or a user-specified phylogenetic tree (Table S1)), the modelling of protein evolution and the ABC method to calculate the probability of every studied substitution model. The framework is distributed with a command line version, a GUI and a cluster version, which can run the simulations in parallel on a multicore machine, and includes a detailed documentation and several illustrative examples that we recommend to probe.

As illustrative examples of application, we performed the selection among site-dependent SCS models and the best-fitting empirical substitution model selected with *ProtTest3* (Darriba et al. 2011) for several protein families of general interest. We found that all the studied datasets fitted better with a site-dependent substitution model. For example, for a dataset of Mpox TNF receptor we found that a site-dependent SCS model was preferred when compared to the empirical substitution model HIVw selected by *ProtTest3* (Darriba et al. 2011) as best-fitting model. We also explored some interesting therapeutic target proteins such as SARS-CoV endopeptidase C30, SARS-CoV 2'-O-methyltransferase, influenza NS1, HIV-1 GAG and HIV-1 PR. Again, we found that a site-dependent SCS model was selected when compared with the best-fitting substitution model. These results suggest that site-dependent SCS models can explain the real evolutionary process better than the traditional empirical substitution models. Perhaps the currently available set of empirical substitution models is very limited and more empirical substitution models should be developed to better mimic the evolution of the studied protein families. However, we believe that just site-dependent SCS models are much more realistic than any empirical substitution model, that as indicated in the Introduction, assume a same exchangeability matrix for all the protein sites. Also note that proteins often present intramolecular interactions that can promote selection towards specific variants (Woo et al. 2014; Rawi et al. 2015; Codoñer et al. 2017; Priya and Shanker 2021; Ferreiro et al. 2022). On the other hand, we find important to consider that a limitation of the implemented site-dependent SCS models is that they assume a representative protein structure for all the sequences of the query dataset. This assumption could lead to a poor fitting if the query dataset has sequences without homology with the representative protein structure. In this regard, the development of more robust SCS models, for example models that include different protein structures and the evolution of protein structures, and their implementation in frameworks than can be used for phylogenetics remain highly demanded in the field.

Altogether, here we show that ABC can be used for selection among complex substitution models of protein evolution, providing a useful alternative to evaluate substitution models that cannot be incorporated into likelihood functions. We implemented this methodology in the freely available evolutionary framework *ProteinModelerABC*, that includes a command line, a GUI and a cluster versions, and conveniently can run the simulations in parallel. Despite here we present a first application of ABC to the selection among substitution models of protein evolution, we believe that it could be extended for accommodating SCS models that can be developed in the future, for example (if required) by designing additional SS that can be informative about the evolutionary processes and that could be easily incorporated in the framework.

**Data availability**

*ProteinModelerABC* is freely available from <https://github.com/DavidFerreiro/ProteinModelerABC>. The simulated and real data used in the study are available from Zenodo at ….

**Supplementary data**

Supplementary data is available at the journal online.

**Acknowledgments**

This work was supported by the Spanish Ministry of Economy and Competitiveness and Ministry of Science and Innovation through the Grant [PID2019-107931GA-I00/AEI/10.13039/501100011033] and by the Xunta de Galicia through the Grant [ED481A-2020/192].

**Conflicts of Interest**

None declared.

**References**

Abascal F, Posada D, Zardoya R. 2007. MtArt: A New Model of Amino Acid Replacement for Arthropoda. *Mol. Biol. Evol.* 24:1–5.

Adachi J, Hasegawa M. 1996. Programs for Molecular Phylogenetics Based on Maximum Likelihood. *Comput Sci Monogr* 28:1–150.

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.

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

Arenas M. 2012. Simulation of Molecular Data under Diverse Evolutionary Scenarios.Lewitter F, editor. *PLoS Comput. Biol.* 8:e1002495.

Arenas M. 2015a. Trends in substitution models of molecular evolution. *Front. Genet.* [Internet] 6. Available from: http://journal.frontiersin.org/Article/10.3389/fgene.2015.00319/abstract

Arenas M. 2015b. Advances in Computer Simulation of Genome Evolution: Toward More Realistic Evolutionary Genomics Analysis by Approximate Bayesian Computation. *J. Mol. Evol.* 80:189–192.

Arenas M. 2022. ProteinEvolverABC: coestimation of recombination and substitution rates in protein sequences by approximate Bayesian computation.Schwartz R, editor. *Bioinformatics* 38:58–64.

Arenas M, Bastolla U. 2019. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. *Methods Ecol. Evol.* 11:248–257.

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

Arenas M., Lopes JS, Beaumont MA, Posada D. 2015. CodABC: A Computational Framework to Coestimate Recombination, Substitution, and Molecular Adaptation Rates by Approximate Bayesian Computation. *Mol. Biol. Evol.* 32:1109–1112.

Arenas Miguel, Sánchez-Cobos A, Bastolla U. 2015. Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. *Mol. Biol. Evol.* 32:2195–2207.

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.

Beaumont MA. 2010. Approximate Bayesian Computation in Evolution and Ecology. *Annu. Rev. Ecol. Evol. Syst.* 41:379–406.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian Computation in Population Genetics. *Genetics* 162:2025–2035.

Blum MGB, François O. 2010. Non-linear regression models for Approximate Bayesian Computation. *Stat. Comput.* 20:63–73.

Bordner AJ, Mittelmann HD. 2014. A New Formulation of Protein Evolutionary Models that Account for Structural Constraints. *Mol. Biol. Evol.* 31:736–749.

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.

Carvajal-Rodriguez A. 2006. Recombination Estimation Under Complex Evolutionary Models with the Coalescent Composite-Likelihood Method. *Mol. Biol. Evol.* 23:817–827.

Chaurasia S, Dutheil JY. 2022. The Structural Determinants of Intra-Protein Compensatory Substitutions.Rogers R, editor. *Mol. Biol. Evol.* 39:msac063.

Codoñer FM, Peña R, Blanch-Lombarte O, Jimenez-Moyano E, Pino M, Vollbrecht T, Clotet B, Martinez-Picado J, Draenert R, Prado JG. 2017. Gag-protease coevolution analyses define novel structural surfaces in the HIV-1 matrix and capsid involved in resistance to Protease Inhibitors. *Sci. Rep.* 7:3717.

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.

Dang CC, Si Le Q, Gascuel O, Sy Le V. 2010. FLU, an amino acid substitution model for influenza proteins. *BMC Evol Biol*:11.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 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.

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

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure. Vol. 5. Dayhoff, M.O. Edition. Washington DC: National Biomedical Research Foundation. p. 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.

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.

Echave J. 2019. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity.Thorne J, editor. *Mol. Biol. Evol.* 36:613–620.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* 17:109–121.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Ferreiro D, Khalil R, Gallego MJ, Osorio NS, Arenas M. 2022. The evolution of the HIV-1 protease folding stability. *Virus Evol.* 8:veac115.

Franzosa EA, Xia Y. 2009. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Mol. Biol. Evol.* 26:2387–2395.

Goldstein RA, Pollock DD. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nat. Ecol. Evol.* 1:1923–1930.

Grahnen JA, Liberles DA. 2012. CASS: Protein sequence simulation with explicit genotype-phenotype mapping. *Trends Evol. Biol.* 4:9.

Grantham R. 1974. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* 185:862–864.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89:10915–10919.

Hoban S, Bertorelle G, Gaggiotti OE. 2012. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* 13:110–122.

Jiménez-Santos MJ, Arenas M, Bastolla U. 2018. Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution. *PeerJ* 6:e5549.

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

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

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.

Kingman JFC. 1982. The coalescent. *Stoch. Process. Their Appl.* 13:235–248.

Kosiol C, Goldman N. 2005. Different Versions of the Dayhoff Rate Matrix. *Mol. Biol. Evol.* 22:193–199.

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.

Larson G, Thorne JL, Schmidler S. 2020. Incorporating Nearest-Neighbor Site Dependence into Protein Evolution Models. *J. Comput. Biol.* 27:361–375.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* 25:1307–1320.

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

Leuenberger C, Wegmann D. 2010. Bayesian Computation and Model Selection Without Likelihoods. *Genetics* 184:243–252.

Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning APJ, Dokholyan NV, Echave J, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* 21:769–785.

Lopes JS, Arenas M, Posada D, Beaumont MA. 2014. Coestimation of recombination, substitution and molecular adaptation rates by approximate Bayesian computation. *Heredity* 112:255–264.

Luo A, Qiao H, Zhang Y, Shi W, Ho SY, Xu W, Zhang A, Zhu C. 2010. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evol. Biol.* 10:242.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation.Steel M, editor. *Syst. Biol.* 52:674–683.

Minning J, Porto M, Bastolla U. 2013. Detecting selection for negative design in proteins through an improved model of the misfolded state: Detecting Selection for Negative Design. *Proteins Struct. Funct. Bioinforma.* 81:1102–1112.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.

Moshe A, Wygoda E, Ecker N, Loewenthal G, Avram O, Israeli O, Hazkani-Covo E, Pe’er I, Pupko T. 2022. An Approximate Bayesian Computation Approach for Modeling Genome Rearrangements.Kim Y, editor. *Mol. Biol. Evol.* 39:msac231.

Müller T, Vingron M. 2000. Modeling Amino Acid Replacement. *J. Comput. Biol.* 7:761–776.

Neverov AD, Popova AV, Fedonin GG, Cheremukhin EA, Klink GV, Bazykin GA. 2021. Episodic evolution of coadapted sets of amino acid sites in mitochondrial proteins.Zhang J, editor. *PLOS Genet.* 17:e1008711.

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. 2007. HIV-Specific Probabilistic Models of Protein Evolution.Pybus O, editor. *PLoS ONE* 2:e503.

Pandey A, Braun EL. 2020. Protein evolution is structure dependent and non-homogeneous across the tree of life. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Virtual Event USA: ACM. p. 1–11. Available from: https://dl.acm.org/doi/10.1145/3388440.3412473

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.

Priya P, Shanker A. 2021. Coevolutionary forces shaping the fitness of SARS-CoV-2 spike glycoprotein against human receptor ACE2.

Puller V, Sagulenko P, Neher RA. 2020. Efficient inference, potential, and limitations of site-specific substitution models. *Virus Evol.* 6:veaa066.

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18:S71–S77.

Rawi R, Kunji K, Haoudi A, Bensmail H. 2015. Coevolution Analysis of HIV-1 Envelope Glycoprotein Complex.He Y, editor. *PLOS ONE* 10:e0143245.

Ripplinger J, Sullivan J. 2010. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. *Mol. Biol. Evol.* 27:2790–2803.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne J. 2003. Protein Evolution with Dependence Among Codons Due to Tertiary Structure. *Mol. Biol. Evol.* 20:1692–1704.

Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci.* 102:9541–9546.

Shakhnovich E, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature* 379:96–98.

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.

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2012. New and continuing developments at PROSITE. *Nucleic Acids Res.* 41:D344–D347.

Sousa VC, Beaumont MA, Fernandes P, Coelho MM, Chikhi L. 2012. Population divergence with or without admixture: selecting models using an ABC approach. *Heredity* 108:521–530.

Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci.* 25:1204–1218.

Sullivan J, Joyce P. 2005. Model Selection in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36:445–466.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11.Battistuzzi FU, editor. *Mol. Biol. Evol.* 38:3022–3027.

Thorne JL. 2000. Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* 10:602–605.

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.

Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P. 2009. Rapid Evolution and the Importance of Recombination to the Gastroenteric Pathogen Campylobacter jejuni. *Mol. Biol. Evol.* 26:385–397.

Woo J, Robertson DL, Lovell SC. 2014. Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants. *Virology* 454–455:34–39.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press Available from: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198567028.001.0001/acprof-9780198567028

Yang Z, Goldman N, Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* [Internet]. Available from: https://academic.oup.com/mbe/article/11/2/316/1113048/Comparison-of-models-for-nucleotide-substitution

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

Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Mol. Biol. Evol.* 31:135–139.

Yu J, Thorne JL. 2006. Dependence among Sites in RNA Evolution. *Mol. Biol. Evol.* 23:1525–1537.

Zhang J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* 16:868–875.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* 44:S139–S146.

**Tables**

**Table 1. Substitution model selection performed with *ProteinModelerABC* for the studied real protein families.** For every studied protein family, the table shows the accession code, the number of sequences and the sequence length of the dataset, the sequence identity, the prior for the population substitution rate (*θ*) and the corresponding approximate range of sequence identity (*SeqID*), a representative protein structure (PDB code) of the dataset, the best-fitting empirical substitution model selected with *ProtTest3*, the probability of selecting every substitution model (empirical, *Fitness* site-dependent SCS or *Neutral* site-dependent SCS model) and the execution time of the analysis.

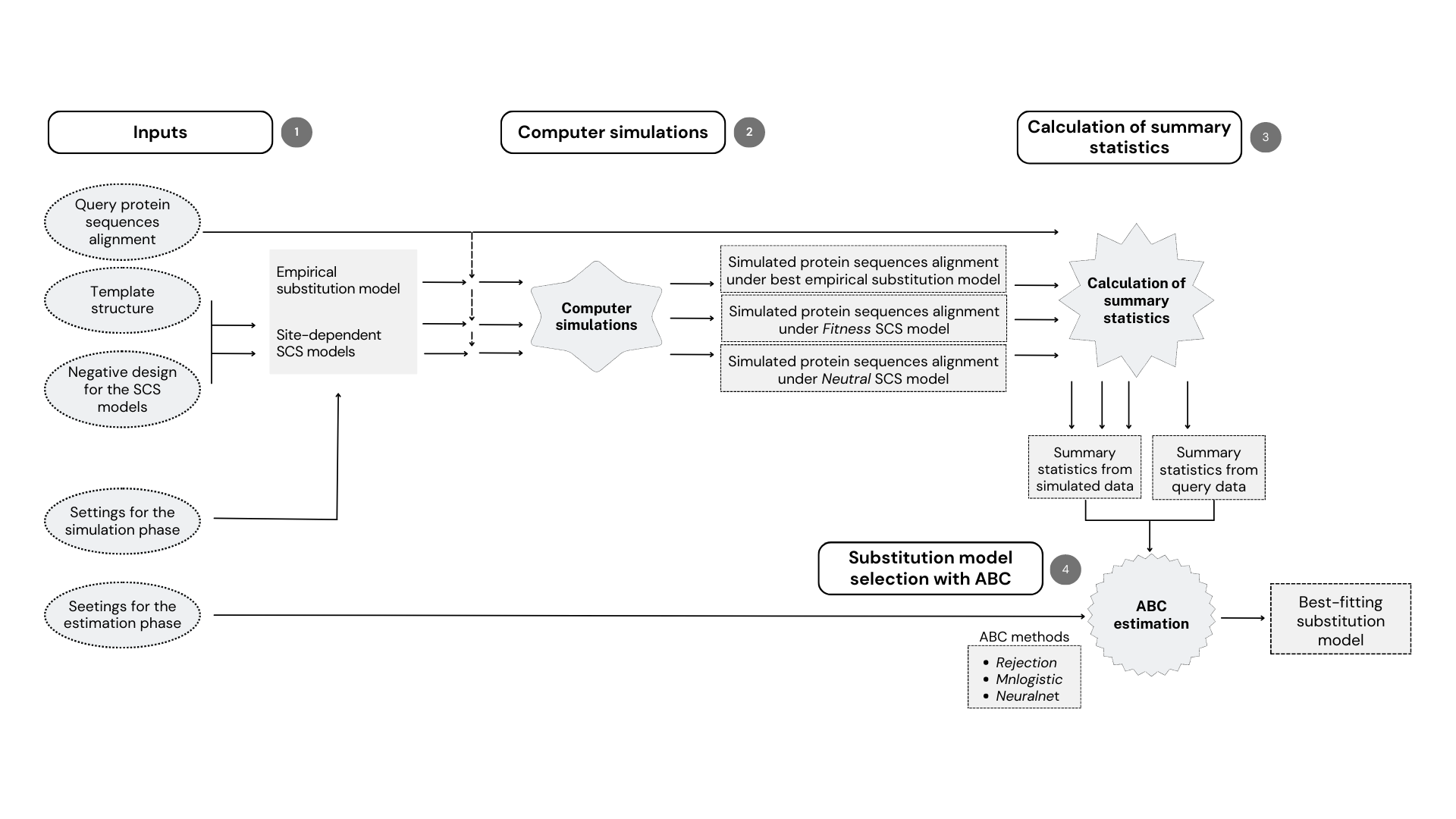
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Protein family*** | ***Sequences database entry*** | ***Number of sequences; sequences length*** | ***Sequence identity*** | ***Prior for the population substitution rate (θ); (SeqID)*** | ***Template protein structure*** | ***Best-fitting empirical substitution model*** | ***Probability of substitution model selection*** | | | ***Ejecution time (hours)*** |
| Tumor necrosis factor monkeypox | self-created1 | 10, 160 | 0.95 | Uniform | 3on9 | HIVw | Fitness | HIVw | **Neutral** | 0.22 |
| (0-100) | 0.33 | 0.15 | **0.52** |
| (1.00-0.65) |
| HIV protease (PR) | PS50175 | 95, 99 | 0.91 | Uniform | 1tcx | HIVb | **Fitness** | HIVb | Neutral | 1.53 |
| (0-150) | **0.43** | 0.25 | 0.32 |
| (1.00-0.46) |
| HIV gag polyprotein | PF00540 | 128, 288 | 0.69 | Uniform | 1l6n | RtRev | **Fitness** | Neutral | RtRev | 15.45 |
| (0-500) | **0.48** | 0.27 | 0.25 |
| (1.00-0.41) |
| Influenza NS1 | PF00600 | 25, 202 | 0.83 | Uniform | 4oph | JTT | Fitness | JTT | **Neutral** | 0.43 |
| (0-200) | 0.01 | 0.25 | **0.74** |
| (1.00-0.54) |
| Coronavirus  endopeptidase C30 | PF05409 | 30, 299 | 0.53 | Uniform | 1lvo | LG | **Fitness** | LG | Neutral | 0.48 |
| (0-500) | **0.86** | 0.03 | 0.11 |
| (1.00-0.42) |
| Coronavirus 2'-O-methyltransferase | PF06460 | 28, 298 | 0.62 | Uniform | 7c2i | LG | **Fitness** | LG | Neutral | 0.45 |
| (0-500) | **0.65** | 0.14 | 0.21 |
| (1.00-0.42) |
| Toll-Interleukin receptor domain | PF01582 | 23, 171 | 0.3 | Uniform | 5ku7 | WAG | **Fitness** | Neutral | WAG | 0.43 |
| (0-700) | **0.98** | 0.01 | 0.01 |
| (1.00-0.25) |
| Mitochondria membrane translocase | PF08038 | 54, 50 | 0.51 | Uniform | 6ucv | WAG | Fitness | **Neutral** | WAG | 0.59 |
| (0-500) | 0.33 | **0.41** | 0.26 |
| (1.00-0.14) |
| Squalene epoxidase | PF08491 | 12, 450 | 0.66 | Uniform | 6c6n | WAG | **Fitness** | Neutral | WAG | 0.21 |
| (0-500) | **0.97** | 0.03 | 0 |
| (1.00-0.50) |
| Ebola nucleoprotein | PF05505 | 8, 373 | 0.67 | Uniform | 6c54 | LG | Fitness | LG | **Neutral** | 0.17 |
| (0-500) | 0 | 0.01 | **0.99** |
| (1.00-0.47) |

1Sequences accession number in the GenBank: AAB94354, AAB94356, AAB94388, ADZ29547, YP\_010085450, AXN75227, AIE41152, AAB94364, URF91555 and AAB94363

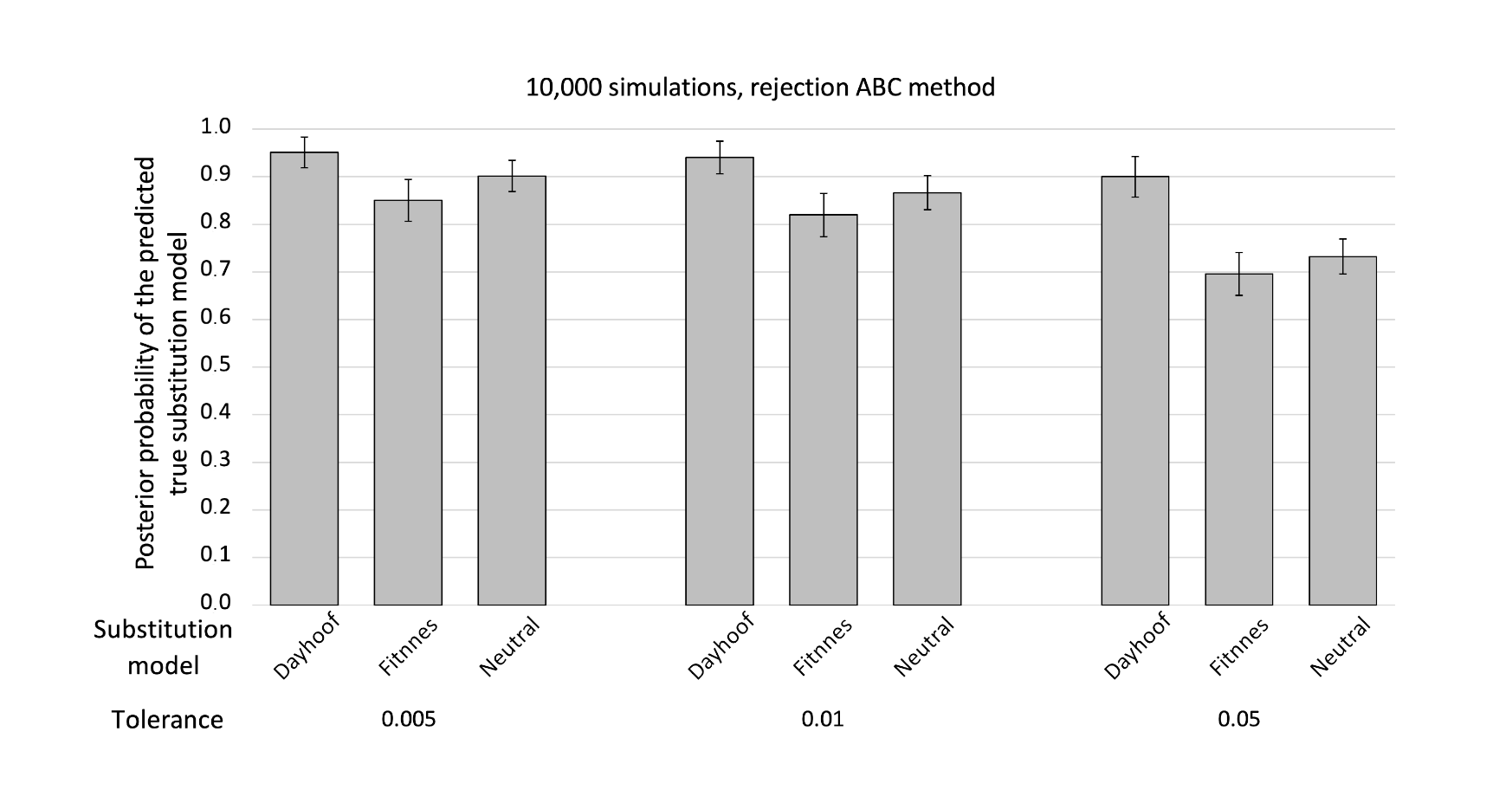
**Figure captions**

**Figure 1. Pipeline of substitution model selection with *ProteinModelerABC*.** The framework starts reading the query multiple alignment of protein sequences and diverse user-specified information such as the substitution models to be evaluated and their parameters [including a negative design for the SCS models where the predicted folding stabilities consider contacts information observed in other protein structures of the PDB (Arenas et al. 2013)], the evolutionary history (simulated with the coalescent or a user-specified phylogenetic tree), the number of simulations, the ABC estimation method, among others]. Next, the framework simulates protein sequences evolution under the specified substitution models (with an equal number of simulations performed under each model) upon the specified evolutionary history. In a subsequent step, it calculates the SS for the query and simulated datasets. Finally, the framework predicts the best-fitting substitution model, among the studied substitution models, using a user-specified ABC method.

**Figure 2. Evaluation of substitution model selection with *ProteinModelerABC* as a function of the number of simulations and tolerance**. The figure shows the probability for predicting a true substitution model (including the substitution models *Dayhoff*, *Fitness* site-dependent SCS and *Neutral* site-dependent SCS models as true models) with the different ABC estimation methods (rejection “*rejection*”, weighted multiple linear regression *“mnlogistic”* and neural networks *“neuralnet”*) and using 100 pseudo-observed datasets per model at different levels of tolerance (0.005, 0.01 and 0.05) and number of simulations (10000, 50000 and 100000). Error bars indicate 95% confidence intervals from the mean of the predictions from the pseudo-observed datasets.



*Figure 1.*



*Figure 2.*