**Supplementary Information**

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

**Index**

[Supplementary tables 2](#_Toc127189694)

[Supplementary figures 9](#_Toc127189695)

# Supplementary tables

**Table S1. Input parameters implemented in *ProteinModelerABC*.** For every parameter, the table includes its related step (i.e., general, associated to the evolutionary history, substitution model or ABC estimation, among others), if the parameter is mandatory or optional to perform an analysis, and useful comments. Further details and recommendations for the specification of every parameter are provided in the documentation distributed with the framework.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Method** | **Mandatory** | **Comments** |
| Query multiple alignment of protein sequences | General | Yes | Input file in *phylip* format. |
| Consideration of indels | General | Yes | How to consider indels (gaps). They can be ignored or considered as a new state. |
| Protein structure template | General | Yes | Representative protein structure (PDB file) of the studied multiple sequence alignment. It is required to predict folding free energies. |
| Protein structure chain | General | Yes | Chain of the representative protein structure. It is required to predict folding free energies. |
| Amount of output information shown on the screen | General | Yes | Amount of output information shown on the screen when running the program. |
| Number of simulations | Simulations | Yes | Total number of simulations. |
| Number of processors | Simulations | Yes | Number of processors used to run the simulations (note that the simulations can run in parallel). |
| Save simulations | Simulations | Yes | Save the simulated data in a compressed folder if desired. |
| Coalescent or phylogeny | Simulations | Yes | Indicate coalescent simulation (with user-specified population genetics parameters) or a rooted phylogenetic tree (user-specified), upon which protein evolution will be simulated. |
| Haploid/diploid | Simulations | Yes if coalescent | Indicate if the data belongs to a haploid or diploid organism. |
| Amino acid substitution rate | Simulations | Yes if coalescent | It can be fixed or sampled from a user-specified distribution. |
| Population size (*N*) | Simulations | Yes if coalescent | Effective population size. |
| Dated tips | Simulations | No | Time at which the tip nodes of the tree are sampled (*2N* generations for diploids). |
| Generation time | Simulations | No | Time for each generation. It can be fixed or sampled from a uniform distribution. |
| Population growth rate | Simulations | No | The population size can change over time by this exponential growth rate. |
| Migration model | Simulations | No | Migration model and type of population structure. |
| Migration rate | Simulations | No | The migration rate among demes (subpopulations). It can be constant or variable over time. |
| Convergence demes | Simulations | No | Events of convergence of demes. |
| Phylogenetic tree | Simulations | Yes if phylogeny | User-specified phylogenetic tree. |
| Substitution models of amino acid evolution | Substitution model | Yes | Empirical and/or site-dependent SCS models to be evaluated. |
| Amino acid frequencies | Substitution model | Yes | Amino acid frequencies at the equilibrium. |
| Heterogeneity of the substitution rate among sites (+G) | Substitution model | No | Substitution rate among sites according to a Gamma distribution. |
| Proportion of invariable sites (+I) | Substitution model | No | Proportion of invariable sites in the sequences. |
| ABC iterations | Estimation | Yes | Number of simulations used to perform the analysis. |
| ABC tolerance | Estimation | Yes | Fraction of simulations closer to the observed data that are retained for the ABC analysis. |
| ABC method | Estimation | Yes | ABC estimation method used for the ABC analysis. |
| Summary statistics | Estimation | Yes | Summary statistics used to perform the ABC analysis. |
| Multiple pages | Estimation | Yes | It provides informative PDF documents with multiple plots showing results. |

**Table S2. Summary statistics implemented in *ProteinModelerABC*.** For every summary statistic the table includes an identification (ID) and a brief description. See the documentation of the framework for further details.

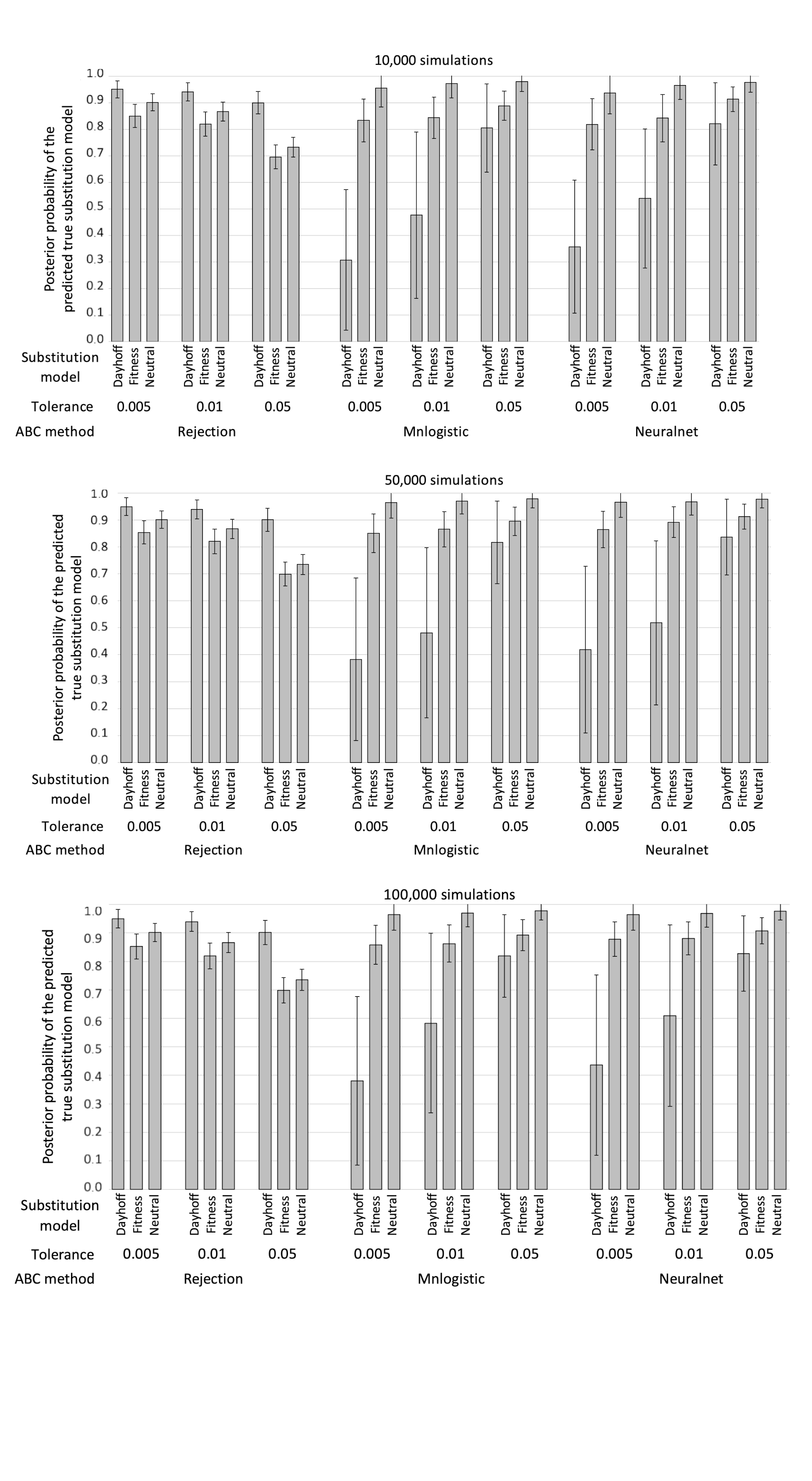
|  |  |  |
| --- | --- | --- |
| ID | Name | Description |
| 1 | *DGREM\_mean* | Mean of folding stability (free energy) of the proteins of the dataset. |
| 2 | *DGREM\_sd* | Standard deviation of folding stability among the proteins of the dataset. |
| 3 | *SegSites* | Number of segregating sites. |
| 4 | *Grantham\_mean\_Position* | Mean of the Grantham distance from amino acid replacements at every protein site (position). |
| 5 | *Grantham\_sd\_Position* | Standard deviation of the Grantham distance from amino acid replacements at every protein site. |
| 6 | *Grantham\_sk\_Position* | Skewness of the Grantham distance from amino acid replacements for every protein site. |
| 7 | *Grantham\_ku\_Position* | Kurtosis of the Grantham distance from amino acid replacements at every protein site. |

**Table S3. *ProteinModelerABC* validation cross-validation based on 100 simulations.** Frequency of predicting the true model for every considered substitution model (Dayhoff, *Fitness* site-dependent SCS and *Neutral* site-dependent SCS models) using each ABC estimation method (rejection “*rejection*”, weighted multiple linear regression *“mnlogistic”* and neural networks *“neuralnet”*) with different ABC tolerance (0.005, 0.01 and 0.05) and number of simulations per studied substitution model (10000, 50000 and 100000).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number of  simulations** | **Tolerance** | **ABC method** | **True model** | **Frequency (%) of predicting  the true model (confusion matrix)** | | |
|  | | | | **Dayhoff** | **Fitness** | **Neutral** |
| 10,000 | 0.005 | Rejection | Dayhoff | 93 | 99 | 96 |
| Fitness | 91 | 95 | 96 |
| Neutral | 92 | 94 | 88 |
| Mnlogistic | Dayhoff | 95 | 100 | 95 |
| Fitness | 95 | 95 | 95 |
| Neutral | 96 | 97 | 98 |
| Neuralnet | Dayhoff | 94 | 98 | 96 |
| Fitness | 94 | 97 | 95 |
| Neutral | 98 | 99 | 96 |
| 0.01 | Rejection | Dayhoff | 93 | 97 | 86 |
| Fitness | 92 | 96 | 95 |
| Neutral | 89 | 99 | 95 |
| Mnlogistic | Dayhoff | 97 | 97 | 98 |
| Fitness | 97 | 98 | 97 |
| Neutral | 96 | 94 | 96 |
| Neuralnet | Dayhoff | 97 | 95 | 93 |
| Fitness | 95 | 97 | 96 |
| Neutral | 99 | 99 | 95 |
| 0.05 | Rejection | Dayhoff | 91 | 97 | 79 |
| Fitness | 89 | 98 | 79 |
| Neutral | 84 | 94 | 89 |
| Mnlogistic | Dayhoff | 96 | 94 | 94 |
| Fitness | 97 | 96 | 91 |
| Neutral | 95 | 95 | 95 |
| Neuralnet | Dayhoff | 96 | 96 | 97 |
| Fitness | 90 | 97 | 92 |
| Neutral | 96 | 94 | 94 |
| 50,000 | 0.005 | Rejection | Dayhoff | 93 | 97 | 95 |
| Fitness | 94 | 98 | 92 |
| Neutral | 92 | 98 | 90 |
| Mnlogistic | Dayhoff | 94 | 99 | 97 |
| Fitness | 97 | 97 | 99 |
| Neutral | 97 | 98 | 99 |
| Neuralnet | Dayhoff | 97 | 98 | 96 |
| Fitness | 100 | 98 | 96 |
| Neutral | 97 | 95 | 97 |
| 0.01 | Rejection | Dayhoff | 93 | 99 | 89 |
| Fitness | 93 | 98 | 90 |
| Neutral | 96 | 100 | 96 |
| Mnlogistic | Dayhoff | 93 | 98 | 97 |
| Fitness | 98 | 98 | 96 |
| Neutral | 97 | 95 | 96 |
| Neuralnet | Dayhoff | 94 | 96 | 96 |
| Fitness | 98 | 97 | 98 |
| Neutral | 97 | 99 | 95 |
| 0.05 | Rejection | Dayhoff | 91 | 98 | 86 |
| Fitness | 81 | 96 | 83 |
| Neutral | 81 | 97 | 93 |
| Mnlogistic | Dayhoff | 93 | 98 | 94 |
| Fitness | 97 | 96 | 94 |
| Neutral | 98 | 93 | 96 |
| Neuralnet | Dayhoff | 92 | 99 | 94 |
| Fitness | 97 | 96 | 97 |
| Neutral | 98 | 96 | 96 |
| 100,000 | 0.005 | Rejection | Dayhoff | 93 | 97 | 95 |
| Fitness | 94 | 98 | 92 |
| Neutral | 92 | 98 | 90 |
| Mnlogistic | Dayhoff | 94 | 99 | 97 |
| Fitness | 97 | 97 | 99 |
| Neutral | 97 | 98 | 99 |
| Neuralnet | Dayhoff | 97 | 98 | 96 |
| Fitness | 100 | 98 | 96 |
| Neutral | 97 | 95 | 97 |
| 0.01 | Rejection | Dayhoff | 93 | 99 | 89 |
| Fitness | 93 | 98 | 90 |
| Neutral | 96 | 100 | 96 |
| Mnlogistic | Dayhoff | 93 | 98 | 97 |
| Fitness | 98 | 98 | 96 |
| Neutral | 97 | 95 | 96 |
| Neuralnet | Dayhoff | 94 | 96 | 96 |
| Fitness | 98 | 97 | 98 |
| Neutral | 97 | 99 | 95 |
| 0.05 | Rejection | Dayhoff | 91 | 98 | 86 |
| Fitness | 81 | 96 | 83 |
| Neutral | 81 | 97 | 93 |
| Mnlogistic | Dayhoff | 93 | 98 | 94 |
| Fitness | 97 | 96 | 94 |
| Neutral | 98 | 93 | 96 |
| Neuralnet | Dayhoff | 92 | 99 | 94 |
| Fitness | 97 | 96 | 97 |
| Neutral | 98 | 96 | 96 |

# Supplementary figures

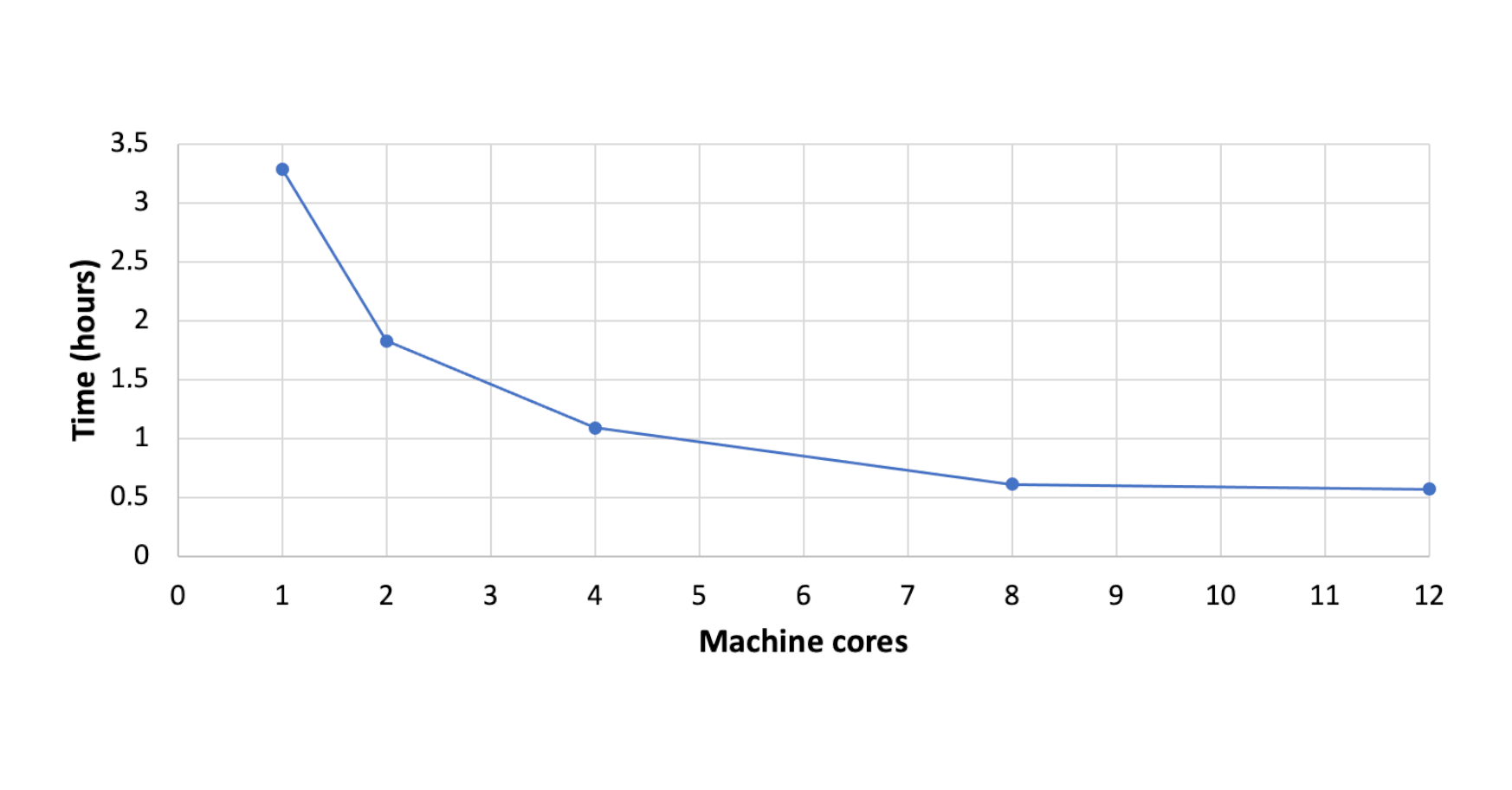
**Figure S1. Evaluation of substitution model selection under different number of simulations and tolerance thresholds for different true substitution models**. The figure shows the probability for predicting every true substitution model (Dayhoff, *Fitness* site-dependent SCS and *Neutral* site-dependent SCS models) using 100 pseudo-observed datasets per model at different levels of tolerance (0.005, 0.01 and 0.05) with the different ABC estimation methods (rejection “*rejection*”, weighted multiple linear regression *“mnlogistic”* and neural networks *“neuralnet”*) and different number of simulations (10000, 50000 and 100000, which correspond with the upper, middle and lower plots, respectively). Error bars indicate 95% confidence intervals from the mean of the predictions from the pseudo-observed datasets.



**Figure S2. Illustrative examples of evaluation of substitution model selection.** This figure shows the most informative plots derived from the analyses of the real dataset 1 (see Table 1). It includes: (1) the histogram of the simulations’ goodness-of-fit analysis for every substitution model studied, (2) the principal components analysis (PCA) about the best-fitting substitution model estimation and (3) the histograms with the estimates of summary statistics from the simulated proteins, including the folding energy mean (*DGREM\_mean*)and standard deviation *(DGREM\_sd)*, the number of segregating sites (*SegSites)* and the Grantham distance mean *(Grantham\_mean\_Position),* standard deviation *(Grantham\_sd\_Position),* skeawness (*Grantham\_sk\_Position)* and kurtosis *(Grantham\_ku\_Position)* per protein position for every substitution model studied. In the histograms the real data SS is represented with a blue line while in the PCA plot they are represented with a black cross.

|  |  |  |
| --- | --- | --- |
| **1.** | | |
|  |  |  |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
| **2.** | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
| **3.** | | |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Figure S3. Total computer time required for an analysis with *ProteinModelerABC*** **using a different number of cores.** The analysis was carried out using a local machine (2.6 GHz Intel Core i7) simulating 1,000 simulations with 1, 2, 4, 8 and 12 cores. The dataset included 10 sequences of 160 amino acids (see real dataset 1 in Table 1). When running the simulations in parallel the RAM memory is shared among cores reducing the computer time. However, the decrease of computer time may not follow a linear function if the memory is shared among cores and because some steps of the analysis are not parallelized.

****