**Supplementary Data**

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

**Index**

[Supplementary tables 2](#_Toc108596990)

[Supplementary images 5](#_Toc108596991)

# Supplementary tables

**Table S1. Parameters implemented in *ProtModel*.** The table includes the name, the type and if the parameter is needed to run an analysis as well as some other comments. See documentation for more information.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Parameter*** | ***Type*** | ***Mandatory*** | ***Comments*** |
| Multiple protein sequence alignment | General settings | Yes | Filename (i.e., no pathway) in PHYLIP format |
| Number of simulations | General settings | Yes | Total number of simulations (number of samples simulated under the prior distributions) |
| Indels | General settings | Yes | Consideration of indels (gaps). It can be "Ignored" (indels are ignored), or "NewState" (indels are considered as a new state) |
| Number of processors | General settings | Yes | Number of processors to run the simulations (it allows running the simulations on parallel) |
| Save simulations | General settings | Yes | Save simulated alignments in a compressed folder |
| Show running information | General settings | Yes | Reduce the amount of information displayed in the terminal |
| Haploid/diploid | Demographic settings | Yes | Data from haploid or diploid organism |
| Population size | Demographic settings | Yes | Effective population size |
| Dated tips | Demographic settings | No | Time at which the tip nodes of the tree were sampled (years) |
| Generation time | Demographic settings | No | Time for each generation. It can be fixed or sampled from a uniform distribution |
| Growth rate | Demographic settings | No | It can be exponential or periods can be defined |
| Migration model | Demographic settings | No | Migration model and population structure |
| Migration rate | Demographic settings | No | It can be constant or variable during time according to temporal periods |
| Convergence demes | Demographic settings | No | Events of convergence of demes |
| Amino acid substitution rate per site | Genetic settings | Yes | It can be fixed or sampled along a distribution |
| Empirical substitution model of amino acid evolution | Genetic settings | Yes | The user has to specify the desire empirical substitution models |
| SCS model of amino acid evolution | Genetic settings | Yes | The user has to specify the desire SCS models |
| Amino acid frequencies | Genetic settings | Yes | Distribution of frequencies for each amino acid site along sequences |
| Rate of heterogeneity across sites (+G) | Genetic settings | No | Distribution of heterogeneity along sequences |
| Proportion of invariable sites (+I) | Genetic settings | No | Distribution of invariable sited along sequences |
| Template | Genetic settings | Yes | PDB protein structure used to structural substitution models and to calculate proteins free energy |
| Chain | Genetic settings | Yes | PDB protein chain used to structural substitution models and to calculate proteins free energy |
| GMRCA | Genetic settings | No | Sequence of the root of the alignment coalescent history |
| Multiple pages | Graphical settings | Yes | PDF documents with multiple plots per page |
| ABC iterations | ABC settings | Yes | Number of simulations to consider in the ABC analysis |
| ABC tolerance | ABC settings | Yes | Proportion of simulations closest to real data to retain in the ABC procedure |
| ABC method | ABC settings | Yes | ABC algorithm to use for the ABC estimation |
| Summary statistics | ABC settings | Yes | Summary statistics to use for the ABC estimation by specifying their numeric identifiers |

**Table S2. Summary statistics implemented in *ProtModel*.** For every summary statistic the table includes a little description and the corresponding ID. See documentation for more information.

|  |  |  |
| --- | --- | --- |
| *Name* | *Description* | *ID* |
| DGREM\_Mean | Mean of alignment folding stability | 1 |
| DGREM\_sd | Standard deviation of alignment folding stability | 2 |
| SegSites | Number of segregation sites | 3 |
| Grantham\_mean\_Position | Mean of the Grantham distance between aa replacements per protein site | 4 |
| Grantham\_sd\_Position | Standard deviation of the Grantham distance between aa replacements per protein site | 5 |
| Grantham\_sk\_Position | Skeaness of the Grantham distance between aa replacements per protein site | 6 |
| Grantham\_ku\_Position | Kurtosis of the Grantham distance between aa replacements per protein site | 7 |

**Table S3. Goodness of fit of the substitution models analyzed in the illustrative examples (Table 1).** The table includes the protein family, the substitution models considered and the corresponding P-value. Note that a good result must present, at least, one model with a P-value higher than 0.05 or 0.1, indicating that this model represents properly the real data.

|  |  |  |
| --- | --- | --- |
| *Protein family* | *Models* | *Model goodness of fit P-value* |
| Monkeypox tumour necrosis receptor | Fitness, HIVw and Neutral | Fitness: 0.04  HIVw: 0.04  Neutral: 0.05 |
| Monkeypox tumour necrosis receptor | Fitness, HIVw and Neutral | Fitness: 0.18  HIVw: 0.15  Neutral: 0.32 |
| HIV protease | Fitness, JTT and Neutral | Fitness: 0.35  JTT: 0.21  Neutral: 0.18 |
| HIV Gag polyprotein | Fitness, HIVb and Neutral | Fitness: 0.66  HIVb: 0.52  Neutral: 0.37 |
| Influenza NS1 | Fitness, JTT and Neutral | Fitness: 0.24  JTT: 0.43  Neutral: 0.28 |
| Coronavirus endopeptidase C30 | Fitness, LG and Neutral | Fitness: 0.07  LG: 0.23  Neutral: 0.04 |
| Coronavirus endopeptidase C30 | Fitness, LG and Neutral | Fitness: 0.26  LG: 0.36  Neutral: 0.13 |
| Coronavirus 2'-O-methyltransferase | Fitness, LG and Neutral | Fitness: 0.16  LG: 0.51  Neutral: 0.08 |
| Calcium-binding EGF domain | Blosum62, Fitness and Neutral | Blosum62: 0.5  Fitness: 0.32  Neutral: 0.04 |
| Toll-Interleukin receptor domain | Fitness, Neutral and WAG | Fitness: 0.46  Neutral: 0.12  WAG: 0.32 |

# Supplementary figures

**Figure S1. Elapsed time of one *ProtModel*** **analysis using different number computer or cluster cores.** The analysis using a computer was carryed out simulating 1,000 simulations with 1, 2, 4, 8 and 12 cores and for the analysis using the cluster we perform 10,000 simulations using 50, 100, 200, 250 and 500 cores.

