Supplementary Material

ProteinEvolverABC: Coestimation of Recombination and Substitution Rates in Protein Sequences by approximate Bayesian computation

The supplementary material includes the Tables S1-S2, Figures S1-S19 and the literature cited in the Supplementary Material.

Table S1. Parameters and prior distributions implemented in *ProteinEvolverABC*. The table shows a list of relevant parameters implemented in the framework. For each parameter it shows the type of parameter (i.e., general setting, simulation or estimation), prior distributions (if not shown the parameter must be fixed by the user) and basic information (including if the parameter is optional or required). The implemented *Gamma*, *Beta*, *Normal* and *Exponential* prior distributions can be truncated by the used if desired.

Parameter	Type of parameter	Prior distributions	Comments
Number of simulations	General settings	-	Required. Total number of simulations (number of samples obtained from the prior distributions)
Number of processors	General settings	-	Number of processors to run the simulations (it allows running the simulations on parallel)
Haploid or diploid data	Evolutionary history	-	Required. Type of input and simulated data
Population size	Evolutionary history	-	Required. Population size used for the coalescent simulations
Longitudinal sampling	Evolutionary history	-	Samples collected at different times
Generation time	Evolutionary history	Fix, Uniform, Gamma, Beta, Normal, Exponential	Required in case of specifying longitudinal sampling. Nuisance parameter
Recombination rate	Evolutionary history	Fix, Uniform, Gamma, Beta, Normal, Exponential	Required. Parameter to be estimated
Substitution rate	Molecular evolution	Fix, Uniform, Gamma, Beta, Normal, Exponential	Required. Parameter to be estimated

Substitution model	Molecular evolution	-	Required. The implemented models 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 and 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 (Muller and Vingron, 2000), WAG (Whelan and Goldman, 2001) and, user-specified exchangeability matrix
Amino acid frequencies	Molecular evolution	Fix, Dirichlet	Required. Nuisance parameter
Heterogeneous rate of change among sites (+G)	Molecular evolution	Fix, Uniform, Gamma, Beta, Normal, Exponential	Nuisance parameter
Proportion of invariable sites (+I)	Molecular evolution	Fix, Uniform, Gamma, Beta, Normal, Exponential	Nuisance parameter
ABC iterations	ABC estimation	-	Required. Number of simulations considered for the estimation
ABC tolerance	ABC estimation	-	Required. Chosen number of simulations closest to real data for the estimation
ABC method	ABC estimation	-	Required. ABC estimation method (rejection and regression)
Summary statistics	ABC estimation	-	Required. Chosen summary statistics for the estimation (it is recommended the use of all the implemented 16 summary statistics)

Table S2. Summary statistics implemented in *ProteinEvolverABC*. The table shows summary statistics implemented in the framework classified in several groups: fast recombination tests, entries 1-3; amino acid diversity, entries 4-7; heterozygosity, entries 8-11; number of segregating sites, entry 12; pairwise amino acid sequence identity, entries 13-16. Each summary statistic includes an identification code that is cited in the software documentation. Sd, Sk and Ku refer to standard deviation, skewness and kurtosis, respectively.

Entry	Description	Code
1	Pairwise homoplasy index PHI	Phi
2	Neighbor similarity scope NSS	NSS
3	Maximum chi-squared χ^2	ChiSq
4	Mean of amino acid diversity	p_av
5	Sd of amino acid diversity	p_sd
6	Sk of amino acid diversity	p_sk
7	Ku of amino acid diversity	p_ku
8	Mean of amino acid heterozygosity	H_av
9	Sd of amino acid heterozygosity	H_sd
10	Sk of amino acid heterozygosity	H_sk
11	Ku of amino acid heterozygosity	H_ku
12	Number of amino acid segregating sites	S
13	Mean of pairwise amino acid sequence identity	si_av
14	Sd of pairwise amino acid sequence identity	si_sd
15	Sk of pairwise amino acid sequence identity	si_sk
16	Ku of pairwise amino acid sequence identity	si_ku

Figure S1. Accuracy of *ProteinEvolverABC* in the estimation of the recombination and substitution rates under different evolutionary scenarios and based on ABC with 100,000 simulations. For each studied combination of ρ and θ (evolutionary scenario evaluated with 100 test datasets) the figure shows the estimates of ρ (above) and θ (below). The black bars indicate the true value. Clear and dark grey bars correspond to the mode of the estimated posterior distributions (using the rejection and multiple linear regression approaches, respectively, both based on 100,000 simulations) and error bars indicate the 95% confidence interval.

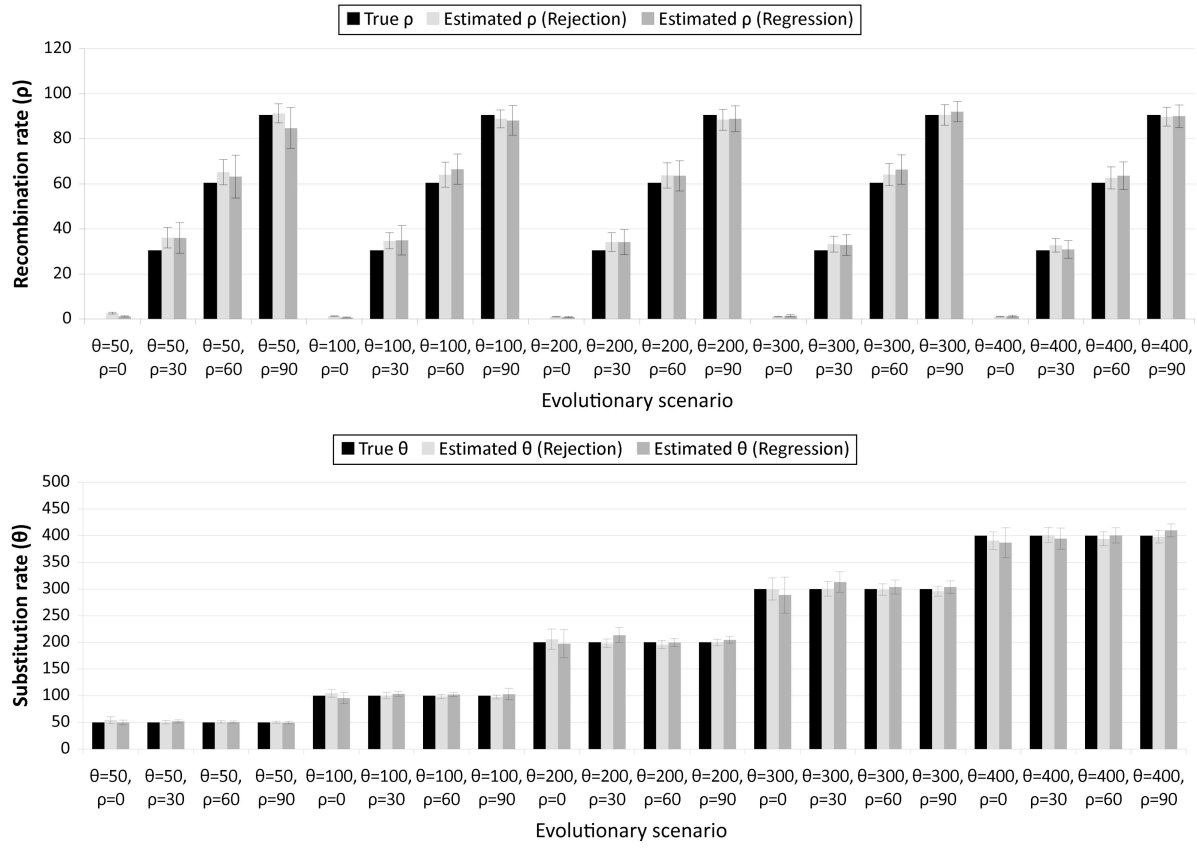


Figure S2. *ProteinEvolverABC* computing times for the analysis of real data. Computer times for the analysis of 8 real datasets (protein families) with *ProteinEvolverABC* using a different number of processors. Each protein family (legend, further details are shown in Table 1) is identified with its PFAM code and, in parenthesis, number of sequences and sequence length in amino acids, respectively. Prior distributions are the following, ρ: *Uniform*(0,120) and θ: *Uniform*(0,500). The analyses were ran on an Intel® Core i7 2.5GHz with 4 cores. The decline of computer time with the number of processors is not linear because parallelizing can still share some tasks among processors (i.e., storage) and also other phases of the estimation do not run in parallel.

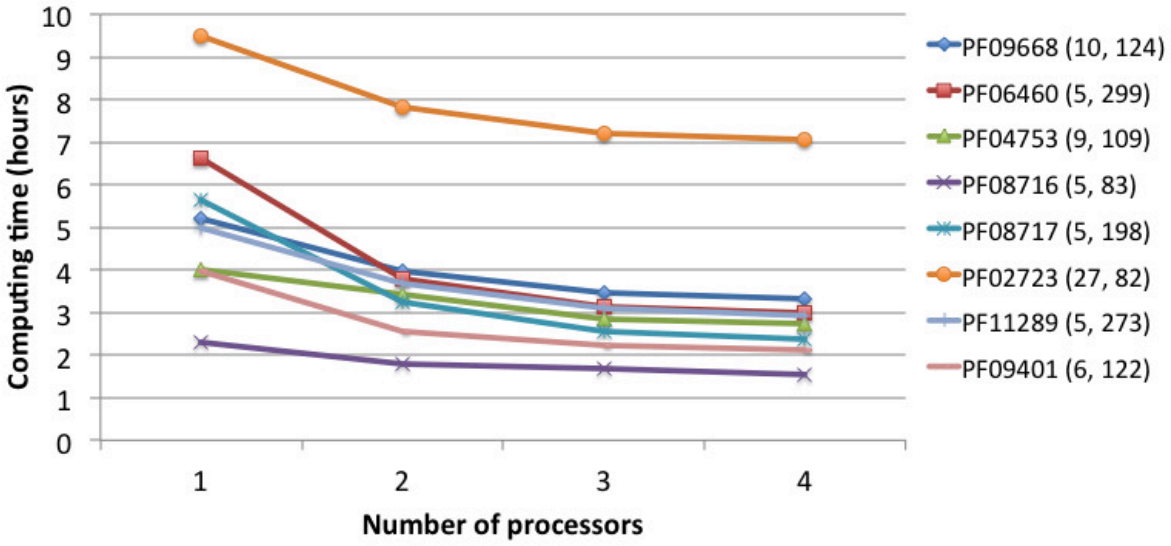


Figure S3. *ProteinEvolverABC* computing times for the analysis the protein family *coronavirus replicase NSP7* under different number of sequences, sequence length and substitution model of protein evolution. Computer times for the analysis of different scenarios applied to the *coronavirus replicase NSP7* (PFAM code: PF08716) that originally presents 5 sequences with 83 amino acids. A: Evaluation of computer times as a function of the number of sequences by artificially duplicating the number of sequences. B: Evaluation of computer times as a function of the sequence length by artificially duplicating the sequence length. C: Evaluation of computer times for diverse substitution models of protein evolution [including models based on nuclear (i.e., JTT), mitochondrial (i.e., MtMam) and virus (i.e., HIVb) proteins] (Arenas, 2015; Yang, 2006) [+G indicates variation of the substitution rate among sites according to a gamma distribution (Yang et al., 1998)]. Prior distributions are the following, ρ*: Uniform*(0,120) and θ*: Uniform*(0,500). The analyses were ran on an Intel® Core i7 2.5GHz using one core.

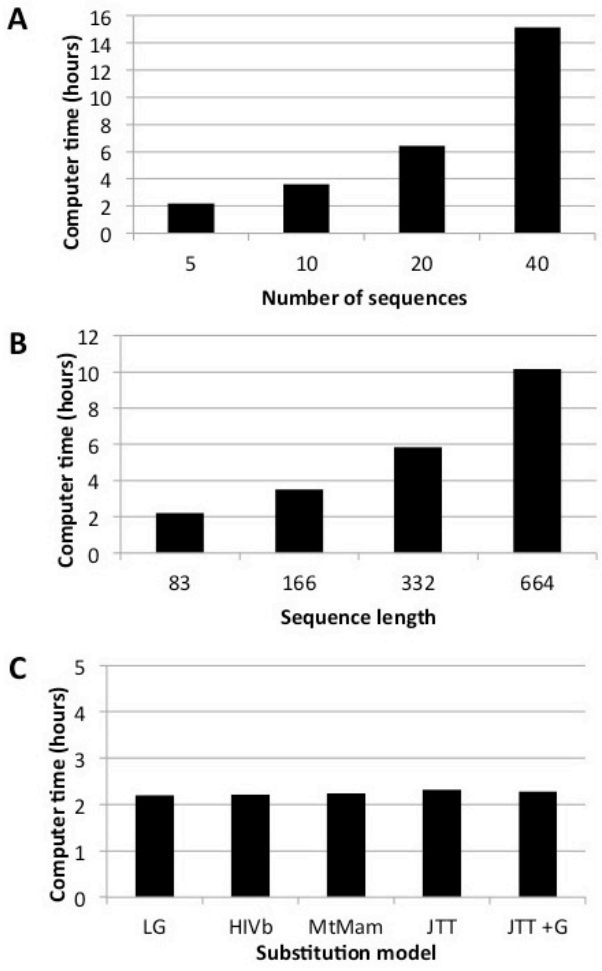


Figure S4. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF02723. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset are inside of SS of the retained simulations.

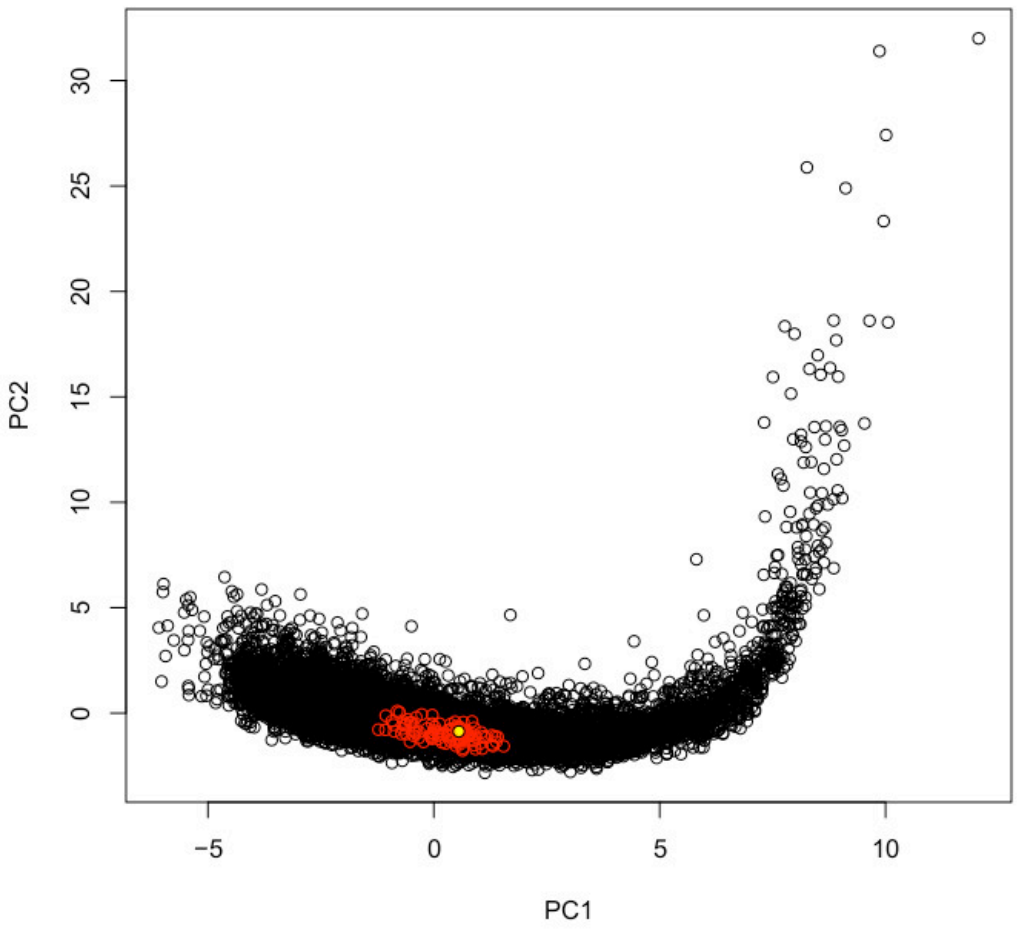


Figure S5. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF06460. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset are inside of SS of the retained simulations.

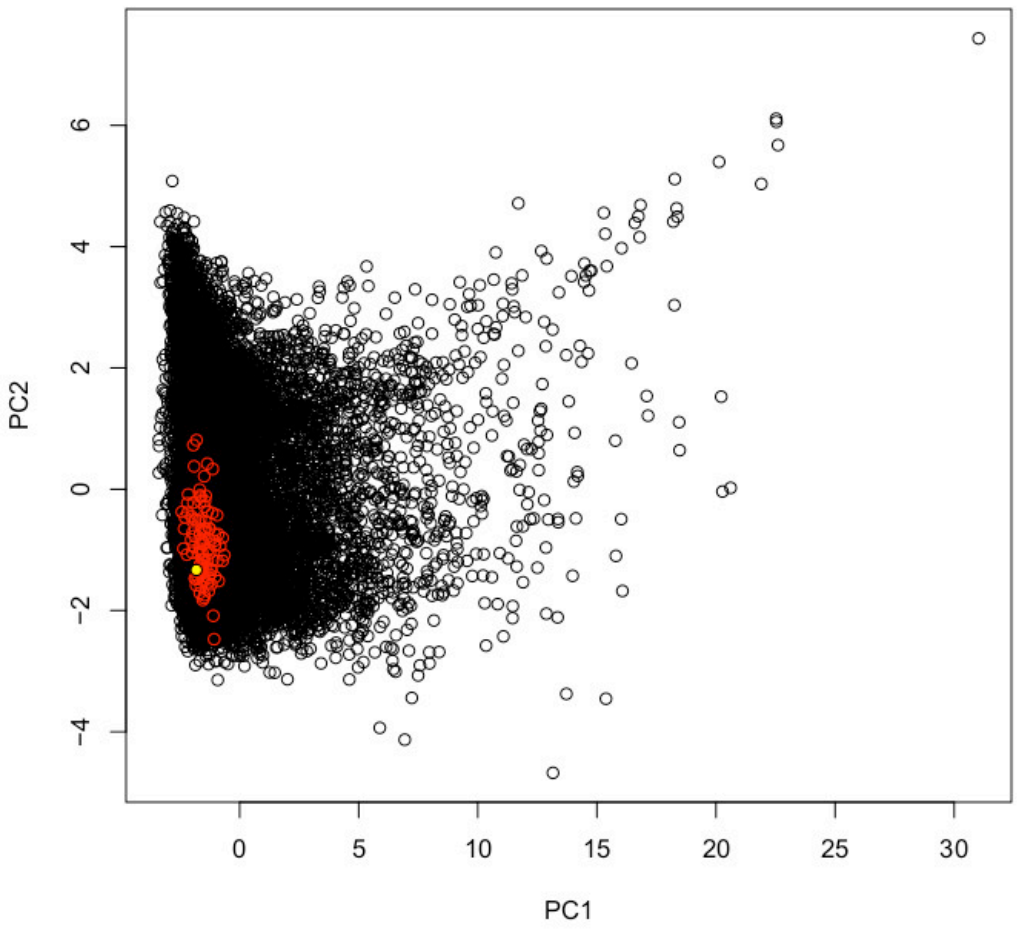


Figure S6. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF04753. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset are inside of SS of the retained simulations.

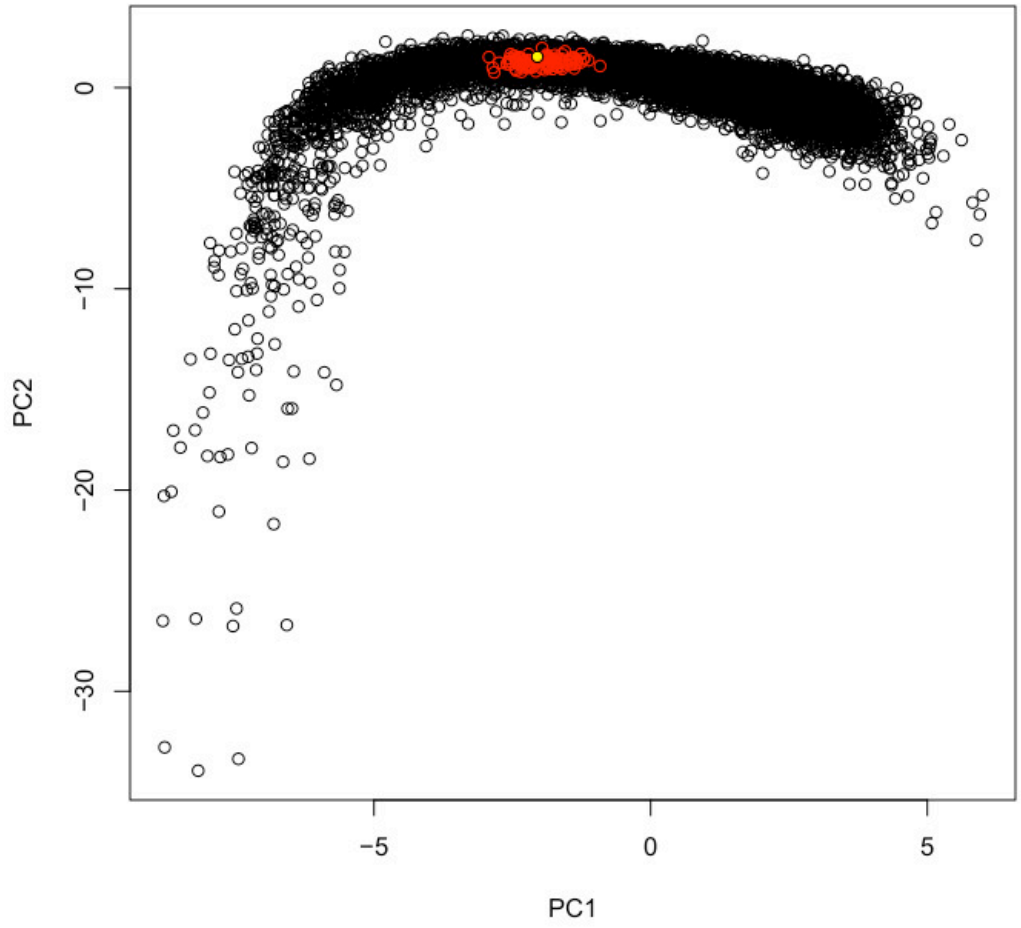


Figure S7. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF08716. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset are inside of SS of the retained simulations.

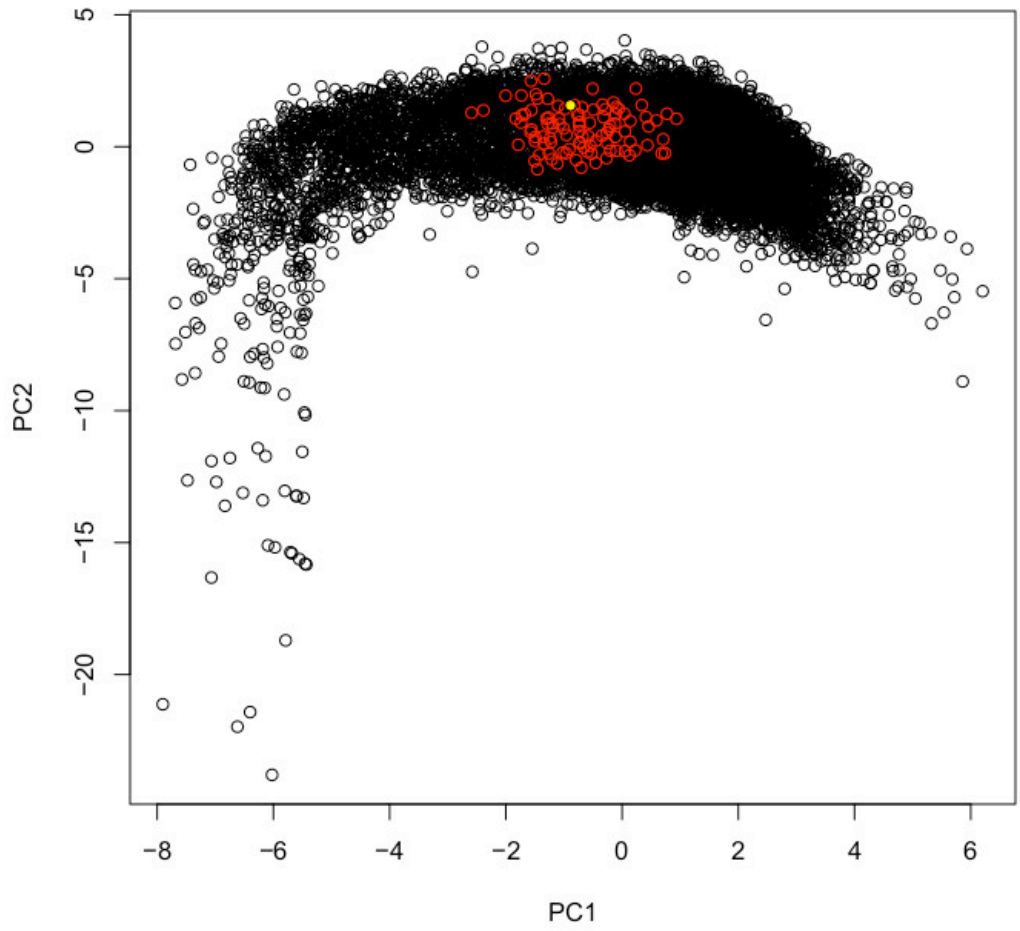


Figure S8. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF08717. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset is inside of the SS of the retained simulations.

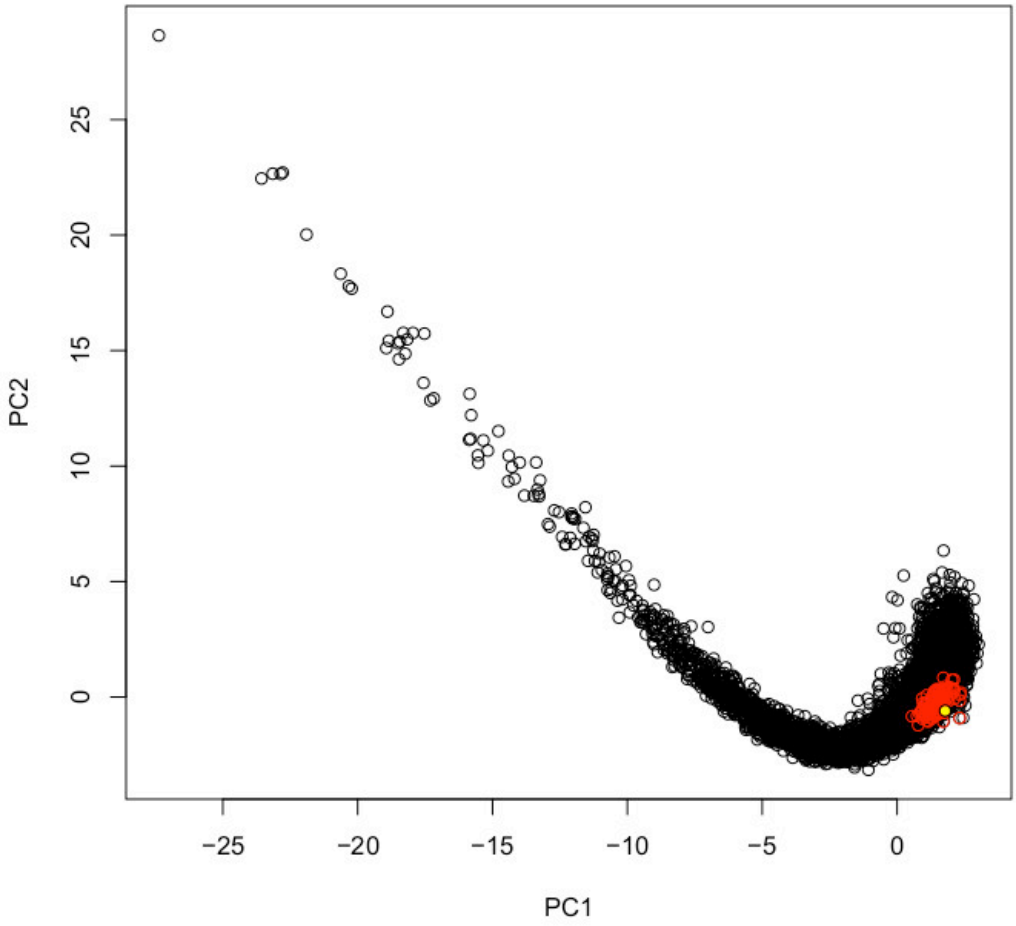


Figure S9. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF09401. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset is inside of the SS of the retained simulations.

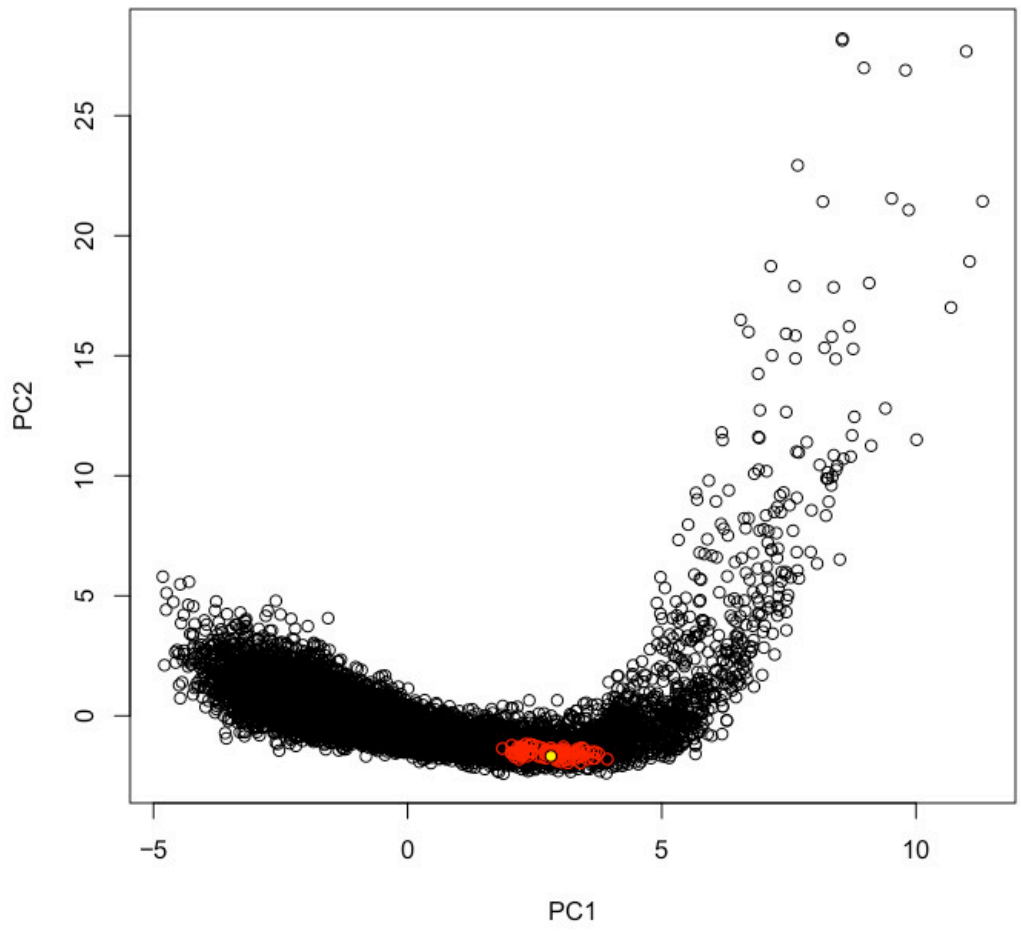


Figure S10. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF11289. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset is inside of the SS of the retained simulations.

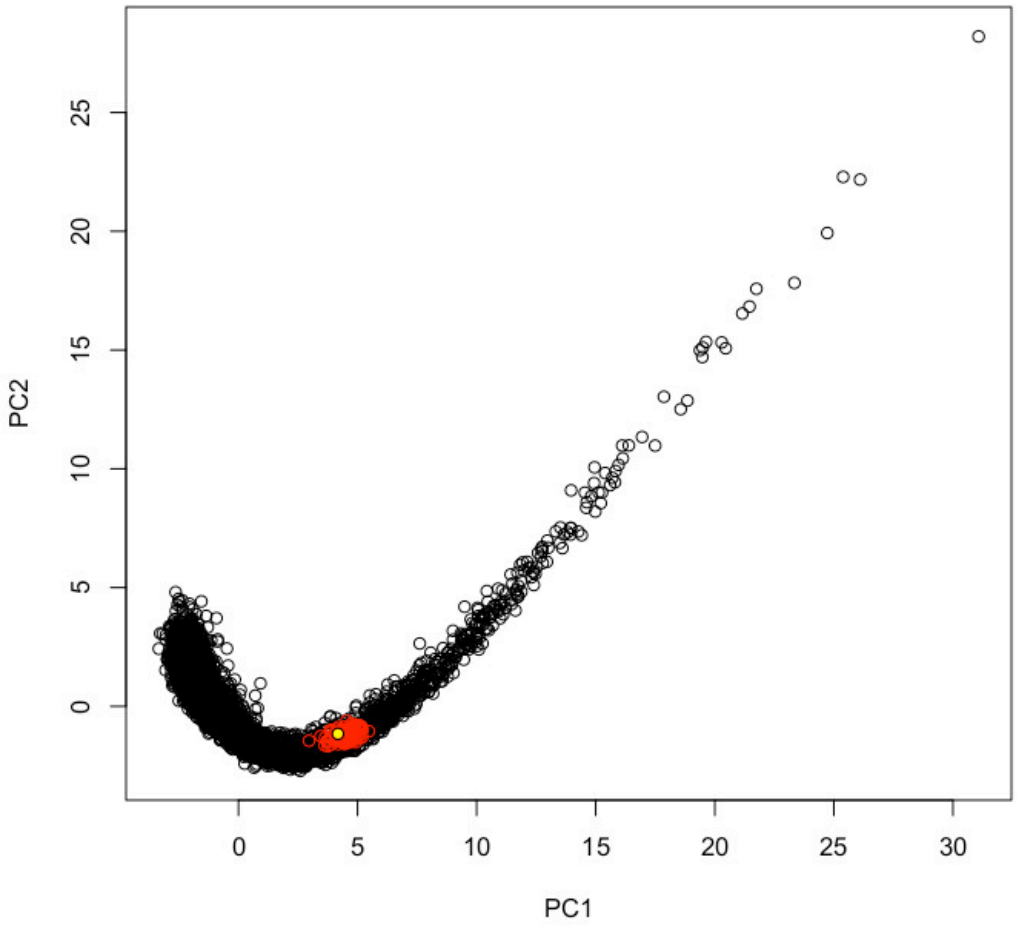


Figure S11. Goodness of fit analyses using PCA based on summary statistics for the PFAM protein family PF09668. The plot shows the two first principal components of a principal component analysis applied to the summary statistics (SS) obtained from the simulated and real data. It includes a sample of all the simulations (black points), all the retained simulations (red points) and the target dataset (yellow point). SS of retained simulations should be inside of SS of the sample of all the simulations. A good fitting of the simulation model with the target dataset implies that SS of the target dataset is inside of the SS of the retained simulations.

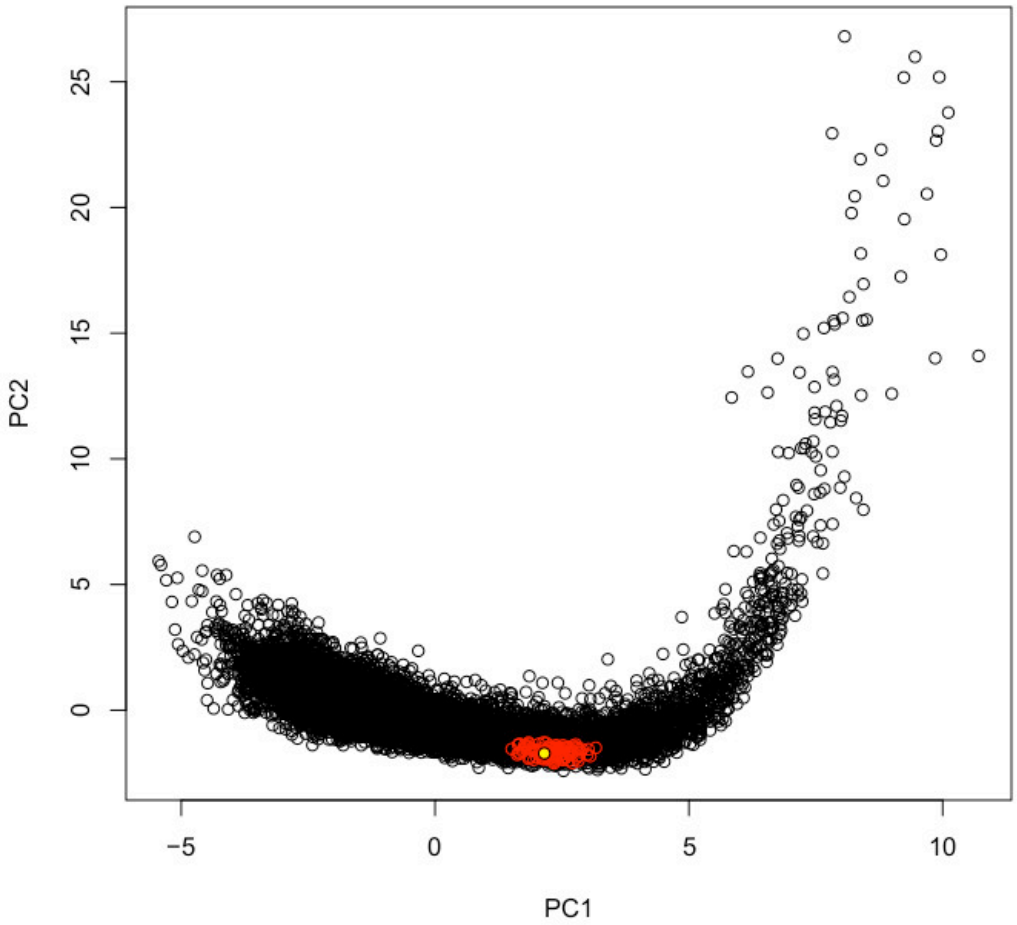


Figure S12. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF02723. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

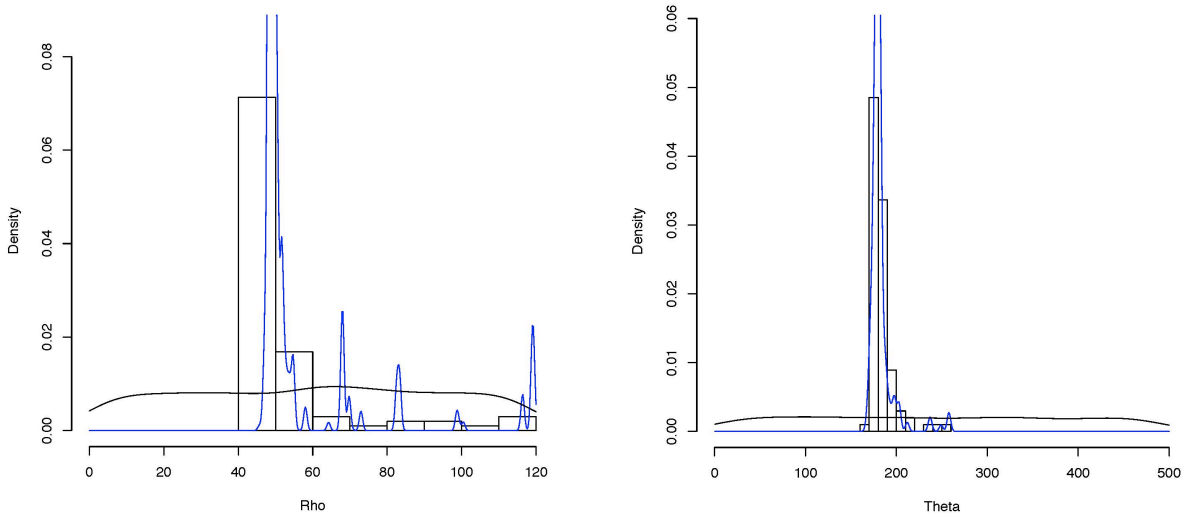


Figure S13. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF06460. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

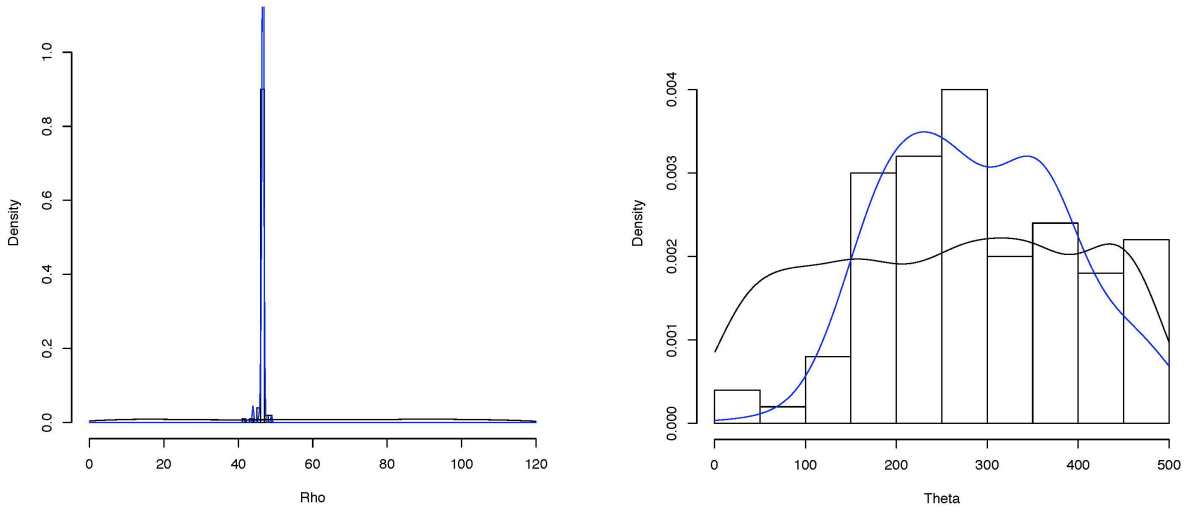


Figure S14. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF04753. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

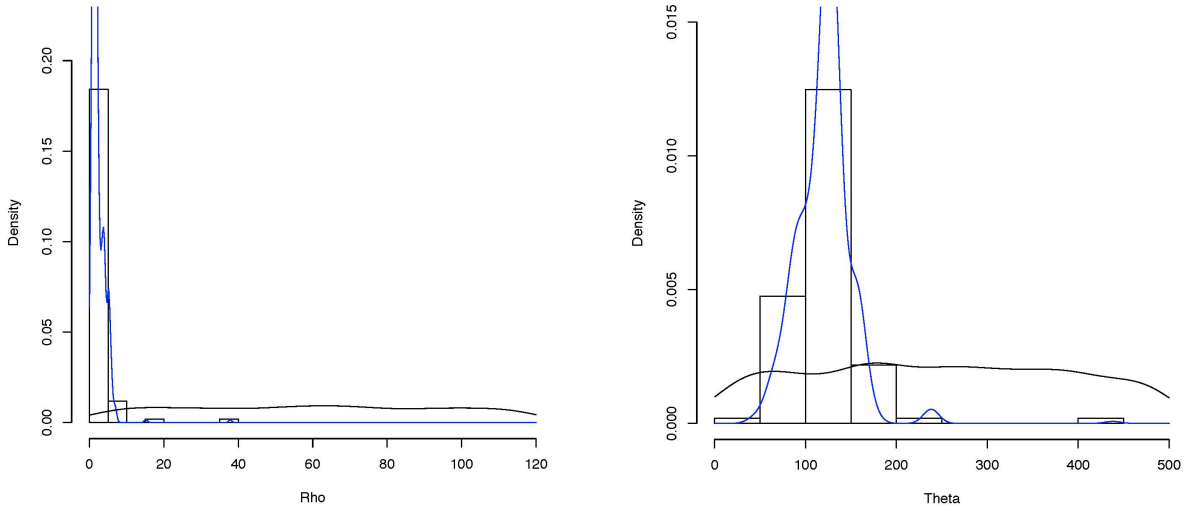


Figure S15. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF08716. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

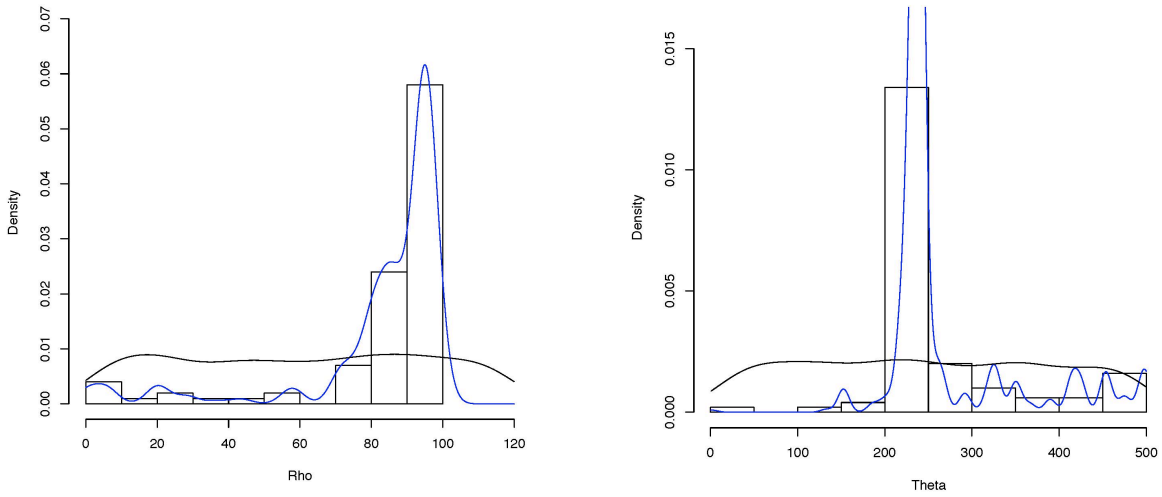


Figure S16. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF08717. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

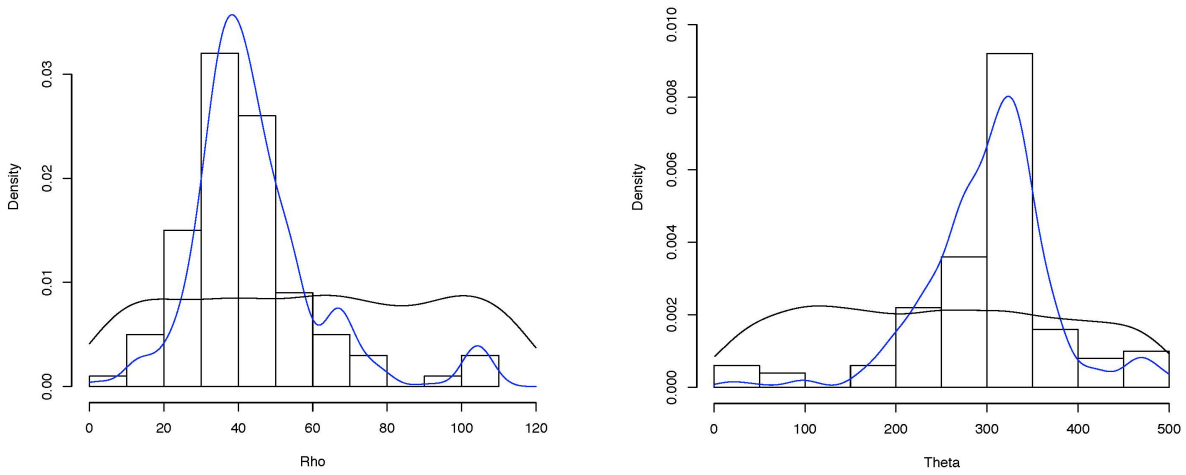


Figure S17. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF09401. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

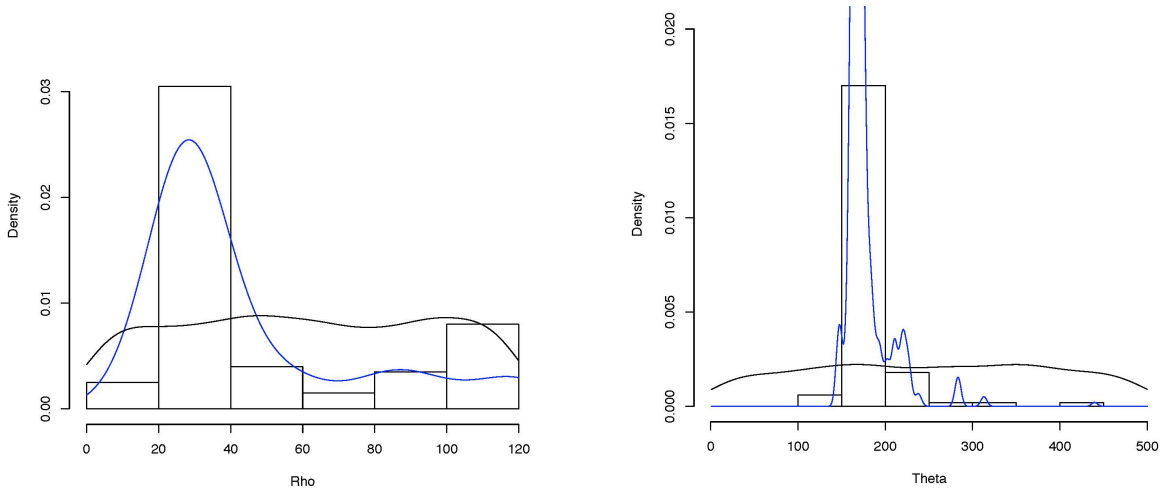


Figure S18. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF11289. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.

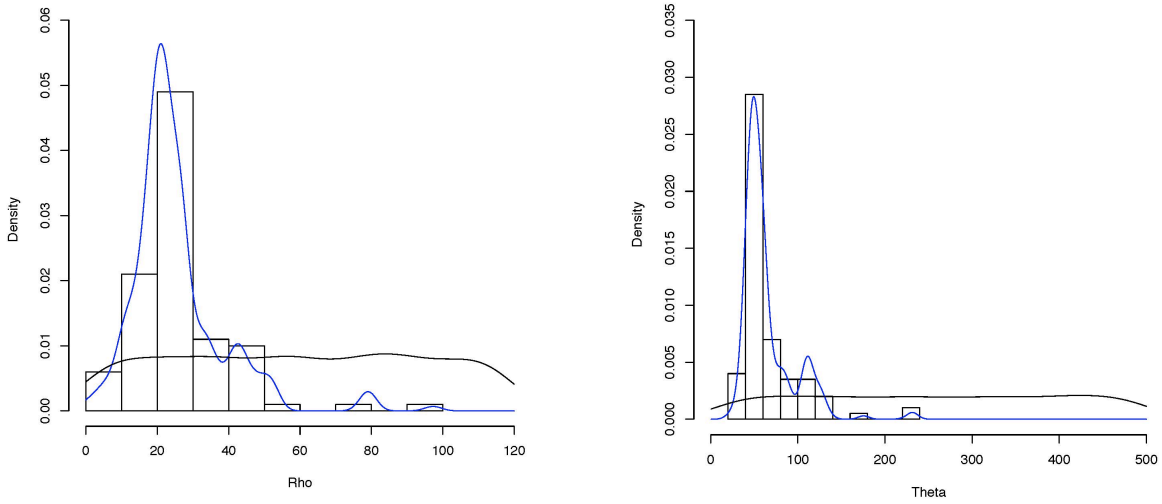
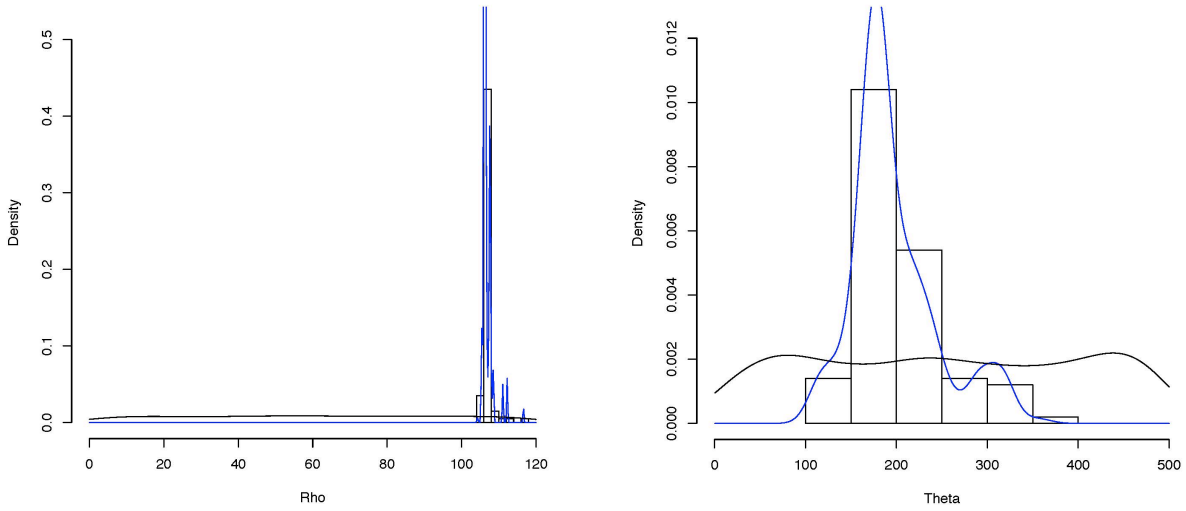


Figure S19. Posterior distribution for the recombination and substitution rates for the PFAM protein family PF09668. The posterior distribution is shown with the blue line and the histogram. The black line represents the prior distribution. Left: Estimation of the recombination rate (ρ). Right: Estimation of the substitution rate (θ). Additional information about the posterior distribution is shown in Table 1.



References

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

Adachi, J. and Hasegawa, M. (1996) MOLPHY version 2.3: programs for molecular phylogenetics based in maximum likelihood. *Comput Sci Monogr*, 28, 1-150.

Adachi, J.*, et al.* (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J Mol Evol*, 50, 348-358.

Arenas, M. (2015) Trends in substitution models of molecular evolution. *Front Genet*, 6, 319.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. A model of evolutionary change in proteins. In: Dayhoff, M.O., editor, *Atlas of protein sequence and structure*. Washington D. C.; 1978. p. 345-352.

Dimmic, M.W.*, et al.* (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol*, 55, 65-73.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89, 10915-10919.

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

Kosiol, C. and Goldman, N. (2005) Different versions of the Dayhoff rate matrix. *Mol Biol Evol*, 22, 193-199.

Le, S.Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol Biol Evol*, 25, 1307-1320.

Muller, T. and Vingron, M. (2000) Modeling amino acid replacement. *J Comput Biol*, 7, 761-776.

Nickle, D.C.*, et al.* (2007) HIV-specific probabilistic models of protein evolution. *PLoS One*, 2, e503.

Whelan, S. and 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.

Yang, Z. Computational Molecular Evolution. Oxford, England.: Oxford University Press; 2006.

Yang, Z., Nielsen, R. and Masami, H. (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15, 1600-1611.