Bioinformatics, 38(1), 2022, 58–64
doi: 10.1093/bioinformatics/btab617
Advance Access Publication Date: 27 August 2021
Original Paper

Phylogenetics

ProteinEvolverABC: coestimation of recombination and substitution rates in protein sequences by approximate Bayesian computation

Miguel Arenas [1,2,3]

[1]CINBIO, Universidade de Vigo, 36310 Vigo, Spain, [2]Universidade de Vigo, Departamento de Bioquimica, Xenetica e Inmunoloxia, 36310 Vigo, Spain and [3]Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain

Received on May 15, 2021; revised on July 24, 2021; editorial decision on August 20, 2021; accepted on August 24, 2021

Abstract

Motivation: The evolutionary processes of mutation and recombination, upon which selection operates, are fundamental to understand the observed molecular diversity. Unlike nucleotide sequences, the estimation of the recombination rate in protein sequences has been little explored, neither implemented in evolutionary frameworks, despite protein sequencing methods are largely used.

Results: In order to accommodate this need, here I present a computational framework, called *ProteinEvolverABC*, to jointly estimate recombination and substitution rates from alignments of protein sequences. The framework implements the approximate Bayesian computation approach, with and without regression adjustments and includes a variety of substitution models of protein evolution, demographics and longitudinal sampling. It also implements several nuisance parameters such as heterogeneous amino acid frequencies and rate of change among sites and, proportion of invariable sites. The framework produces accurate coestimation of recombination and substitution rates under diverse evolutionary scenarios. As illustrative examples of usage, I applied it to several viral protein families, including coronaviruses, showing heterogeneous substitution and recombination rates.

Availability and implementation: *ProteinEvolverABC* is freely available from https://github.com/miguelarenas/proteinevolverabc, includes a graphical user interface for helping the specification of the input settings, extensive documentation and ready-to-use examples. Conveniently, the simulations can run in parallel on multicore machines.

Contact: marenas@uvigo.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Mutation and recombination play a central role in acquiring molecular diversity during the evolution of organisms. Consequently, the analysis of molecular evolution through the estimation of recombination and substitution (fixed mutation) rates constitutes a common object of study in evolutionary biology. Despite a variety of evolutionary methods are available to estimate the recombination rate in DNA sequences (Arenas, 2021; Martin *et al.*, 2011), none of them allow it in protein sequences. However, methods to study molecular evolution in protein sequences are necessary because most of protein sequencing methodologies [i.e. based on RMN and X-ray diffraction, among others (see for a review Edman and Begg, 1967; Smith, 2003), some of them developed before than DNA sequencing (see for a review Strasser, 2010)] and protein databases [i.e. PFAM, PDB and CATH, among others (see for a review Kwon *et al.*, 2006)], do not involve DNA sequences. Analyzing protein evolution allows us to understand the direct consequences of selection from protein stability and function (Bastolla *et al.*, 2007; Pál *et al.*, 2006; Taverna and Goldstein, 2002; Wilke, 2012) and resolve phylogenetic relationships (e.g. Alvarez-Ponce, 2021; Arenas, 2020; Bastolla and Arenas, 2019; Gabaldón, 2005; Gupta, 1998; Pazos and Valencia, 2008), among others (Patthy, 2008). In this concern, molecular signatures of recombination were detected in proteins, with strong selection against intra-domain recombination (compared to recombination in protein domain boundaries) to conserve protein function and folding stability (Meyer *et al.*, 2006; Voigt *et al.*, 2002), but still the presence of recombination should be quantified by the estimation of the recombination rate.

The estimation of the recombination rate is not trivial, especially in data presenting low genetic variation (note that recombination cannot be detected in absence of substitution events) or lack of phylogenetic tree discordance among recombinant fragments [a number of recombination estimation methods are based on this feature (e.g. de Oliveira Martins *et al.*, 2008) but not always the recombined fragments display different evolutionary histories (Arenas and

Posada, 2010)]. Thus, the recombination rate can be underestimated in certain evolutionary scenarios (Martin *et al.*, 2011; Posada and Crandall, 2001). A recent example is the problematic detection of recombination along the SARS-CoV-2 genome due to the relatively low levels of genetic diversity displayed by this virus (VanInsberghe *et al.*, 2021). The estimation of the recombination rate can be affected by the substitution rate and *vice versa* [note that the recombination rate can exceed the substitution rate in some organisms such as HIV (Perez-Losada *et al.*, 2015; Shriner *et al.*, 2004) and affect estimations of diverse evolutionary parameters such as the synonymous and non-synonymous substitution rates (Anisimova *et al.*, 2003; Arenas and Posada, 2010; Del Amparo *et al.*, 2021; Shriner *et al.*, 2003)]. Therefore, the coestimation of both parameters is convenient due to accounting for their parameters interactions (Lopes *et al.*, 2014).

Indeed, the estimation of the recombination rate could be more complex at the protein level than at the nucleotide level because of the usually lower levels of sequence diversity in proteins compared to DNA (i.e. note that synonymous substitutions are not observed at the protein level). However, the accuracy of recombination rate estimates from protein sequences respect to estimates from DNA sequences was not yet formally evaluated. Next, at the protein level, probabilistic phylogenetic methods to analyze protein sequences traditionally mimic the substitution process with substitution models of evolution composed by a 20 × 20 exchangeability matrix of relative rates of change among amino acids and the 20 amino acid frequencies (Arenas, 2015; Yang, 2006), which differ from the non-coding DNA substitution models (4 × 4 exchangeability matrix and 4 nucleotide frequencies) and coding DNA substitution models (61 × 61 exchangeability matrix and 61 codon frequencies, stop codons are often excluded) (Arenas, 2015; Yang, 2006). These models can also include variation of the rate of change among sites and a proportion of invariable sites (Yang *et al.*, 1998). In this concern, it is known that accounting for the best fitting substitution model of protein evolution is convenient for evolutionary inferences (e.g. Arenas and Bastolla, 2020; Bordner and Mittelmann, 2014; Lemmon and Moriarty, 2004) although there is a recent discussion concerning its specific application to phylogenetic tree reconstructions (Spielman, 2020). In any case, accounting for the substitution model of evolution that best fits the data in the estimation of the recombination rate with probabilistic methods was found necessary (Lopes *et al.*, 2014). However, it is yet unclear if the estimation of the recombination rate with probabilistic methods could be affected by analyzing different molecular markers (i.e. molecular evolution is mimicked by different substitution models of evolution).

In general, there is a need for methods that coestimate multiple evolutionary parameters (especially for evolutionary analyses of protein sequences) and their implementation into useful evolutionary frameworks. Coestimation methods usually require algorithms based on complex models of evolution and present limitations concerning the derivation of analytical formulae or can lead to computationally too expensive evaluations of likelihood functions (Beaumont and Rannala, 2004). In such situations, an alternative is the approximate Bayesian computation (ABC) approach (Beaumont, 2010; Csillery *et al.*, 2010), which considers computer simulations under complex evolutionary scenarios followed by a statistical adjustment for the estimation of parameters without the need of a likelihood function. Hence, taking advantage of ABC, we previously developed an ABC framework for the joint estimation of recombination, selection [by the non-synonymous/synonymous rate ratio, *dN/dS* (Del Amparo *et al.*, 2021)] and substitution rates in coding sequences that outperformed other methods (including maximum-likelihood methods) in the estimation of the recombination rate (Arenas *et al.*, 2015; Lopes *et al.*, 2014). Here, I adapted that ABC method to the evolutionary analysis of protein sequences and I implemented it in a user-friendly computational framework called *ProteinEvolverABC*. *ProteinEvolverABC* provides the coestimation of substitution and recombination rates from multiple alignments of protein sequences under ABC based on computer simulations of coalescent evolutionary histories (coalescent modified with recombination, demographics and serial sampling) followed by protein evolution (under a variety of substitution models of evolution). The framework includes diverse nuisance parameters (simulated under user-specified prior distributions but not estimated) that are optional to provide a more realistic modeling (by accounting for their uncertainty), multiple summary statistics designed to extract the evolutionary information from protein sequences and ABC estimation under both rejection and multiple linear regression approaches. The framework was evaluated under different levels of substitution and recombination rates and showed an acceptable estimation performance. In addition, the framework was able to accurately estimate the recombination rates observed in coding sequences. As illustrative practical examples, I applied it to the analysis of some viral protein families of general interest.

2 System and methods

2.1 *ProteinEvolverABC* framework

The framework *ProteinEvolverABC* follows the standards of ABC to perform parameters estimation through four main steps: specification of input information including prior distributions, simulation of evolutionary histories and protein sequences evolution (upon those evolutionary histories), computation of summary statistics and, joint estimation of recombination and substitution rates with the rejection and regression statistical approaches. Details about these steps are provided below.

(1) Specification of input information. Despite the ABC approach has some advantages respect to other analytical approaches in terms of flexibility when dealing with complex models (Arenas, 2015; Beaumont, 2010), it also requires some decisions (to be made by the user) that can affect the estimations. One of them is the specification of prior distributions for the parameters to be estimated and, optionally, for nuisance parameters. A list with the main parameters implemented in the framework is shown in Supplementary Table S1, Supplementary Material. The prior distribution for any parameter should be wide and sampled enough to include a sufficient number of simulated data with parameter values close to the true value (Beaumont, 2010). The user of *ProteinEvolverABC* has to specify prior distributions for the recombination and substitution rates and, optionally, for some nuisance parameters (i.e. amino acid frequencies, heterogeneity in the substitution rate across sites and proportion of invariable sites) that could help provide more realistic simulations. Indeed, the user has to specify a substitution model of protein evolution [that can be previously identified with frameworks such as *ProtTest* (Darriba *et al.*, 2011)] and some parameters for the ABC estimation such as the tolerance (number of simulations used for the estimation with summary statistics closer to the summary statistics of the study dataset) or the estimation approach (i.e. rejection and multiple linear regression) (Beaumont *et al.*, 2002). The software documentation includes detailed recommendations for specifying every input parameter properly.

(2) The simulation of protein sequences is internally performed with an adapted version of the simulator *ProteinEvolver* (Arenas *et al.*, 2013) to ABC. This simulator produces multiple alignments of protein sequences by the following two steps: (i) simulation of evolutionary histories with the coalescent modified with recombination and demographics (Hudson, 1983; Kingman, 1982), where note that the coalescent allows a rapid computation (Arenas, 2012) that is convenient for ABC methods that usually require a large number of simulated data to properly explore the parameters landscape (Beaumont, 2010) and, (ii) simulation of protein sequence evolution upon the previously simulated evolutionary histories under a user-specified substitution model of protein evolution (Supplementary Table S1) (Arenas, 2012; Yang, 2006), optionally including substitution rate variation among sites and proportion of invariable sites (Yang, 1994). *ProteinEvolver* is based on previous coalescent simulators (Arenas and Posada, 2007; Arenas and Posada, 2010) and has been widely used and validated (e.g. Arenas and Bastolla, 2020; Arenas *et al.*, 2013; Arenas *et al.*, 2017; Pascual-Garcia *et al.*, 2019). The simulations are parameterized sampling from the user-specified prior distributions (Supplementary Table S1) and, conveniently, can run in parallel on multicore computers.

(3) The framework *ProteinEvolverABC* implements a total of 16 summary statistics to extract the evolutionary information from the observed and simulated protein data (Supplementary Table S2; Supplementary Material). The summary statistics comprise the mean, standard deviation, skewness and kurtosis of diversity, heterozygosity and pairwise sequence identity. They also include the number of segregating sites and three fast recombination tests [pairwise homoplasy index (Bruen *et al.*, 2006), neighbor similarity score (Jakobsen and Easteal, 1996) and maximum chi-squared (Maynard Smith, 1992)]. Some of these summary statistics (i.e. heterozygosity) are based on a previous ABC study to estimate substitution and recombination rates from codon sequences with acceptable error (Lopes *et al.*, 2014) and were adapted here to the analysis of protein sequences.

(4) Finally, *ProteinEvolverABC* estimates the substitution and recombination rates with the *abc R* library (Csillery *et al.*, 2012). In particular, the framework estimates (i) the population recombination rate $\rho = 4Nrl$ (McVean *et al.*, 2004), where N is the effective population size, r is the recombination rate per site and l is the sequence length; (ii) the population amino acid substitution rate $\theta = 4N\mu l$, where μ is the substitution rate per site. Before performing the estimation, the framework includes some goodness-of-fit analyses that are useful to evaluate the fitting of the simulation model with the observed data and the information provided by every summary statistic. The implemented goodness-of-fit analyses include (i) distance between the distribution of summary statistic values from retained simulations and summary statistic value of the observed data for every summary statistic, (ii) evaluation of summary statistic values (for every summary statistic) with corresponding parameters (recombination and substitution rates) values of the simulation and, (iii) the two first principal components of a principal component analysis based on all the summary statistics from a sample of all the simulations, retained simulations and observed data (Csillery *et al.*, 2012). Further details about the implemented goodness-of-fit analyses can be found in the documentation and illustrative examples distributed with *ProteinEvolverABC*. The ABC estimation can be performed with the rejection and weighted multiple linear regression approaches, which are well established in the field (Beaumont, 2010; Beaumont *et al.*, 2002; Blum and François, 2010; Csillery *et al.*, 2010).

In practice, the user has to specify the number of simulations, the prior distributions for the parameters of interest, the tolerance level and the estimation approach, as well as other minor options. The framework is distributed with a documentation that includes detailed recommendations, and illustrative examples, about the specifications (see also next section). In general, a total of 50 000 simulations can be a good starting point (see next section), but of course some datasets (especially those presenting low sequence identity) may require more simulations.

The framework *ProteinEvolverABC* runs on the command line but includes a graphical user interface (GUI) to facilitate the input parameterization of the whole estimation procedure. Because the simulation step is usually slow, the framework can run the simulations in parallel on multicore machines, which allows a reduction of the computation time (see later). *ProteinEvolverABC* consists of a pipeline written in Java (GUI), C, Perl and R, and it is freely available from https://github.com/miguelarenas/proteinevolverabc. The package includes executable files, source code, detailed documentation and illustrative examples with input and output files.

2.2 *ProteinEvolverABC* validation

Despite a previous ABC study already showed accurate coestimation of recombination and substitution rates (even more accurate than maximum-likelihood methods) in the analysis of protein-coding data (Lopes *et al.*, 2014), here I performed an evaluation of *ProteinEvolverABC* under different evolutionary scenarios based on multiple combinations of substitution and recombination rates. I simulated protein sequences under different levels of ρ [0, 30, 60 and 90, which are levels of recombination that can be observed in nature (e.g. Arenas *et al.*, 2016; Carvajal-Rodriguez *et al.*, 2006; Lopes *et al.*, 2014)] and θ [50, 100, 200, 300 and 400, which involve a wide range of sequence identity (from above 0.9 to below 0.6) fitting with observations in nature (e.g. Cornish-Bowden, 1977; Pascual-Garcia *et al.*, 2010)], for alignments of 25 sequences with 500 amino acids, assuming a fixed effective population size of 1000 individuals (Carvajal-Rodriguez *et al.*, 2006) and the JTT empirical substitution model of protein evolution (Jones *et al.*, 1992). For every combination of parameters [$4 \times 5 = 20$ combinations, which are more combinations than those evaluated in (Lopes *et al.*, 2014)], I simulated 100 multiple sequence alignments (test datasets). Next, the framework was used to estimate ρ and θ in the test datasets under ABC based on a total of 50 000 simulations parameterized under the following prior distributions: $\rho = Uniform(0,120)$ and $\theta = Uniform(0,500)$, which are prior distributions wider than those used in (Lopes *et al.*, 2014) and encompass values that are commonly observed in real data (e.g. Carvajal-Rodriguez *et al.*, 2006; Monteiro *et al.*, 2021; Perez-Losada *et al.*, 2011; Perez-Losada *et al.*, 2009; Stumpf and McVean, 2003). Following previous works (Arenas *et al.*, 2015; Lopes *et al.*, 2014), ABC estimates were obtained assuming an acceptance rate of 0.2% (100 points) in the adjustment with the rejection and the weighted multiple linear regression approaches. The validation of *ProteinEvolverABC* showed that the parameter estimates are generally accurate (the true values usually fell within the 95% confidence interval of the estimates; Fig. 1) and in good agreement with previous ABC methods (Arenas *et al.*, 2015; Lopes *et al.*, 2014). In particular, the results indicate that both parameters can be accurately estimated using 50 000 computer simulations under any implemented ABC estimation approach (rejection and multiple linear regression). As expected, the estimation of the recombination rate was more difficult at low levels of substitution and recombination rates (in agreement with Posada and Crandall, 2001) (see Fig. 1A with $\theta = 50$ and $\rho = 30$) showing a small overestimation. Increasing the number of simulations to 100 000 overall improved the accuracy of the estimations (also reduced the overestimation of the recombination rate at low levels of substitution and recombination rates) (Supplementary Fig. S1; Supplementary Material) but still, in general, 50 000 simulations were enough to obtain acceptable estimates concerning accuracy.

Since additional comparisons of accuracy of estimated recombination and substitution rates between *ProteinEvolverABC* and other evolutionary frameworks cannot be performed yet due to the lack of frameworks implementing these estimations in protein sequences, next I evaluated the accuracy of *ProteinEvolverABC* in the estimation of the recombination and substitution rates present in coding data. The aim is to evaluate the bias of estimating these parameters from protein sequences respect to the codon level. I applied the *CoalEvol* framework (Arenas and Posada, 2014) to simulate coding sequences under different levels of ρ (0 and 60) and θ (50, 200 and

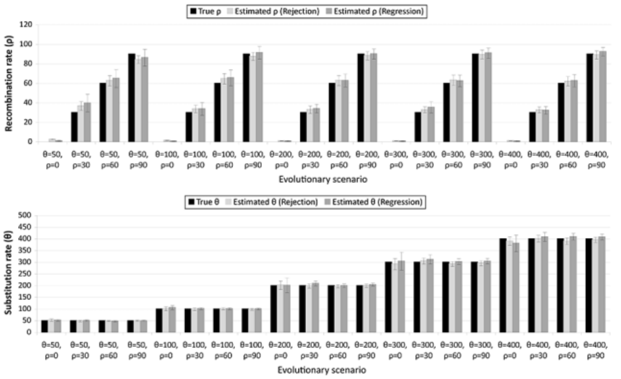

Fig. 1. Accuracy of *ProteinEvolverABC* in the estimation of recombination and substitution rates under different evolutionary scenarios and based on ABC with 50 000 simulations. For each studied combination of ρ and θ (evolutionary scenario based on 100 simulations) the figure shows the estimates of 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 50 000 simulations) and error bars indicate the 95% confidence interval

400) for alignments of 25 sequences with 500 codons (1500 nucleotides), assuming a fixed effective population size of 1000 individuals (Carvajal-Rodriguez *et al.*, 2006) and the GY94 substitution model of codon evolution with $dN/dS = 1$ (Goldman and Yang, 1994). For every combination of parameters ($2 \times 3 = 6$ combinations), I simulated 100 multiple sequence alignments. Next, the multiple alignments of codon sequences were translated to protein sequences considering the universal genetic code. Finally, *ProteinEvolverABC* (using the previously described parameterization of prior distributions, total number of simulations and acceptance rate) was applied to estimate the recombination and substitution rates. The accuracy of *ProteinEvolverABC* in estimating recombination and substitution rates observed in coding data differed between these parameters. The recombination rate was accurately estimated (Fig. 2), suggesting that the framework can capture signatures of recombination observed at the nucleotide level. However, the substitution rate was underestimated, especially when it is large (Fig. 2). This means that the substitution rate occurring at the nucleotide level cannot be estimated at the protein level. This is not surprising because note that synonymous codon substitutions are not transferred to protein sequences. In any case, users of *ProteinEvolverABC* should take this issue into account.

2.3 Illustrative application to viral protein families

As illustrative practical examples of the application of *ProteinEvolverABC* (and also to provide an idea about running times) I analyzed 8 real datasets of viral protein families (Table 1). These protein families include coronaviruses [motivated by the current pandemic and since they have shown genetic signatures of recombination (e.g. Graham and Baric, 2010; Zhu *et al.*, 2020)] and aspartyl proteases [due to its interest as a molecular target of antiretroviral therapies (e.g. Ghosh *et al.*, 2016; van Leeuwen *et al.*, 2003) and where recombination could favor the emergence of resistance to therapies (e.g. Fraser, 2005; Shi *et al.*, 2010)]. Indeed, note that the coalescent can be used to mimic virus evolution (Bouckaert *et al.*, 2019; Pybus and Rambaut, 2009). All the protein families were downloaded from the *PFAM* database (Finn *et al.*, 2014) and realigned with *MAFFT* (Katoh and Standley, 2013). I ran a total of 50 000 simulations under the prior distributions used for the analysis of the simulated data (see above). The analyses of these datasets took several hours and, as expected, they ran faster when using more cores (Supplementary Fig. S2; Supplementary Material). In particular, increasing the number of processors always reduced computer times, especially in big datasets where simulations (which are parallelized) require more time. However, the decrease of computer time as a function of the number of processors was not linear (Supplementary Fig. S2) because other ABC phases could not be parallelized. Also as expected, bigger datasets, with more and longer sequences, required longer computer times (Supplementary Figs S2 and S3AB; Supplementary Material). By contrast, the computer time was similar for different substitution models of protein evolution (Supplementary Fig. S3C), which was also expected since all the empirical substitution models of protein evolution present exchangeability matrices with same dimension (20×20) (Arenas, 2015; Yang, 2006). The goodness-of-fit analyses produced acceptable results with summary statistics of real data within the distribution of summary statistics of simulated data (including retained simulated data; Supplementary Figs S4–S11, Supplementary Material). The results (Table 1) showed that both recombination and substitution rates are heterogeneous among coronavirus protein families and that the aspartyl protease protein family presents a high recombination rate in agreement with previous observations (e.g. Speranskaya *et al.*, 2012; Sun *et al.*, 2019).

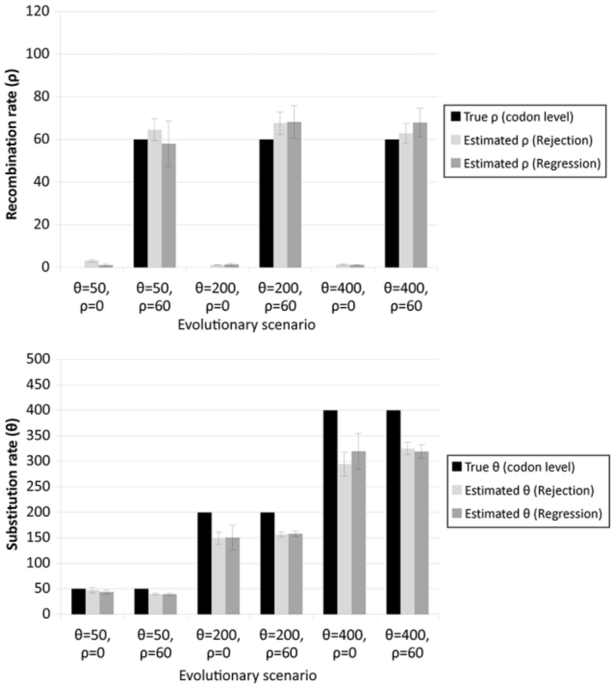

Fig. 2. Accuracy of *ProteinEvolverABC* in the estimation of recombination and substitution rates present in coding data. For each studied combination of ρ and θ (evolutionary scenario based on 100 simulations) the figure shows the estimates of ρ (above) and θ (below). The black bars indicate the true value (recombination and substitution rates present in coding data). 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 50 000 simulations) and error bars indicate the 95% confidence interval

3 Discussion

There is a general need for evolutionary frameworks implementing the estimation of the recombination rate in protein data. Motivated by this aspect, here I introduce an ABC framework for the coestimation of substitution and recombination rates from a multiple alignment of protein sequences. The user can specify diverse prior distributions for a variety of evolutionary parameters (including nuisance parameters, Supplementary Table S1) and perform ABC estimation under both rejection and multiple linear regression approaches considering protein data simulated under an empirical substitution model of protein evolution (among a variety of implemented substitution models; Table S1) upon coalescent evolutionary histories with and without recombination. An advantage of using computer simulations in an analytical approach is to consider the influence of parameters interactions in their estimation, which is often omitted in other analytical methods such as maximum likelihood (Li and Stephens, 2003). However, ABC estimation methods usually require computational efforts due to the large number of simulations necessary to cover the parameters landscape described by the prior distributions (although if the user has prior information can reduce the range of the prior distribution and/or choose an informative prior distribution sampling frequently near to the true parameter value, thus reducing the required number of simulations), but the accuracy of the estimates may justify the effort. Similarly to the preceding ABC methods of *ProteinEvolverABC* (Arenas *et al.*, 2015; Lopes *et al.*, 2014), I found that *ProteinEvolverABC* provides accurate estimates under different evolutionary conditions. In addition, it can estimate the recombination rate that is observed at the nucleotide level, although (as expected) it underestimates the substitution rate observed at the nucleotide level. However, but comparably to other estimation methods, the user has to consider that the accuracy of the estimates provided by *ProteinEvolverABC* could be affected by processes ignored in the framework such as a small population size (note that the coalescent requires a much larger population size compared to the sample size), among others. Moreover, analyzing protein fragments (i.e. domains) is not recommended, especially if the fragment is conserved, because of the lack of evolutionary information that could produce statistically unsupported estimates. In general, the user of an evolutionary framework (including *ProteinEvolverABC*) should be cautious interpreting the results, take into account the estimated credible interval

Table 1. Recombination and substitution rates estimated with *ProteinEvolverABC* for the studied protein families

Description of the protein family	PFAM code, number of sequences, sequence length	Mode	Mean	Median	97.5% HPDI
Coronavirus small envelope protein E	PF02723, 27, 82	(ρ) 48.68 (θ) 179.40	(ρ) 53.95 (θ) 182.51	(ρ) 48.79 (θ) 179.44	(ρ) 47.64–119.00 (θ) 170.45–237.06
Coronavirus 2′-O-methyltransferase	PF06460, 5, 299	(ρ) 46.67 (θ) 230.62	(ρ) 46.53 (θ) 290.14	(ρ) 46.63 (θ) 272.73	(ρ) 45.49–47.01 (θ) 132.93–477.30
Coronavirus non-structural protein NS12.7	PF04753, 9, 109	(ρ) 1.30 (θ) 125.72	(ρ) 2.27 (θ) 120.98	(ρ) 1.69 (θ) 123.10	(ρ) 0.40–5.82 (θ) 67.07–162.22
Coronavirus replicase NSP7	PF08716, 5, 83	(ρ) 95.02 (θ) 237.74	(ρ) 82.39 (θ) 267.56	(ρ) 92.75 (θ) 238.56	(ρ) 5.89–98.06 (θ) 186.48–495.16
Coronavirus replicase NSP8	PF08717, 5, 198	(ρ) 38.45 (θ) 323.13	(ρ) 44.99 (θ) 303.82	(ρ) 41.83 (θ) 309.27	(ρ) 14.59–103.27 (θ) 167.88–468.66
Coronavirus RNA synthesis protein NSP10	PF09401, 6, 122	(ρ) 28.45 (θ) 166.01	(ρ) 44.19 (θ) 176.39	(ρ) 29.52 (θ) 169.29	(ρ) 19.45–119.99 (θ) 146.82–237.30
Betacoronavirus viroporin	PF11289, 5, 273	(ρ) 21.03 (θ) 49.43	(ρ) 25.49 (θ) 65.17	(ρ) 22.00 (θ) 54.80	(ρ) 8.34–51.98 (θ) 36.07–127.61
Aspartate protease	PF09668, 10, 124	(ρ) 106.23 (θ) 178.27	(ρ) 106.61 (θ) 197.65	(ρ) 106.26 (θ) 183.42	(ρ) 105.58–111.09 (θ) 115.15–357.84

Note: The first column provides information about the protein family and the second column includes the PFAM code, number of sequences and sequence length in amino acids, respectively. For each parameter (recombination rate ρ or substitution rate θ), the table presents the mode, mean, median and 97.5% HPDI (highest posterior density interval) of the estimated posterior distribution. The full posterior distributions are shown in Supplementary Figures S12–S19, Supplementary Material.

and consider that other biological processes can affect the estimates. Indeed, care should be taken when specifying ABC settings such as number of simulations, prior distributions or tolerance. In this concern, *ProteinEvolverABC* is distributed as a package with a GUI that helps to specify the entire analysis, a detailed documentation that includes theory, usage and recommendations about the input parameters and, several illustrative practical examples. Concerning the prior distributions, in absence of previous biological information the user may specify uniform distributions. However, if there is any prior information (i.e. a suspected low substitution rate) the user can define more precise prior distributions that could improve the accuracy of the estimation (Beaumont, 2010). In any case, the prior distributions should include biologically reasonable parameter values and prior expectations of the parameter values in the study data. In general, I recommend applying by default the settings presented in the validation section (50 000 simulations and acceptance rate of 0.2%). However, for complex real data (i.e. data presenting very high or low genetic diversity), increasing the number of simulations and acceptance rate [i.e. up to 10 times (following Lopes *et al.*, 2014)] can be useful to obtain more accurate estimates. Indeed, repeating an analysis increasing the number of simulations, exploring different acceptance rates or even a different set of the implemented summary statistics (evaluated with the goodness-of-fit analyses), can help in finding a proper input parameterization to obtain reliable estimates. As illustrative examples, the framework was applied to several viral protein families, especially from coronavirus. Coronavirus protein families showed contrasting recombination and substitution rates. It is remarkable the low recombination rate estimated for the protein families NS12.7 (studies involving recombination in this protein were not found) and viroporins [in agreement with the minor recombination events detected in (Lulla and Firth, 2020) and that can be expected due to conservation for maintaining viral pathogenicity (Nieva *et al.*, 2012) and thus being ideal target for antiviral treatments (Nieva *et al.*, 2012)] and, the large recombination rate estimated for the replicase NSP7 [where recombinant forms of SARS-CoV NSP7 and NSP8 could greatly improve the association of these proteins (te Velthuis *et al.*, 2012)]. Concerning the aspartyl protease protein family, it presented a high recombination rate that agrees with previous studies (e.g. Speranskaya *et al.*, 2012; Sun *et al.*, 2019).

Data Availability

ProteinEvolverABC is freely available at https://github.com/miguelarenas/proteinevolverabc.

Funding

This work was supported by the Spanish Ministerio de Ciencia e Innovación through the Grants [RYC-2015-18241] and [PID2019-107931GA-I00]. Funding for open access charge: Universidade de Vigo/CISUG. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

References

Alvarez-Ponce,D. (2021) Richard Dickerson, molecular clocks, and rates of protein evolution. *J. Mol. Evol.*, **89**, 122–126.

Anisimova,M. *et al.* (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**, 1229–1236.

Arenas,M. (2012) Simulation of molecular data under diverse evolutionary scenarios. *PLoS Comput. Biol.*, **8**, e1002495.

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

Arenas,M. (2020) Protein evolution in the flaviviruses. *J. Mol. Evol.*, **88**, 473–476.

Arenas,M. (2021) Computational analysis of recombination in viral nucleotide sequences. In: Bamford D. and Zuckerman M. (eds.) *Encyclopedia of Virology*, 4th edn., Academic Press, Oxford, pp. 108–115.

Arenas,M. and Bastolla,U. (2020) ProtASR2: ancestral reconstruction of protein sequences accounting for folding stability. *Methods Ecol. Evol.*, **11**, 248–257.

Arenas,M. *et al.* (2013) Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics*, **29**, 3020–3028.

Arenas,M. *et al.* (2015) CodABC: a computational framework to coestimate recombination, substitution, and molecular adaptation rates by approximate Bayesian computation. *Mol. Biol. Evol.*, **32**, 1109–1112.

Arenas,M. *et al.* (2016) Influence of mutation and recombination on HIV-1 in vitro fitness recovery. *Mol. Phylogenet. Evol.*, **94**, 264–270.

Arenas,M. and Posada,D. (2007) Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography. *BMC Bioinformatics*, **8**, 458.

Arenas,M. and Posada,D. (2010) Coalescent simulation of intracodon recombination. *Genetics*, **184**, 429–437.

Arenas,M. and Posada,D. (2014) Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. *Mol. Biol. Evol.*, **31**, 1295–1301.

Arenas,M. *et al.* (2017) ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability. *Syst. Biol.*, **66**, 1054–1064.

Bastolla,U. and Arenas,M. (2019) The influence of protein stability on sequence evolution: applications to phylogenetic inference. In: Sikosek T. (ed.) *Computational Methods in Protein Evolution.* Springer, New York, pp. 215–231.

Bastolla,U. *et al.* (2007) *Structural Approaches to Sequence Evolution.* Springer, Berlin, Heidelberg.

Beaumont,M.A. (2010) Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.*, **41**, 379–405.

Beaumont,M.A. and Rannala,B. (2004) The Bayesian revolution in genetics. *Nat. Rev. Genet.*, **5**, 251–261.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Blum,M.G.B. and François,O. (2010) Non-linear regression models for approximate Bayesian computation. *Stat. Comput.*, **20**, 63–73.

Bordner,A.J. and Mittelmann,H.D. (2014) A new formulation of protein evolutionary models that account for structural constraints. *Mol. Biol. Evol.*, **31**, 736–749.

Bouckaert,R. *et al.* (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **15**, e1006650.

Bruen,T.C. *et al.* (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics*, **172**, 2665–2681.

Carvajal-Rodriguez,A. *et al.* (2006) Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. *Mol. Biol. Evol.*, **23**, 817–827.

Cornish-Bowden,A. (1977) Assessment of protein sequence identity from amino acid composition data. *J. Theor. Biol.*, **65**, 735–742.

Csillery,K. *et al.* (2010) Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.*, **25**, 410–418.

Csillery,K. *et al.* (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.*, **3**, 475–479.

Darriba,D. *et al.* (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164–1165.

de Oliveira Martins,L. *et al.* (2008) Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. *PLoS One*, **3**, e2651.

Del Amparo,R. *et al.* (2021) Analysis of selection in protein-coding sequences accounting for common biases. *Brief. Bioinf.*, bbaa431.

Edman,P. and Begg,G. (1967) A protein sequenator. *Eur. J. Biochem.*, **1**, 80–91.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Fraser,C. (2005) HIV recombination: what is the impact on antiretroviral therapy? *J. R. Soc. Interface*, **2**, 489–503.

Gabaldón,T. (2005) Evolution of proteins and proteomes: a phylogenetics approach. *Evol. Bioinf.*, **1**, 51–61.

Ghosh,A.K. *et al.* (2016) Recent progress in the development of HIV-1 protease inhibitors for the treatment of HIV/AIDS. *J. Med. Chem.*, **59**, 5172–5208.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

Graham,R.L. and Baric,R.S. (2010) Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J. Virol.*, **84**, 3134–3146.

Gupta,R.S. (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev. MMBR*, **62**, 1435–1491.

Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.

Jakobsen,I.B. and Easteal,S. (1996) A program for calculating and displaying compatibility matrices as an aid to determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.*, **12**, 291–295.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kingman,J.F.C. (1982) The coalescent. *Stochastic Processes Appl.*, **13**, 235–248.

Kwon,M.S. *et al.* (2006) Protein databases. In: Ganten,D. and Ruckpaul,K (eds.) *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine.* Springer, Berlin, pp. 1483–1487.

Lemmon,A.R. and Moriarty,E.C. (2004) The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.*, **53**, 265–277.

Li,N. and Stephens,M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Lopes,J.S. *et al.* (2014) Coestimation of Recombination, Substitution and Molecular Adaptation rates by approximate Bayesian computation. *Heredity*, **112**, 255–264.

Lulla,V. and Firth,A.E. (2020) A hidden gene in astroviruses encodes a viroporin. *Nat. Communic.*, **11**, 4070.

Martin,D.P. *et al.* (2011) Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.*, **11**, 943–955.

Maynard Smith,J. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.

McVean,G.A. *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Meyer,M.M. *et al.* (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Select.*, **19**, 563–570.

Monteiro,B. *et al.* (2021) Evolutionary dynamics of the human pseudoautosomal regions. *PLoS Genet.*, **17**, e1009532.

Nieva,J.L. *et al.* (2012) Viroporins: structure and biological functions. *Nat. Rev. Microbiol.*, **10**, 563–574.

Pál,C. *et al.* (2006) An integrated view of protein evolution. *Nat. Rev. Genet.*, **7**, 337–348.

Pascual-Garcia,A. *et al.* (2010) Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. *Proteins*, **78**, 181–196.

Pascual-Garcia,A. *et al.* (2019) The molecular clock in the evolution of protein structures. *Syst. Biol.*, **68**, 987–1002.

Pathy,L. (2008) *Protein Evolution.* Wiley-Blackwell, Oxford.

Pazos,F. and Valencia,A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.

Perez-Losada,M. *et al.* (2015) Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.*, **30C**, 296–307.

Perez-Losada,M. *et al.* (2011) Phylodynamics of HIV-1 from a phase III AIDS vaccine trial in Bangkok, Thailand. *PLoS One*, **6**, e16902.

Perez-Losada,M. *et al.* (2009) Ethnic differences in the adaptation rate of HIV gp120 from a vaccine trial. *Retrovirology*, **6**, 67.

Posada,D. and Crandall,K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA*, **98**, 13757–13762.

Pybus,O.G. and Rambaut,A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.*, **10**, 540–550.

Shi,B. *et al.* (2010) Evolution and recombination of genes encoding HIV-1 drug resistance and tropism during antiretroviral therapy. *Virology*, **404**, 5–20.

Shriner,D. *et al.* (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.*, **81**, 115–121.

Shriner,D. *et al.* (2004) Pervasive genomic recombination of HIV-1 in vivo. *Genetics*, **167**, 1573–1583.

Smith,B.J. (2003) *Protein Sequencing Protocols.* Humana Press, Totowa, NJ.

Speranskaya,A.S. *et al.* (2012) Impact of recombination on polymorphism of genes encoding Kunitz-type protease inhibitors in the genus Solanum. *Biochimie*, **94**, 1687–1696.

Spielman,S.J. (2020) Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. *Mol. Biol. Evol.*, **37**, 2110–2123.

Strasser,B.J. (2010) Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of protein sequence and structure, 1954–1965. *J. Hist. Biol.*, **43**, 623–660.

Stumpf,M.P. and McVean,G.A. (2003) Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.*, **4**, 959–968.

Sun,D. *et al.* (2019) Biochemical characterization of recombinant Avihepatovirus 3C protease and its localization. *Virol. J.*, **16**, 54.

Taverna,D.M. and Goldstein,R.A. (2002) Why are proteins marginally stable? *Proteins*, **46**, 105–109.

Te Velthuis,A.J.W. *et al.* (2012) The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. *Nucleic Acids Res.*, **40**, 1737–1747.

van Leeuwen,R. *et al.* (2003) A randomized trial to study first-line combination therapy with or without a protease inhibitor in HIV-1-infected patients. *AIDS*, **17**, 987–999.

VanInsberghe,D. *et al.* (2021) Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. Virus Evol., veab059.

Voigt,C.A. *et al.* (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.*, **9**, 553–558.

Wilke,C.O. (2012) Bringing molecules back into molecular evolution. *PLoS Comput. Biol.*, **8**, e1002572.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yang,Z. (2006) *Computational Molecular Evolution*. Oxford University Press, Oxford.

Yang,Z. *et al.* (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.*, **15**, 1600–1611.

Zhu,Z. *et al.* (2020) Genomic recombination events may reveal the evolution of coronavirus and the origin of SARS-CoV-2. *Sci. Rep.*, **10**, 21617.