Tools and Resources article

Forecasting protein evolution by integrating birth-death population models with structurally constrained substitution models

David Ferreiro[1,2], Luis Daniel González-Vázquez[1,2], Ana Prado-Comesaña[1] and Miguel Arenas[1,2,*]

[1]CINBIO, Universidade de Vigo, 36310 Vigo, Spain.

[2]Department of Biochemistry, Genetics and Immunology, Universidade de Vigo, 36310 Vigo, Spain.

Email addresses:

DF: david.ferreiro.garcia@uvigo.es

LDGV: luisdaniel.gonzalez@uvigo.gal

APC: ana.maria.prado.comesana@uvigo.gal

MA: marenas@uvigo.es

*** Corresponding author:**
Miguel Arenas
Department of Biochemistry, Genetics and Immunology
University of Vigo
36310 Vigo, Spain
Email: marenas@uvigo.es
Phone: +34 986 130047

Running head: Forecasting protein evolution

Keywords: Molecular evolution, Forecasting evolution, Birth-death process, Substitution model, Protein folding stability, Phylogenetics, Molecular adaptation

Abstract

Evolutionary studies in population genetics and ecology were mainly focused on predicting and understanding past evolutionary events. Recently, however, a growing trend explores the prediction of evolutionary trajectories toward the future promoted by its wide variety of applications. In this context, we introduce a forecasting protein evolution method that integrates birth-death population models with substitution models that consider selection on protein folding stability. In contrast to traditional population genetics methods that usually make the unrealistic assumption of simulating molecular evolution separately from the evolutionary history, the present method combines both processes to simultaneously model forward-in-time birth-death evolutionary trajectories and protein evolution under structurally constrained substitution models that outperformed traditional empirical substitution models. We implemented the method into a freely available computer framework. We evaluated the accuracy of the predictions with several monitored viral proteins of broad interest. Overall, the method showed acceptable errors in predicting the folding stability of the forecasted protein variants but, expectedly, the errors were larger in the prediction of the corresponding sequences. We conclude that forecasting protein evolution is feasible in certain evolutionary scenarios and provide suggestions to enhance its accuracy by improving the underlying models of evolution.

Introduction

Molecular evolution is traditionally investigated through inferences about past evolutionary events, such as phylogenetic tree and ancestral sequence reconstructions, and predictions about the future were considered as inaccessible for a long time because they can be affected by complex processes such as environmental change. Nevertheless, a variety of biological systems display a Darwinian evolutionary process where selection operates toward a limited set of adapted variants. These variants, and in extension the evolutionary trajectories to reach them, would be positively selected and could present a certain degree of predictability (Lässig *et al*, 2017; Wortel *et al*, 2023). The progress made in developing more accurate models of evolution (Arenas, 2015b) and the benefits from predicting the outcome of evolution (i.e., to understand the course of evolution or to prepare for the future (Wortel *et al.*, 2023)) motivated a variety of investigations on forecasting evolution in diverse fields including medicine, agriculture, biotechnology, and conservation biology, among others (e.g., Barton *et al*, 2016; Bull & Molineux, 2008; de Visser *et al*, 2018; Diaz-Uriarte & Vasallo, 2019; Fischer *et al*, 2015; Gerrish & Sniegowski, 2012; Lassig & Luksza, 2014; Lind *et al*, 2019; Luksza & Lassig, 2014; Morris *et al*, 2018; Munck *et al*, 2014; Neher *et al*, 2014; Wortel *et al.*, 2023). Unfortunately, forecasting evolution not always is achievable. Under neutral evolution, all the molecular variants are equally likely to be present in the population, showing lack of repeatability and disallowing accurate prediction of future variants. Thus, forecasting evolution requires a system with measurable selection pressures, and where certain positively selected variants could produce more descendants than other variants and expand in the population (Desai & Fisher, 2007; Goyal *et al*, 2012; Neher & Hallatschek, 2013; Neher *et al.*, 2014). Actually, a rougher fitness landscape resulting from selection can lead to greater accuracy in evolutionary predictions (Papkou *et al*, 2023; Rubin *et al*, 2023; Van Cleve &

Weissman, 2015). Overall, an evolutionary process could be predictable to some extent (prediction errors are inevitable with any method and in any evolutionary scenario) depending on the strength of selection driving evolution and the heterogeneity in fitness among different variants.

Here, we focus on forecasting protein evolution because it involves molecular evolutionary processes driven by selection pressures and where the fitness of each variant can be parameterized and predicted (Carneiro & Hartl, 2010; Gilson *et al*, 2017). Traditionally, evolutionary histories and ancestral sequences of proteins are inferred using probabilistic methods based on advanced substitution models of protein evolution (e.g., Arenas, 2015b; Arenas & Bastolla, 2020; Ferreiro *et al*, 2022; Malcom *et al*, 1990; Moreira *et al*, 2023; Stackhouse *et al*, 1990; Thornton *et al*, 2003; Ugalde *et al*, 2004). The accuracy of these inferences is affected by the accuracy of the applied substitution model, where substitution models that better fit with the study data usually produce more accurate phylogenetic trees and ancestral sequences (Arenas & Bastolla, 2020; Del Amparo & Arenas, 2022a, 2023; Lemmon & Moriarty, 2004). These findings suggest that accurate substitution models of evolution are also convenient for forecasting protein evolution. In this regard, a variety of studies showed that structurally constrained substitution (SCS) models of protein evolution provide more accurate evolutionary inferences than the traditional empirical substitution models of protein evolution, in terms of phylogenetic likelihood, distribution of amino acid frequencies among protein sites, rates of molecular evolution and folding stability of reconstructed proteins, among other aspects (e.g., Arenas & Bastolla, 2020; Arenas *et al*, 2015; Bastolla *et al*, 2006; Bordner & Mittelmann, 2013; Del Amparo *et al*, 2023; Echave & Wilke, 2017; Ferreiro *et al*, 2024a; Ferreiro *et al*, 2024b; Fornasari *et al*, 2002; Parisi & Echave, 2001; Pascual-Garcia *et al*, 2019; Rodrigue *et al*, 2005), although SCS models

usually demand more computational resources than substitution models that only include information from the protein sequence. Notice that the protein structure provides information about the location and molecular interactions of amino acids at different protein sites, which could be far from each other in the sequence but close in the three-dimensional structure and interact affecting their evolution (Morcos *et al*, 2011; Ruiz-Gonzalez & Fares, 2013). Indeed, selection from the protein folding stability is relevant in the evolution of multiple proteins, including those in microbial and viral systems (e.g., Ferreiro *et al.*, 2022; Gong *et al*, 2013; Jacquier *et al*, 2013; Rodrigues *et al*, 2016; Wylie & Shakhnovich, 2011; Zeldovich *et al*, 2007). Therefore, we believe that it should be taken into account for forecasting protein evolution in such systems.

Predictions about future evolutionary events can be performed with simulation-based methods (e.g., Eccleston *et al*, 2023; Neher *et al*., 2014; Yoshida *et al*, 2023). In order to simulate molecular evolution, traditional population genetics methods apply two separate steps (Arenas, 2012; Hoban *et al*, 2012). First, the simulation of the evolutionary history (i.e., a phylogenetic tree) using approaches such as the coalescent and birth-death population processes (Gernhard, 2008; Hudson, 1990; Kingman, 1982; Stadler, 2010). Afterward, the forward-in-time simulation of molecular evolution is performed, from the root node to the tip nodes, upon the previously simulated evolutionary history (Yang, 2006). This methodology was implemented into a variety of population genetics frameworks that simulate molecular evolution (Arenas, 2012; Hoban *et al*., 2012). However, for technical and computational simplicity, it assumes that the simulation of the evolutionary history is independent from the simulation of molecular evolution, which can produce biological incoherences (i.e., the evolutionary history is usually simulated under neutral evolution while molecular evolution is usually simulated with substitution models that consider selection). To enhance the realism of

this modeling, here we merged both processes into a single one where evolutionary history influences molecular evolution and *vice versa*. In particular, we adopted a birth-death population genetics method to simulate the forward-in-time evolutionary history (already used for forecasting evolution (Lassig & Luksza, 2014; Neher *et al.*, 2014)), taking into account the fitness of the molecular variant (through evolutionary constraints from the protein folding stability (Bastolla & Demetrius, 2005; Gong *et al.*, 2013; Liberles *et al*, 2012; Zeldovich *et al.*, 2007)) at the corresponding node, to determine its subsequent birth or death event, and we integrated this process with SCS models to model protein evolution along the derived phylogenetic branches. The method is detailed below, and we implemented it into a new version of our computer framework *ProteinEvolver* (Arenas *et al*, 2013), which is freely available from https://github.com/MiguelArenas/proteinevolver. *ProteinEvolver2* includes detailed documentation and a variety of ready-to-use examples. Next, considering the potential applications of forecasting evolution to design vaccines and therapies against pathogens, we evaluated and applied the method to forecasting protein evolution in several real protein data of viruses monitored over time.

Materials and Methods

A method for forecasting protein evolution by combining birth-death population genetics with structurally constrained substitution models of protein evolution

Following previous methods for forecasting evolution based on simulations (Lassig & Luksza, 2014; Neher *et al.*, 2014), we developed a method to simulate the forward-in-time evolutionary history of a protein sample with a birth-death process that considers the fitness of the protein variant (based on folding stability) at every temporal node. The method derives the birth and death rates for a protein variant based on the fitness of that variant, where a high

fitness results in a high birth rate and a low death rate, which can lead to a large number of descendants, and the opposite leading to a few or none (extinction) descendants. Thus, the fitness of the molecular variant at every node drives its corresponding forward in time birth-death evolutionary history. The details of this simulation process are outlined below.

First, similarly to common simulators of molecular evolution (Arenas, 2012; Hoban *et al.*, 2012; Yang, 2006), a given protein sequence and structure (hereafter, protein variant) is assigned to the root node. The fitness (f) of the protein variant (A) is calculated from its folding stability (free energy, ΔG) following the Boltzmann distribution (Goldstein, 2013) (I, which takes values from 0 to 1),

$$f(A) = \frac{1}{1 + e^{\Delta G / kT}} \qquad \text{(I)}$$

Protein folding stability constrains protein evolution and is commonly used to obtain protein fitness (Bastolla *et al*, 2007; Goldstein, 2013; Liberles *et al.*, 2012; Lobkovsky *et al*, 2010; Mendez *et al*, 2010; Sella & Hirsh, 2005; Zeldovich *et al.*, 2007). The user can alternatively choose whether the fitness of the modeled protein variant is determined solely by its folding stability or by its similarity to the stability of a real protein variant (i.e., a protein structure from the Protein Data Bank, PDB). We believe the latter can be more realistic, as in nature, high folding stability does not necessarily indicate high fitness, but a stability that closely resembles that of a real protein may suggest high fitness since the real stability is the result of a selection process (which also incorporates negative design). If the fitness is derived from only the folding stability of the protein variant, the birth rate (b) is considered equal to the fitness. Alternatively, if the fitness is determined based on the similarity in folding stability between the modeled variant and a real variant, the birth rate is assumed to be 1 minus the root mean square deviation (RMSD, which offers advantages such as minimizing the influence of small deviations while amplifying larger differences, thereby enhancing the

detection of remarkable molecular changes) in folding stability. Notice that the smaller this difference, the higher the birth rate. In both cases, the death rate (*d*) is considered as *1-b* to allow a constant global (birth-death) rate. In this model, the fitness influences reproductive success, where protein variants with higher fitness have higher birth rates leading to more birth events, while those with lower fitness have higher death rates leading to more extinction events. This parameterization is meaningful in the context of protein evolution because the fitness of a protein variant can affect its survival (birth or extinction) without necessarily altering its rate of evolution. Although a higher growth rate can sometimes correlate with higher fitness, a variant with high fitness does not necessarily accumulate substitutions more rapidly.

Additionally, we incorporated another birth-death model that follows the proposal by Neher et al (2014), in which the death rate is fixed at 1 and the birth rate is modeled as *1+fitness*. In this model, fitness not only affects reproductive success but also influences the global birth-death rate, which can vary among lineages.

The birth-death process is simulated forward in time deciding whether every next event is a birth or a death event (Harmon, 2019; Stadler, 2010; Stadler, 2011) according to the fitness of the corresponding molecular variant, and it ends when a user-specified criterion, such as a particular sample size (considering or ignoring extinction nodes) or a certain evolutionary time (t_e), is reached. Starting from an "active" node (i.e., the root node) at current time t_c, the time to the next event (birth to produce two descendants or death to produce the extinction of the node) can be calculated (details below). In contrast to standard birth-death processes where birth and death rates are constant over time and among lineages, the present method considers heterogeneity where each protein variant at a node has specific birth and death rates

according to its corresponding fitness. The method is described below and summarized in Figure 1.

1. The process starts at the root node, assigning a user-specified protein sequence and corresponding protein structure (i.e., obtained from the PDB) to that node. In general, for every protein variant assigned to an active node, the corresponding birth and death rates are calculated following the indications presented above.
2. Calculation of the time to the next birth or death event. Following common birth-death methods, the time to the next event t_n is calculated through an exponential distribution with rate based on the number of active nodes (s) and the sum of the birth and death rates (Harmon, 2019) (II),

$$t_n = e^{(s(b+d))} \qquad \text{(II)}$$

   One of the birth-death models that we implemented considers that $b+d=1$ at each node, allowing variation of the reproductive success among nodes while keeping t_n consistent among them, according to (Harmon, 2019). In contrast, in the other birth-death model we implemented, $b+d$ can vary among nodes (Neher *et al.*, 2014), thereby allowing variation of both reproductive success and t_n among nodes.
3. Evaluate whether the simulation concludes before the next event occurs $(t_c + t_n)$.
4. If it does conclude (i.e., $t_c + t_n$ higher than t_e), the simulation of the evolutionary history finishes.
5. If it does not conclude, a random active node is selected, and its protein variant is analyzed to determine its type of next event (birth or death). The probability of a birth event is $P_b = b/(b + d)$ and the probability of a death event is $P_d = d/(b + d)$. A random sample from those probabilities is taken to determine the type of evolutionary event.

6. If a birth event is selected. Two descendant active nodes from the study node are incorporated with branch lengths t_n, and the study node is then considered inactive. Next, molecular evolution is simulated from the original node to each descendant node based on the specified SCS model of protein evolution (Arenas *et al.*, 2013). The integration of SCS models to evolve protein variants along given branch lengths followed standards approaches of molecular evolution in population genetics, in which the branch length and the substitution model inform the number and type of substitution events, respectively (Arenas, 2012; Carvajal-Rodriguez, 2010; Hoban *et al.*, 2012; Yang, 2006). Thus, the process results in a protein variant for every descendant node. Finally, the folding stability and subsequent fitness of these descendant protein variants are calculated.
7. If a death event is selected. Then, the study node is considered inactive.
8. Return to step 1 while the user-specified criterion for ending the birth-death process is not satisfied and at least one active node exists in the evolutionary history. Otherwise, the simulation ends.

This process simultaneously simulates, forward in time, evolutionary history and protein evolution, with protein evolution influencing the evolutionary history through selection from the folding stability. Indeed, selection can vary among protein variants at their corresponding nodes of the evolutionary history. The process produces a forward in time birth-death phylogenetic history, that encompasses nodes that reached the ending time, internal nodes and nodes that were extinct at some time, along with the protein variant associated with each node.

The method includes several optional capabilities listed in Table S1 (Supplementary Material) and in the software documentation, with some summarized below.

- The user can fix the birth and death rates along the evolutionary history or specify that they are based on the fitness of the corresponding protein variant as described before. For the latter, the fitness can be based on the folding stability of the analyzed protein variant or on the similarity in folding stability between the analyzed and real protein variants. In addition, the global birth-death rate can be constant or vary among lineages depending on the specified birth-death model.
- The birth-death simulation can finish when any of the following criteria is met: (a) A specified sample size, including extinction nodes derived from death events. (b) A specified sample size, excluding extinction nodes. (c) A specified evolutionary time, measured from the root to a tip node.
- Extinction nodes can either be preserved or removed from the evolutionary history.
- Evaluation of the simulated birth-death evolutionary history in terms of tree balance using the Colless index (Colless, 1982; Lemant *et al*, 2022).
- Possible variation of the site-specific substitution rate according to user specifications.
- Customizable substitution model featuring site-specific exchangeability matrices (relative rates of change among amino acids and their frequencies at the equilibrium).
- Implementation of several SCS models and a variety of empirical substitution models of protein evolution. The implemented SCS models were evaluated in our previous work (Arenas *et al*., 2013) and, the implemented empirical substitution models were properly identified using simulated data with *ProtTest3* (Darriba *et al*, 2011).

The implemented SCS models consider molecular energy functions based on amino acid contact matrices and configurational entropies per residue in unfolded and misfolded proteins

(Arenas *et al.*, 2013; Bastolla *et al.*, 2007; Mendez *et al.*, 2010). These models incorporate both positive and negative design strategies. In particular, the evaluation of the target protein structure while taking into account a database of residue contacts from alternative protein structures in the PDB, thus considering background genetic information that helps reduce prediction biases (Minning *et al*, 2013). Technical details about these SCS models are presented in our previous study (Arenas *et al.*, 2013). Next, SCS models outperformed models that ignore structural evolutionary constraints in terms of phylogenetic likelihood, among other properties (Arenas *et al.*, 2013; Arenas *et al.*, 2015; Bordner & Mittelmann, 2013). The method also implements common empirical substitution models of protein evolution (i.e., *Blosum62*, *CpRev*, *Dayhoff*, *DayhoffDCMUT*, *FLU*, *HIVb*, *HIVw*, *JTT*, *JonesDCMUT*, *LG*, *Mtart*, *Mtmam*, *Mtrev24*, *RtRev*, *VT* and *WAG*; Table S1) and the user can specify any particular exchangeability matrix for all sites or for each site, allowing for heterogeneity of the substitution process among sites (details in Table S1 and in the software documentation). In addition, the framework implements heterogeneous substitution rates among sites by the traditional Gamma distribution (+G) (Yang, 1996) and proportion of invariable sites (+I) (Fitch & Margoliash, 1967), and also the user can directly alter the substitution rate at each site for any empirical or SCS model (Table S1). Regarding the evolutionary history, in addition to the birth-death process presented before, the user can specify a particular phylogenetic tree or simulate a coalescent evolutionary history (Hudson, 1983; Kingman, 1982) (Table S1). In this regard, we maintained the capabilities of the previous version, including the coalescent with recombination (Hudson, 1983) [which can be homogeneous or heterogeneous along the sequence (according to Wiuf & Posada, 2003)], variable population size over time (growth rate or demographic periods), several migration models [i.e., island (Hudson, 1998), stepping-stone (Kimura & Weiss, 1964) and continent-island (Wright, 1931)]

with temporal variation of migration rates and convergence of demes or subpopulations, simulation of haploid or diploid data, and longitudinal sampling (Navascues *et al*, 2010) (details in Table S1 and in the software documentation). The framework outputs a simulated multiple sequence alignment with the protein sequences of the internal and tip nodes, as well as their folding stabilities, and the evolutionary history, among other information (Table S1 and software documentation).

Study data for evaluating the method for forecasting protein evolution

We evaluated the accuracy of the method for forecasting protein evolution using viral proteins sampled over time (longitudinal sampling). Specifically, we used protein sequences from previous experiments of virus evolution monitored over time, which contain consensus molecular data (avoiding rare variants) that belong to different evolutionary time points.

- The matrix (MA) protein of HIV-1, with data obtained from an *in vitro* cell culture experiment where samples were collected at different times (Arenas *et al*, 2016). This data included 21 MA protein consensus sequences collected at times (population passages, *T*) *T1* (initial time and includes an initial sequence) and *T31* (time after 31 passages and includes 20 sequences), which accumulated 48 amino acid substitutions (sequence identity 0.973; Figure S1, Supplementary Material).
- The main protease (Mpro) and papain-like protease (PLpro) proteins of SARS-CoV-2. For each protein, the data includes the first sequenced variant (Wuhan) and a sequence built with all the substitutions observed in a dataset of 384 genomes of the Omicron variant of concern collected from the GISAID database. The resulting Mpro and PLpro sequences presented 10 and 22 substitutions (sequence identity 0.967 and 0.930), respectively.

- The non-structural protein 1 (NS1) of the influenza virus. We retrieved the NS1 sequences from the years 2005, 2015 and 2020 from the Influenza Virus Resource (Bao *et al*, 2008). Next, we obtained the consensus sequence of the 2005 dataset as initial time point (*T1*), and the consensus sequences from the 2015 (*T2*) and 2020 (*T3*) datasets as subsequent time points. The resulting consensus sequences for 2015 and 2020 showed 40 and 37 substitutions (sequence identity 0.802 and 0.817), respectively.
- The protease (PR) of HIV-1 with data sampled from patients monitored over time, from 2008 to 2017, available from the Specialized Assistance Services in Sexually Transmissible Diseases and HIV/AIDS in Brazil (Ferreiro *et al.*, 2022; Santos-Pereira *et al*, 2021; Souto *et al*, 2021). These evolutionary scenarios are complex due to the diverse antiretroviral therapies administrated to the patients (Table S2; Supplementary Material), which could vary during the studied time periods and that could promote the fixation of specific mutations (i.e., associated with resistance) (Ferreiro *et al.*, 2022; Santos-Pereira *et al.*, 2021; Souto *et al.*, 2021). For each viral population (patient), the data included a consensus sequence for each of the four or five samples collected at different time points. These consensus sequences exhibited between one and 22 amino acid substitutions respect to the consensus sequence of the corresponding first sample (Table S2).

We identified a representative protein structure for each dataset, which was used to predict the folding stability and to inform the SCS model. In particular, we obtained the protein structures of the sequences at the initial time (*T1*) from the PDB (Table S3, Supplementary Material). For the case of the HIV-1 protease, we obtained the protein structure through homology modelling. We used *SWISS-MODEL* (Arnold *et al*, 2006) to identify the best-fitting templates

of PDB structures (Table S3). Next, we predicted the protein structures by homology modelling with *Modeller* (Sali & Blundell, 1993) using the protein sequences and corresponding best-fitting structural templates.

Forward in time prediction of viral protein variants

The evaluation was performed with the previously presented real data, which includes a real protein variant present at the initial time (*T1*) and subsequent protein variants present at a later time (*Tn*). We applied the method for forecasting protein evolution to predict the most likely protein variants at *Tn* derived from the real protein variant observed at the initial time (*T1*). The prediction error was determined by measuring the distance between the real protein variants and the predicted variants corresponding to time *Tn*.

Thus, we assigned to the initial node (*T1*) the corresponding real protein variant (including its sequence and structure), and its evolution was simulated forward in time until it reached the number of substitutions observed in the real data at time *Tn*. Thus, we considered the number of observed substitutions as a measure of evolutionary time to allow proper comparisons between real and predicted variants. We simulated 100 alignments of protein sequences, each containing the same number of sequences as the real data at *Tn*. This included 20 HIV-1 MA protein sequences, a consensus sequence for the SARS-CoV-2 Mpro, a consensus sequence for the SARS-CoV-2 PLpro, a consensus sequence of the influenza NS1 protein for each time point and, a consensus sequence of HIV-1 PR for each viral population. To avoid rare variants, each sequence of the simulated multiple sequence alignments was obtained as the consensus of 100 linked simulated sequences.

To investigate the effect of selection on the predictions, we compared the accuracy of forecasting protein evolution when selection from the protein structure is considered and,

when it is ignored (neutral evolution). If selection is considered, as previously presented, the probability of birth and death events was based on the fitness of the protein variants, and protein evolution was modeled using an SCS model (Arenas *et al.*, 2013). In the case of neutral evolution, all protein variants equally fit and are allowed. Since variants are observed, we allowed birth events. However, it assumed the absence of death events as no information independent of fitness is available to support their inclusion, thereby avoiding the imposition of arbitrary death events based on an arbitrary death rate. Also, to model neutral evolution, we used an exchangeability matrix with the same relative rates of change to all amino acid pairs.

Accuracy of the predicted protein variants

We assessed the accuracy of the method for forecasting protein evolution by comparing the predicted and real protein variants present at time *Tn*, both at the protein sequence and structure levels.

For data containing multiple sequences at time *Tn* (i.e., HIV-1 MA dataset), we calculated the Kullback-Leibler (KL) divergence, which provides a distance between two multiple sequence alignments (the real and predicted data) based on the distribution of amino acid frequencies along the sequences (Equation III, where factors *P* and *Q* denote the distribution of amino acid frequencies in the real and predicted protein sequences at time *Tn*, respectively, *i* refers to protein site) (Kullback & Leibler, 1951). This distance was only calculated for data with a set of sequences at time *Tn* (HIV-1 MA) because a single sequence does not provide site-specific variability.

$$KL\ (P||Q) = \sum_i P_i \times \log(\frac{P_i}{Q_i}) \qquad \text{(III)}$$

We also compared the real and predicted evolutionary trajectories of protein variants using the Grantham distance, which measures the differences between amino acids based on their

physicochemical properties (Grantham, 1974). In particular, for both real and predicted protein variants, we calculated the Grantham distance at each protein site that differs between the two datasets, considering its evolution from *T1* to the subsequent multiple sequence alignment at *Tn*. We examined sites that varied over time, thus the general site-specific Grantham distance *Gi* was calculated as the frequency of each amino acid *f* at site *i* multiplied by the specific Grantham distance between amino acid *m* at time *Tn* and amino acid *n* at time *T1*, normalized with the largest Grantham distance G_{max} to obtain values between 0 and 1 (Equation IV). Next, to compare the real and predicted data, we calculated the site-specific difference of Grantham distance Gb_i between the real *P* and predicted *Q* protein variants (Equation V),

$$G_i = \sum_{m=1}^{20} f_m \times \frac{G_{(m,n)}}{G_{max}} \qquad \text{(IV)}$$

$$Gb_i = |G_{P,i} - G_{Q,i}| \qquad \text{(V)}$$

In addition, we obtained and compared the protein folding stability (ΔG) of the predicted and real protein variants observed at time *Tn*, using their corresponding protein structures, with *DeltaGREM* (Arenas *et al.*, 2015; Minning *et al.*, 2013).

Results

Implementation of the forecasting protein evolution method

We extended the previous version of our framework *ProteinEvolver* (Arenas *et al.*, 2013), maintaining its previous capabilities (i.e., simulation of protein evolution upon user-specified phylogenetic trees and upon phylogenetic trees simulated with the coalescent with or without recombination, migration, demographics and longitudinal sampling, empirical and SCS models, among others; Table S1), by adding, among others (Table S1), the forward in time modeling of protein evolution that combines a birth-death process based on the fitness of

every protein variant (folding stability) at each node to determine its birth and death rates, as well as SCS models of protein evolution. The framework *ProteinEvolver2* is written in C and distributed with a detailed documentation and a variety of illustrative practical examples. The framework is freely available from https://github.com/MiguelArenas/proteinevolver.

Evaluation of predictions of HIV-1 MA evolution

Regarding the evolution of the HIV-1 MA protein, the Grantham distance and the KL divergence between the real variants at time *Tn* and the corresponding predicted variants were low (around 5% and 6%, respectively; Table 1), and they did not differ comparing predictions that consider selection on the folding stability (including birth-death models with constant and variable global birth-death rate among lineages) and predictions that ignore it (Table 1). On the other hand, we found that the folding stability of the protein variants predicted considering selection on the folding stability (again, including birth-death models with constant or variable global birth-death rate among lineages) was closer to the folding stability of the real protein variants than that of the protein variants predicted under neutral evolution (Table 1). In particular, the protein variants predicted ignoring selection were less stable than those predicted considering selection and also less stable than the real protein variants.

Evaluation of predictions of SARS-CoV-2 Mpro and PLpro evolution

The analyses of the SARS-CoV-2 Mpro and PLpro data showed Grantham distances between the real and predicted sequences around 25% and 36%, respectively (Figure 2A). Again, this distance was similar when comparing predictions based on models that consider selection on the protein folding stability (including birth-death models with constant or variable global birth-death rate among lineages) and a model of neutral evolution. Regarding comparisons

based on the protein folding stability, we found again that the models that consider selection from the folding stability produce variants closer to the stability of the real protein variants than the model that ignores selection (Figure 2B). Indeed, protein variants derived from the models that consider selection were more stable than those derived from the model of neutral evolution. Next, we did not find statistically significant differences in sequence similarity or folding stability between variants predicted under birth-death models with constant or variable global birth-death rate among lineages (Figure 2).

Evaluation of predictions of influenza NS1 protein evolution

The evolutionary predictions for the influenza NS1 protein varied depending on the model used. Specifically, at the sequence level and for the two prediction time points studied, Grantham distances between the real and predicted protein sequences were around 23.5% for the models that incorporated structural evolutionary constraints, compared to about 25.5% for the neutral model (Figure 3A). These differences became more pronounced when examining predictions based on protein folding stability. For both time points, models that included selection consistently generated protein variants with stability more similar to that of the observed variants than those predicted by the neutral model (Figure 3B). Indeed, sequences predicted by the model that accounts for selection were generally more stable than those predicted under neutral evolution. Again, we found no statistically significant differences in sequence similarity or folding stability between variants predicted under birth-death models with constant or variable global birth-death rate among lineages (Figure 3).

Evaluation of predictions of HIV-1 PR evolution

In general, the Grantham distance, which compared the evolutionary trajectories of the real and predicted protein variants from time *T1* to later times, varied among viral populations (patients) (Figure 4A). However, for the majority of these populations, the distance remained below 30% and exhibited minor fluctuations over time. One particular population exhibited a notable trend, with the distance increasing from 10% to nearly 60% over time. Considering that the length of the evolutionary trajectories of the protein can differ among the studied populations, we explored whether the accumulated number of amino acid substitutions could affect the accuracy of the predictions, and we found that the number of substitutions varied among populations and this variability did not correlate with the Grantham distance between the real and predicted data (R^2=0.0001, Figure 4B). In general, the folding stability of the predicted protein variants was similar or slightly less stable than that of the real protein variants [with a difference ranging from 0 to 9 kcal/mol and a mean of 3.1 ± 0.9 (95% CI) kcal/mol; Figure 4C]. Indeed, the folding stability exhibited small fluctuations, increasing and decreasing, over time.

Discussion

While reconstructing evolutionary histories and ancestral sequences, among other inferences about the past, was popular in the field, predictions about evolution toward the future were traditionally ignored due to potential high prediction errors. However, in the last decade, forecasting evolution has gained attention because of its variety of applications and advancements in models of evolution (see the reviews Lässig *et al.*, 2017; Wortel *et al.*, 2023). Several studies showed that forecasting evolution can be feasible in some systems, including virus evolution (Luksza & Lassig, 2014; Thadani *et al*, 2023). Here, we also investigated forecasting evolution in viruses but at the molecular level, considering the

relevance of the substitution process to produce molecular variants that affect the viral fitness (Arenas, 2015a; Arenas *et al*., 2016; Bloom & Neher, 2023; Poon *et al*, 2007; Watabe & Kishino, 2010). We believe that forecasting genome evolution remains challenging due to the complexity of its evolutionary processes [i.e., including epistatic interactions, chromosomal rearrangements and heterogeneous substitution patterns among genomic regions (Arbiza *et al*, 2011), among others] that complicate the parameterization and prediction of accurate fitness landscapes. However, we believe that it could be more feasible in structural proteins because of their relatively simpler evolutionary processes that usually include selection on folding stability (Bloom *et al*, 2006; Goldstein, 2011).

To make evolutionary predictions over time, either toward the past or toward the future, considering a substitution model of molecular evolution can be convenient. Actually, a variety of traditional studies showed that the substitution model influences the phylogenetic likelihood generated by probabilistic approaches used for evolutionary inference (Lemmon & Moriarty, 2004; Yang *et al*, 1994; Zhang & Nei, 1997). To study protein evolution, empirical substitution models of protein evolution are routinely used because they allow rapid calculations based on the assumption of site-independent evolution and typically apply the same exchangeability matrix for all the protein sites, which is highly unrealistic (Echave *et al*, 2016). Besides, a small set of empirical substitution models was developed for modeling the evolution of the diverse range of viral proteins (Dang *et al*, 2010; Del Amparo & Arenas, 2022b; Dimmic *et al*, 2002; Le & Vinh, 2020; Nickle *et al*, 2007). As an alternative, some substitution models that consider evolutionary constraints on the protein structure incorporated site-dependent evolution, which allows a more accurate modeling of protein evolution compared to the empirical substitution models (Arenas *et al*., 2013; Arenas *et al.*, 2015; Bordner & Mittelmann, 2013; Ferreiro *et al*., 2024a; Parisi & Echave, 2005).

According to previous methods for forecasting evolution based on computer simulations (Lassig & Luksza, 2014; Neher *et al.*, 2014), we also adopted the birth-death population genetics process to allow a forward-in-time evolutionary process where the birth and death rates can differ among nodes (Stadler, 2010; Stadler, 2011). Traditional population genetics methods to simulate the evolution of molecular data implement a two-step simulation process, where, first, the evolutionary history is simulated [i.e., with the coalescent theory or a forward-in-time simulation approach (Arenas, 2012; Hoban *et al.*, 2012), often assuming neutral evolution] and, in a subsequent step, molecular evolution is simulated upon the previously obtained evolutionary history (Yang, 2006). However, this methodology is unrealistic because the fitness of the data can affect its evolutionary history (i.e., a variant with high fitness is likely to have more descendants than a variant with low fitness). Thus, we designed and implemented a method for forecasting protein evolution that integrates a birth-death population genetics process (including the modeling of constant and variable global birth-death rate among lineages) with a SCS model of protein evolution, where the folding stability of each protein variant is evaluated to predict its future trajectory in terms of both evolutionary history and molecular evolution. We implemented this forward-in-time simulation of protein evolution in a new version of our previous framework *ProteinEvolver* (Arenas *et al.*, 2013), while maintaining its previous capabilities and extending some of them (i.e., incorporation of site-specific exchangeability matrices and additional substitution models of protein evolution, among others; see Table S1 and software documentation). As any other method for forecasting evolution, the present method ignores possible environmental shifts that are inherently unpredictable and that could affect the accuracy of the predictions. Next, we evaluated the forward-in-time evolutionary predictions with real data of HIV-1, SARS-CoV-2 and influenza virus proteins. We determined the prediction errors

between the real and the predicted protein variants by examining dissimilarity in evolutionary trajectory (Grantham distance based on physicochemical properties among the differing amino acids during the evolutionary trajectories), sequence divergence (distribution of amino acid frequencies among sites using the KL divergence), and protein structure (protein folding stability). Additionally, we analyzed the influence of accounting for selection, based on the protein folding stability, on the predictions. We also evaluated birth-death models incorporating either constant or variable global birth-death rates among lineages. Notably, in the variable-rate model, fitness can influence both reproductive success and the rate of molecular evolution.

In general, the sequence and evolutionary trajectory dissimilarities between the real and predicted protein variants were relatively small, with some variations among the study proteins. For the HIV-1 MA protein, the SARS-CoV-2 Mpro and PLpro, and influenza NS1 protein, the sequence dissimilarity was below 10%, 26%, 36% and 26%, respectively. The low prediction errors for the HIV-1 MA protein were expected because this dataset was derived from an *in vitro* cell culture experiment that is not influenced by a variety of external factors that could affect the predictions in *in vivo* experiments. As a rough reference, in traditional ancestral sequence reconstruction (ASR) of protein data with high sequence identity the error was approximately 2% (Arenas & Posada, 2010). Notice that ASR methods can exhibit low prediction errors due to their statistical evaluation based on a set of original sequences, rather than relying on a single original sequence used in the case of forecasting evolution. The contrasting scenario regarding evolutionary complexity was the HIV-1 PR data, which was sampled from populations (patients) undergoing various antiretroviral treatments (Ferreiro *et al.*, 2022). There, the sequence dissimilarity varied among viral populations, although most were below 30%. A viral population exhibited higher

dissimilarities (near 60%) that we believe could be caused by molecular adaptations promoted by the therapies, although this needs formal investigation.

An unexpected result was that the model of neutral evolution produced sequence dissimilarities between the real and predicted protein variants that were quite similar to those obtained with the SCS model (Table 1 and Figures 2 and 3). A few studies indicated that the substitution model has negligible effects on the reconstructed phylogenetic trees (Abadi *et al*, 2019; Spielman, 2020). Subsequent studies found that the influence of the substitution model on phylogenetic reconstructions is dependent on the diversity of the data, where data with high diversity is more sensible to the applied substitution model due to containing more evolutionary information to be modeled (Del Amparo & Arenas, 2022a, 2023). Considering that the studied data present overall low diversity [i.e., sequence identity of 0.973, 0.967, 0.930, 0.802 and 0.817 for the HIV MA data, SARS-CoV-2 Mpro data, SARS-CoV-2 PLpro data, and influenza NS1 protein data (for each prediction time point, T2 and T3), respectively; it is important to note that, in general, longitudinal data derived from monitored evolutionary processes usually show a low diversity because involve relatively short evolutionary histories, among other factors], we believe that the influence of the applied substitution models on the prediction of the sequences was small because of the small number of modeled substitution events, as found in phylogenetic reconstructions (Del Amparo & Arenas, 2022a, 2023). Actually, in the case of the influenza NS1 protein dataset, which had the highest sequence diversity among the study datasets, the sequences predicted under the SCS models were more similar to the real sequences than those derived from predictions under neutral evolution.

Overall, the prediction accuracy varied among the studied evolutionary scenarios; as expected, it was lower in the more complex scenarios. Indeed, datasets with higher sequence diversity contain more evolutionary signals, which can improve prediction quality.

We also evaluated the prediction error between the real and predicted protein variants regarding their folding stability, again comparing the predictions made under a model that considers structural constraints and a model of neutral evolution. In general, the protein variants predicted under the SCS model presented a folding stability close to the folding stability of the respective real protein variants, with differences below 1, 2, 7 and 1 kcal/mol for the HIV MA data, SARS-CoV-2 Mpro data, SARS-CoV-2 PLpro data, and influenza NS1 protein data, respectively. The higher differences (around 9 kcal/mol) were again observed for the HIV-1 PR data. In contrast to the prediction error based on sequence dissimilarity, the prediction error based on folding stability varies between predictions obtained under the SCS model and those obtained under the neutral model. In the studied evolutionary scenarios, the protein variants predicted under the neutral model were less stable and farther from the stability of the real protein variants compared to those predicted under the SCS model. These results were expected because, under SCS models, protein stability can be modeled with greater accuracy than sequence similarity due to selection for maintaining stability in the protein structure despite amino acid changes (Arenas & Bastolla, 2020; Illergard *et al*, 2009; Pascual-Garcia *et al*, 2010). Indeed, previous studies showed that models that ignore structural constraints often produce proteins with unrealistic folding instability (Arenas & Bastolla, 2020; Del Amparo *et al*., 2023), which suggests that accounting for protein folding stability in the modeling of protein evolution is recommended for predicting protein variants with appropriate structural properties.

Therefore, we found a good accuracy in predicting the real folding stability of forecasted protein variants, while predicting the exact sequences was more challenging, which was not surprising considering previous studies. In particular, inferring specific sequences is considerably challenging even for ancestral molecular reconstruction (Arenas & Bastolla,

2020; Arenas *et al*, 2017). Indeed, observed sequence diversity is much greater than observed structural diversity (Illergard *et al.*, 2009; Pascual-Garcia *et al.*, 2010), and substitutions between amino acids with similar physicochemical properties can yield modeled protein variants with more accurate folding stability, even when the exact amino acid sequences differ. Further work is demanded in the field of substitution models of molecular evolution. We also found that datasets with relatively high sequence diversity can improve the accuracy of the predictions due to containing more evolutionary information. Apart from that, forecasting the folding stability of future real proteins is an important advance in forecasting protein evolution, given the essential role of folding stability in protein function (Bloom *et al.*, 2006; Scheiblhofer *et al*, 2017) and its variety of applications.

Variation in the global birth-death rate among lineages showed minor effects on prediction accuracy, suggesting a limited role in protein evolution modeling. Molecular evolution parameters, particularly the substitution model, appear to be more critical in this regard.

In the context of protein evolution, substitution models are a critical factor (Arenas *et al.*, 2017; Bordner & Mittelmann, 2013; Echave *et al.*, 2016; Echave & Wilke, 2017; Liberles *et al.*, 2012; Wilke, 2012), and the presented combination with a birth-death model constitutes a first approximation upon which next studies can build to better understand this evolutionary system. Next, the present method assumes that the protein structure is maintained over the studied evolutionary time, which can be generally reasonable for short timescales where the structure is conserved (Illergard *et al.*, 2009; Pascual-Garcia *et al.*, 2010). Over longer evolutionary timescales, structural changes may occur and, in such cases, modeling the evolution of the protein structure would be necessary. To our knowledge, modeling the evolution of the protein structure remains a challenging task that requires substantial

methodological developments. Recent advances in artificial intelligence, particularly in protein structure prediction from sequence (Abramson *et al*, 2024; Jumper *et al*, 2021), may offer promising tools for addressing this challenge. However, we believe that evaluating such approaches in the context of structural evolution would be difficult, especially given the limited availability of real data with known evolutionary trajectories involving structural change. In any case, this is probably an important direction for future research.

We present a method to simulate forward-in-time protein evolution accounting for evolutionary constraints from the protein structure and a birth-death population process, and where the evolutionary history is influenced by the protein evolution and *vice versa*. The method is implemented in the computer framework *ProteinEvolver2*, which is freely distributed with several practical examples and detailed documentation. We believe that implementing methods into freely available phylogenetic frameworks is important to facilitate practical applications, as well as future improvements and evaluations. We applied the method to forecast protein evolution in some viral proteins. We found that the method provides acceptable approximations to the real evolution, especially in terms of protein folding stability, suggesting that combining structural constraints with birth-death population processes in the modeling of protein evolution is convenient. Still, to advance in methods for forecasting protein evolution, we believe that further efforts should be made in the field to improve the modeling of protein evolution, such as the incorporation of site-dependent evolutionary constraints from the protein activity.

Supplementary Material

The supplementary material is available at the journal online.

Data Availability

The computer framework *ProteinEvolver2* is freely available from https://github.com/MiguelArenas/proteinevolver. The SARS-CoV-2 data is available from GISAID database with DOI https://doi.org/10.55876/gis8.250206gt. The real and predicted protein variants are available from Zenodo repository at the URL https://doi.org/10.5281/zenodo.15548146.

Acknowledgments

This work was supported by the Project PID2023-151032NB-C22 funded by MCIU/AEI/10.13039/501100011033 and by FEDER, UE. D.F. was funded by a fellowship from Xunta de Galicia [ED481A-2020/192]. We thank the "Centro de Supercomputación de Galicia (CESGA)" for the computer resources. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

Abadi S, Azouri D, Pupko T, Mayrose I (2019) Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun* 10: 934

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J *et al* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630: 493-500

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. *PLoS Comput Biol* 8: e1002495

Arenas M (2015a) Genetic Consequences of Antiviral Therapy on HIV-1. *Comput Math Methods Med* 2015: 9

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

Arenas M, Bastolla U (2020) 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, Lorenzo-Redondo R, Lopez-Galindez C (2016) Influence of mutation and recombination on HIV-1 in vitro fitness recovery. *Mol Phylogenet Evol* 94: 264-270

Arenas M, Posada D (2010) The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184: 1133-1139

Arenas M, Sanchez-Cobos A, Bastolla U (2015) Maximum likelihood phylogenetic inference with selection on protein folding stability. *Mol Biol Evol* 32: 2195-2207

Arenas M, Weber CC, Liberles DA, Bastolla U (2017) ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. *Syst Biol* 66: 1054-1064

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

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D (2008) The influenza virus resource at the National Center for Biotechnology Information. *J Virol* 82: 596-601

Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK (2016) Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nature Communications* 7: 11660

Bastolla U, Demetrius L (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng Des Sel* 18: 405-415

Bastolla U, Porto M, Roman HE, Vendruscolo M (2006) A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evol Biol* 6: 43

Bastolla U, Porto M, Roman HE, Vendruscolo M (2007) *Structural approaches to sequence evolution*. Springer, Berlin, Heidelberg

Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* 103: 5869-5874

Bloom JD, Neher RA (2023) Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol* 9: vead055

Bordner AJ, Mittelmann HD (2013) A new formulation of protein evolutionary models that account for structural constraints. *Mol Biol Evol* 31: 736-749

Bull JJ, Molineux IJ (2008) Predicting evolution from genomics: experimental evolution of bacteriophage T7. *Heredity* 100: 453-463

Carneiro M, Hartl DL (2010) Colloquium papers: Adaptive landscapes and protein evolution. *Proc Natl Acad Sci U S A* 107 Suppl 1: 1747-1751

Carvajal-Rodriguez A (2010) Simulation of genes and genomes forward in time. *Curr Genomics* 11: 58-61

Colless DH (1982) Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics. *Systematic Zoology* 31: 100-104

Dang CC, Le QS, Gascuel O, Le VS (2010) FLU, an amino acid substitution model for influenza proteins. *BMC Evol Biol* 10: 99

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

de Visser J, Elena SF, Fragata I, Matuszewski S (2018) The utility of fitness landscapes and big data for predicting evolution. *Heredity (Edinb)* 121: 401-405

Del Amparo R, Arenas M (2022a) Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Mol Biol Evol* 39: msac144

Del Amparo R, Arenas M (2022b) HIV Protease and Integrase Empirical Substitution Models of Evolution: Protein-Specific Models Outperform Generalist Models. *Genes* 13

Del Amparo R, Arenas M (2023) Influence of substitution model selection on protein phylogenetic tree reconstruction. *Gene* 865: 147336

Del Amparo R, González-Vázquez LD, Rodríguez-Moure L, Bastolla U, Arenas M (2023) Consequences of Genetic Recombination on Protein Folding Stability. *Journal of Molecular Evolution* 91: 33-45

Desai MM, Fisher DS (2007) Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics* 176: 1759-1798

Diaz-Uriarte R, Vasallo C (2019) Every which way? On predicting tumor evolution using cancer progression models. *PLoS Comput Biol* 15: e1007246

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

Eccleston RC, Manko E, Campino S, Clark TG, Furnham N (2023) A computational method for predicting the most likely evolutionary trajectories in the stepwise accumulation of resistance mutations. *eLife* 12: e84756

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

Echave J, Wilke CO (2017) Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu Rev Biophys* 46: 85-103

Ferreiro D, Branco C, Arenas M (2024a) Selection among site-dependent structurally constrained substitution models of protein evolution by approximate Bayesian computation. *Bioinformatics* 40: btae096

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

Ferreiro D, Khalil R, Sousa SF, Arenas M (2024b) Substitution Models of Protein Evolution with Selection on Enzymatic Activity. *Mol Biol Evol* 41: msae026

Fischer A, Vazquez-Garcia I, Mustonen V (2015) The value of monitoring to control evolving populations. *Proc Natl Acad Sci U S A* 112: 1007-1012

Fitch WM, Margoliash E (1967) A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem Genet* 1: 65-71

Fornasari MS, Parisi G, Echave J (2002) Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol* 19: 352-356

Gernhard T (2008) The conditioned reconstructed process. *J Theor Biol* 253: 769-778

Gerrish PJ, Sniegowski PD (2012) Real time forecasting of near-future evolution. *J R Soc Interface* 9: 2268-2278

Gilson AI, Marshall-Christensen A, Choi JM, Shakhnovich EI (2017) The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution. *Biophys J* 112: 1350-1365

Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79: 1396-1407

Goldstein RA (2013) Population Size Dependence of Fitness Effect Distribution and Substitution Rate Probed by Biophysical Model of Protein Thermostability. *Genome Biol Evol* 5: 1584-1593

Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2: e00631

Goyal S, Balick DJ, Jerison ER, Neher RA, Shraiman BI, Desai MM (2012) Dynamic Mutation–Selection Balance as an Evolutionary Attractor. *Genetics* 191: 1309-1319

Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185: 862-864

Harmon LJ (2019) Introduction to birth-death models. In: *Phylogenetic Comparative Methods*, p. https://lukejharmon.github.io/pcm/chapter10_birthdeath/.

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

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* 23: 183-201

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7: 1-44

Hudson RR (1998) Island models and the coalescent process. *Mol Ecol* 7: 413-418

Illergard K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* 77: 499-508

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G *et al* (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences* 110: 13067-13072

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583-589

Kimura M, Weiss GH (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics* 49: 561-576

Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications* 13: 235-248

Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22: 79-86

Lassig M, Luksza M (2014) Can we read the future from a tree? *Elife* 3

Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. *Nature Ecology & Evolution* 1: 0077

Le TK, Vinh LS (2020) FLAVI: An amino acid substitution model for flaviviruses. *J Mol Evol* In press

Lemant J, Le Sueur C, Manojlović V, Noble R (2022) Robust, Universal Tree Balance Indices. *Systematic Biology* 71: 1210-1224

Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. *Syst Biol* 53: 265-277

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

Lind PA, Libby E, Herzog J, Rainey PB (2019) Predicting mutational routes to new adaptive phenotypes. *Elife* 8

Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci U S A* 107: 2983-2988

Luksza M, Lassig M (2014) A predictive fitness model for influenza. *Nature* 507: 57-61

Malcom BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345: 86-89

Mendez R, Fritsche M, Porto M, Bastolla U (2010) Mutation bias favors protein folding stability in the evolution of small populations. *PLoS Comput Biol* 6: e1000767

Minning J, Porto M, Bastolla U (2013) Detecting selection for negative design in proteins through an improved model of the misfolded state. *Proteins* 81: 1102-1112

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108: E1293-1301

Moreira F, Arenas M, Videira A, Pereira F (2023) Evolution of TOP1 and TOP1MT Topoisomerases in Chordata. *Journal of Molecular Evolution* 91: 192-203

Morris DH, Gostic KM, Pompei S, Bedford T, Łuksza M, Neher RA, Grenfell BT, Lässig M, McCauley JW (2018) Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. *Trends Microbiol* 26: 102-118

Munck C, Gumpert HK, Wallin AI, Wang HH, Sommer MO (2014) Prediction of resistance development against drug combinations by collateral responses to component drugs. *Science translational medicine* 6: 262ra156

Navascues M, Depaulis F, Emerson BC (2010) Combining contemporary and ancient DNA in population genetic and phylogeographical studies. *Mol Ecol Resour* 10: 760-772

Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences* 110: 437-442

Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. *Elife* 3

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL (2007) HIV-specific probabilistic models of protein evolution. *PLoS One* 2: e503

Papkou A, Garcia-Pastor L, Escudero JA, Wagner A (2023) A rugged yet easily navigable fitness landscape. *Science* 382: eadh3860

Parisi G, Echave J (2001) Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol* 18: 750-756

Parisi G, Echave J (2005) Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene* 345: 45-53

Pascual-Garcia A, Abia D, Mendez R, Nido GS, Bastolla U (2010) Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. *Proteins* 78: 181-196

Pascual-Garcia A, Arenas M, Bastolla U (2019) The Molecular Clock in the Evolution of Protein Structures. *Syst Biol* 68: 987-1002

Poon AF, Kosakovsky Pond SL, Richman DD, Frost SD (2007) Mapping protease inhibitor resistance to human immunodeficiency virus type 1 sequence polymorphisms within patients. *J Virol* 81: 13598-13607

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

Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI (2016) Biophysical principles predict fitness landscapes of drug resistance. *Proceedings of the National Academy of Sciences* 113: E1470-E1478

Rubin IN, Ispolatov Y, Doebeli M (2023) Adaptive diversification and niche packing on rugged fitness landscapes. *J Theor Biol* 562: 111421

Ruiz-Gonzalez MX, Fares MA (2013) Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system GroES-L. *BMC Evol Biol* 13: 156

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234: 779-815

Santos-Pereira A, Triunfante V, Araujo PMM, Martins J, Soares H, Poveda E, Souto B, Osorio NS (2021) Nationwide Study of Drug Resistance Mutations in HIV-1 Infected Individuals under Antiretroviral Therapy in Brazil. *Int J Mol Sci* 22

Scheiblhofer S, Laimer J, Machado Y, Weiss R, Thalhamer J (2017) Influence of protein fold stability on immunogenicity and its implications for vaccine design. *Expert Rev Vaccines* 16: 479-489

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

Souto B, Triunfante V, Santos-Pereira A, Martins J, Araujo PMM, Osorio NS (2021) Evolutionary dynamics of HIV-1 subtype C in Brazil. *Sci Rep* 11: 23060

Spielman SJ (2020) Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. *Molecular Biology and Evolution* 37: 2110-2123

Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA (1990) The ribonuclease from an extinct bovid ruminant. *FEBS Lett* 262: 104-106

Stadler T (2010) Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* 267: 396-404

Stadler T (2011) Simulating trees with a fixed number of extant species. *Syst Biol* 60: 676-684

Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Ritter D, Sander C, Gal Y, Marks DS (2023) Learning from prepandemic data to forecast viral escape. *Nature* 622: 818-825

Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301: 1714-1717

Ugalde JA, Chang BSW, Matz MV (2004) Evolution of Coral Pigments Recreated. *Science* 305: 1433-1433

Van Cleve J, Weissman DB (2015) Measuring ruggedness in fitness landscapes. *Proceedings of the National Academy of Sciences* 112: 7345-7346

Watabe T, Kishino H (2010) Structural considerations in the fitness landscape of a virus. *Mol Biol Evol* 27: 1782-1791

Wilke CO (2012) Bringing molecules back into molecular evolution. *PLoS Comput Biol* 8: e1002572

Wiuf C, Posada D (2003) A coalescent model of recombination hotspots. *Genetics* 164: 407-417

Wortel MT, Agashe D, Bailey SF, Bank C, Bisschop K, Blankers T, Cairns J, Colizzi ES, Cusseddu D, Desai MM *et al* (2023) Towards evolutionary predictions: Current promises and challenges. *Evolutionary Applications* 16: 3-21

Wright S (1931) Evolution in Mendelian populations. *Genetics* 16: 97-159

Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences* 108: 9916-9921

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol Evol* 11: 367-372

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

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11: 316-324

Yoshida K, Hata K, Kawakami K, Hiradate S, Osawa T, Kachi N (2023) Predicting ecosystem changes by a new model of ecosystem evolution. *Scientific Reports* 13: 15353

Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences* 104: 16152-16157

Zhang J, Nei M (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol* 44 Suppl 1: S139-146

Tables

Table 1. Comparison of real and predicted sequences of the HIV-1 MA protein considering predictions based on the SCS and neutral models. For the data simulated under the SCS [including birth-death models with constant (SCS) and variable global birth-death rate among lineages (GlobalBDvar)] and neutral models, the table shows the Grantham distance between the amino acids that changed during the real and predicted evolutionary trajectories and, the Kullback-Leibler (KL) divergence between the real and predicted multiple sequence alignments. Next, it shows the folding stability (ΔG) of the real protein variants at times *T1* and *T31*, and the folding stability of the predicted protein variants at time *T31*. The error corresponds to the 95% confidence interval from the mean of predictions of folding stability.

	Grantham distance	**KL divergence**	**ΔG of the real variant at *T1* (kcal/mol)**	**ΔG of the real variants at *T31* (kcal/mol)**	**ΔG of the predicted variants at *T31* (kcal/mol)**	**ΔΔG (kcal/mol) at *T31* (predicted – real variants)**
SCS model	5%	6%	-9.72	-10.34 ± 0.14	-9.96 ± 0.02	0.38
SCS GlobalBDvar model	5%	6%	-9.72	-10.34 ± 0.14	-10.03 ± 0.03	0.31
Neutral model	5%	6%	-9.72	-10.34 ± 0.14	-9.21 ± 0.07	1.14

Figure captions

Figure 1. Illustrative example of forward in time simulation of protein evolution integrating a birth-death population evolutionary process with fitness from the protein folding stability and the modeling of protein evolution with a structurally constrained substitution model. Given a protein variant assigned to a node at time *t* (blue node), its fitness is calculated considering its protein folding stability. Then, the fitness is used to determine the birth and death rates for that variant, which provide the time to the next birth or death event (horizontal dashed line) that corresponds to the forward-in-time branch length. Next, the variant is evolved forward in time toward each descendant, upon the previously determined branch length, under a SCS model of protein evolution. The process is repeated, forward in time, starting at each new variant. If a death event occurs, the variant of the extinct node (pink node) is obtained but it does have descendants. The process finishes when a particular sample size or simulation time is reached (i.e., $t+n$).

Figure 2. Prediction error of SARS-CoV-2 Mpro and PLpro evolution under SCS and neutral models regarding physicochemical properties of the amino acid changes accumulated during the evolutionary trajectories and protein folding stability. Predictions based on data simulated under the SCS [including birth-death models with constant (SCS) and variable global birth-death rate among lineages (GlobalBDvar)] and neutral models. (A) Grantham distance calculated from the amino acid changes that occurred during the real and predicted evolutionary trajectories based on SCS and neutral models of protein evolution. (B) Variation of protein folding stability ($\Delta\Delta G$) between real and predicted protein variants based on SCS and neutral models of protein evolution. Notice that positive $\Delta\Delta G$ indicates that the real protein variants are more stable than the predicted protein variants and

vice versa. Error bars correspond to the 95% confidence interval of the mean of prediction error from 100 multiple sequence alignments simulated for the corresponding population and time.

Figure 3. Prediction error of influenza NS1 protein evolution under SCS and neutral models regarding physicochemical properties of the amino acid changes accumulated during the evolutionary trajectories and protein folding stability. Predictions were performed for two time points (longitudinal samples T2 and T3). Predictions based on data simulated under the SCS [including birth-death models with constant (SCS) and variable global birth-death rate among lineages (GlobalBDvar)] and neutral models. (A) Grantham distance calculated from the amino acid changes that occurred during the real and predicted evolutionary trajectories based on SCS and neutral models of protein evolution. (B) Variation of protein folding stability (ΔΔG) between real and predicted protein variants based on SCS and neutral models of protein evolution. Notice that positive ΔΔG indicates that the real protein variants are more stable than the predicted protein variants and *vice versa*. Error bars correspond to the 95% confidence interval of the mean of prediction error from 100 multiple sequence alignments simulated for the corresponding population and time.

Figure 4. Prediction error of HIV-1 PR evolution at diverse populations regarding physicochemical properties of the amino acid changes accumulated during the evolutionary trajectories and protein folding stability. (A) For each viral population (patient, represented with a particular color) and time, Grantham distance calculated from the amino acid changes that occurred during the real and predicted evolutionary trajectories. For each population, the mean of distances obtained over time is shown on the right. (B)

Relationship between Grantham distances and accumulated number of substitution events (R^2 = 0.0001, which indicates a lack of correlation between these parameters). (C) Variation of protein folding stability ($\Delta\Delta G$) between real and predicted protein variants at each viral population and time. For each population, the mean of distances obtained over time is shown on the right. Notice that positive $\Delta\Delta G$ indicates that the real protein variants are more stable than the predicted protein variants and *vice versa*. Error bars correspond to the 95% confidence interval of the mean of prediction error from 100 multiple sequence alignments simulated for the corresponding viral population and time.

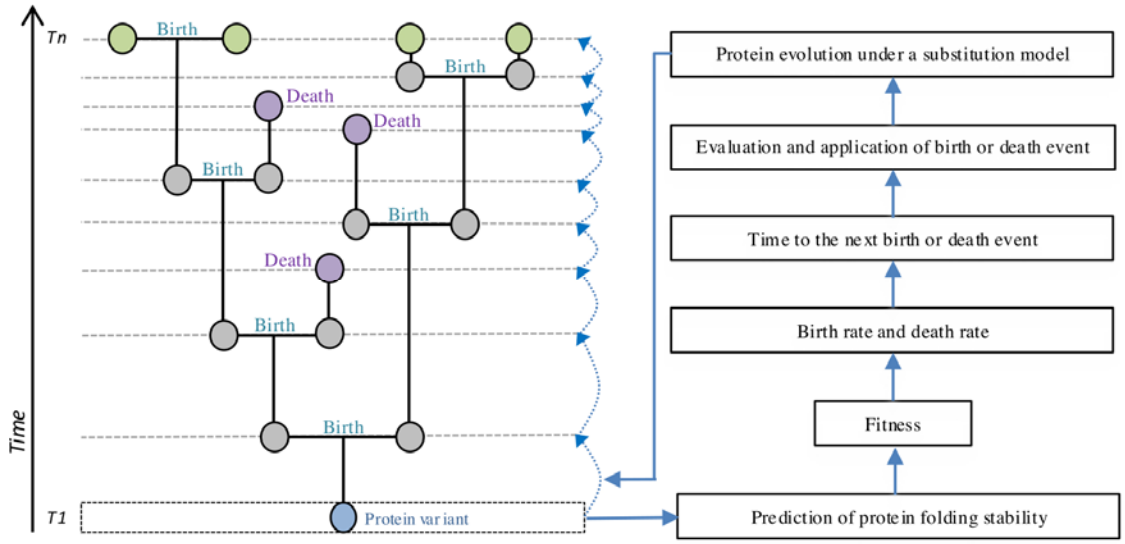
Tn
Time
T1
Birth
Death
Protein variant
Protein evolution under a substitution model
Evaluation and application of birth or death event
Time to the next birth or death event
Birth rate and death rate
Fitness
Prediction of protein folding stability

Figure 1.

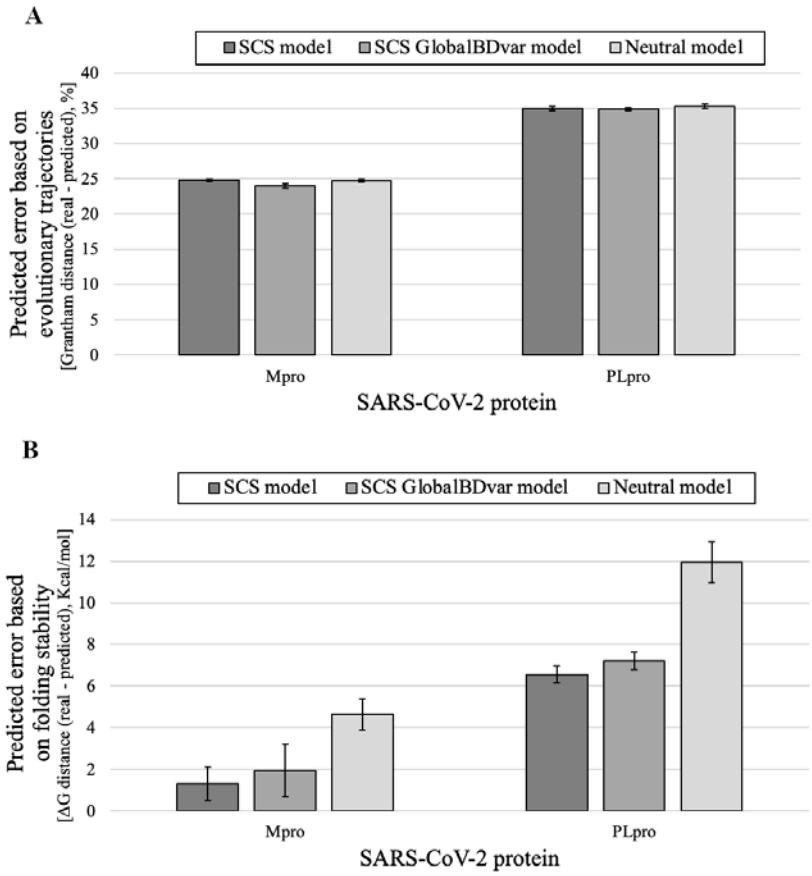
A
SCS model
SCS GlobalBDvar model
Neutral model
Predicted error based on evolutionary trajectories
[Grantham distance (real - predicted), %]
40
35
30
25
20
15
10
5
0
Mpro
PLpro
SARS-CoV-2 protein
B
SCS model
SCS GlobalBDvar model
Neutral model
Predicted error based on folding stability
[ΔG distance (real - predicted), Kcal/mol]
14
12
10
8
6
4
2
0
Mpro
PLpro
SARS-CoV-2 protein

Figure 2.

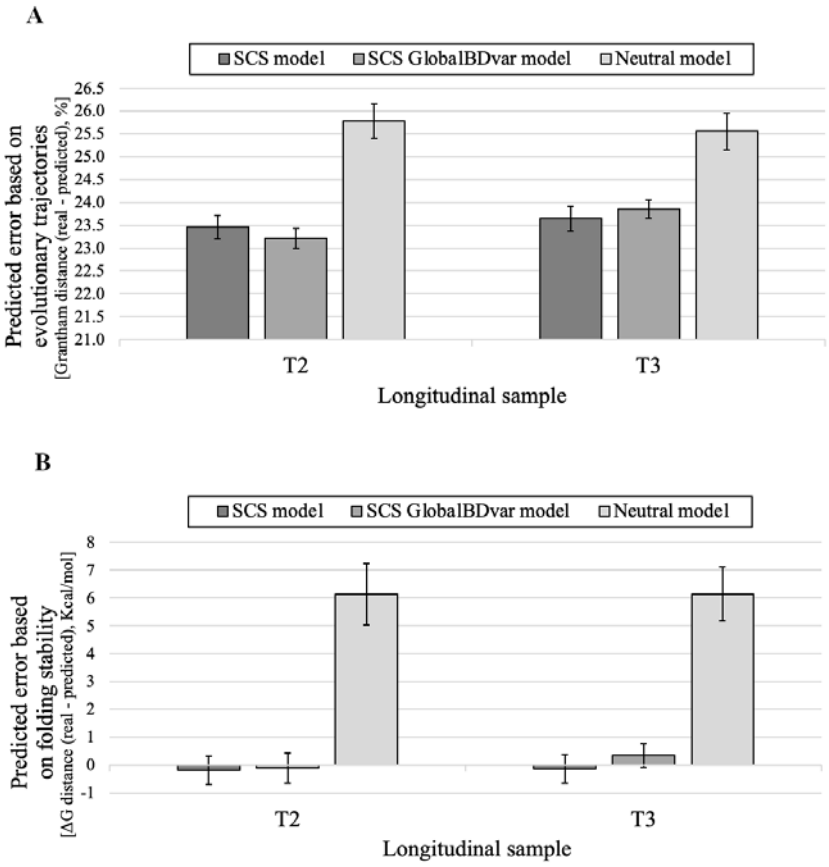
A
SCS model
SCS GlobalBDvar model
Neutral model
Predicted error based on evolutionary trajectories
[Grantham distance (real - predicted), %]
26.5
26.0
25.5
25.0
24.5
24.0
23.5
23.0
22.5
22.0
21.5
21.0
T2
T3
Longitudinal sample
B
SCS model
SCS GlobalBDvar model
Neutral model
Predicted error based on folding stability
[ΔG distance (real - predicted), Kcal/mol]
8
7
6
5
4
3
2
1
0
-1
T2
T3
Longitudinal sample

Figure 3.

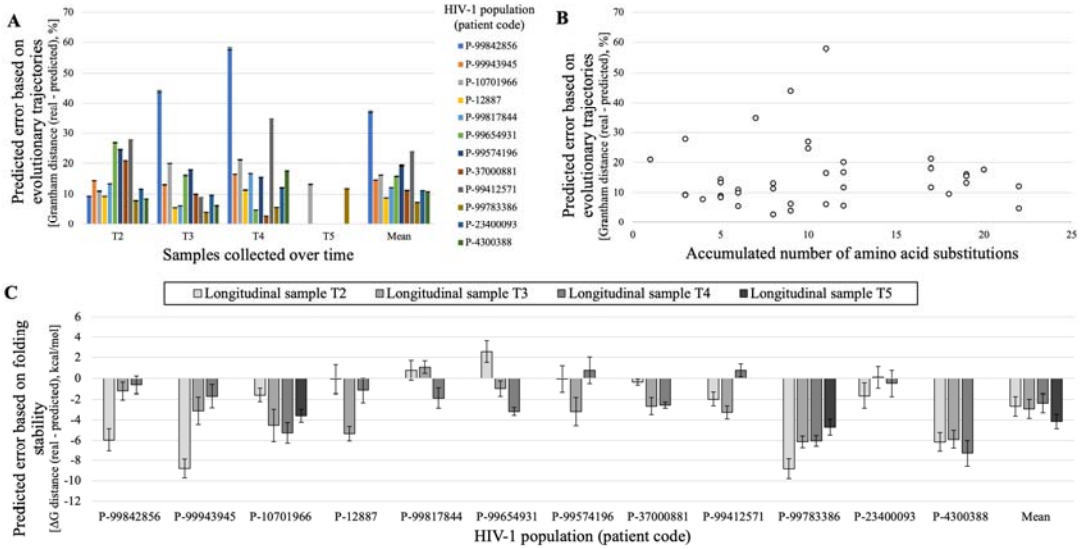
A
Predicted error based on evolutionary trajectories
[Grantham distance (real - predicted), %]
Samples collected over time
T2
T3
T4
T5
Mean
HIV-1 population (patient code)
P-99842856
P-99943945
P-10701966
P-12887
P-99817844
P-99654931
P-99574196
P-37000881
P-99412571
P-99783386
P-23400093
P-4300388
B
Predicted error based on evolutionary trajectories
[Grantham distance (real - predicted), %]
Accumulated number of amino acid substitutions
C
Longitudinal sample T2
Longitudinal sample T3
Longitudinal sample T4
Longitudinal sample T5
Predicted error based on folding stability
[ΔG distance (real - predicted), kcal/mol]
HIV-1 population (patient code)

Figure 4.