Syst. Biol. 0(0):1–17, 2019

DOI:10.1093/sysbio/syz022

The Molecular Clock in the Evolution of Protein Structures

ALBERTO PASCUAL-GARCÍA[1,2,3], MIGUEL ARENAS[1,4] AND UGO BASTOLLA[1,*]

[1]*Centro de Biologia Molecular "Severo Ochoa" CSIC-UAM Cantoblanco, 28049 Madrid, Spain;* [2]*Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK;* [3]*Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland;* [4]*Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain*

**Correspondence to be sent to: Centro de Biologia Molecular "Severo Ochoa" CSIC-UAM Cantoblanco, 28049 Madrid, Spain; E-mail: ubastolla@cbm.csic.es.*

Alberto Pascual-García and Miguel Arenas contributed equally to this article.

Received 8 March 2018; reviews returned 20 March 2019; accepted 9 April 2019
Associate Editor: Rachel Mueller

Abstract.—The molecular clock hypothesis, which states that substitutions accumulate in protein sequences at a constant rate, plays a fundamental role in molecular evolution but it is violated when selective or mutational processes vary with time. Such violations of the molecular clock have been widely investigated for protein sequences, but not yet for protein structures. Here, we introduce a novel statistical test (Significant Clock Violations) and perform a large scale assessment of the molecular clock in the evolution of both protein sequences and structures in three large superfamilies. After validating our method with computer simulations, we find that clock violations are generally consistent in sequence and structure evolution, but they tend to be larger and more significant in structure evolution. Moreover, changes of function assessed through Gene Ontology and InterPro terms are associated with large and significant clock violations in structure evolution. We found that almost one third of significant clock violations are significant in structure evolution but not in sequence evolution, highlighting the advantage to use structure information for assessing accelerated evolution and gathering hints of positive selection. Clock violations between closely related pairs are frequently significant in sequence evolution, consistent with the observed time dependence of the substitution rate attributed to segregation of neutral and slightly deleterious polymorphisms, but not in structure evolution, suggesting that these substitutions do not affect protein structure although they may affect stability. These results are consistent with the view that natural selection, both negative and positive, constrains more strongly protein structures than protein sequences. Our code for computing clock violations is freely available at https://github.com/ugobas/Molecular_clock.[Co-evolution; molecular clock; protein structure evolution; selection.]

In the early days of molecular evolution, the discovery that protein sequences accumulate amino acid substitutions at a roughly constant rate prompted the use of this molecular clock to infer evolutionary events (Zuckerkandl and Pauling 1962; Langley and Fitch 1973). The molecular clock hypothesis can be theoretically justified on the ground of the neutral theory of Kimura (Kimura 1983), while significant accelerations of the substitution rate are often used to assess positive selection (McDonald and Kreitman 1991; Kosakovsky Pond and Frost 2005). The first systematic analysis of the molecular clock was presented by Gillespie (1989), who found pervasive violations of this hypothesis, incompatible with the assumption that the substitution process follows a Poissonian process as supposed by Kimura, a result confirmed by subsequent analysis (Ayala 1999). Computer simulations of neutral protein evolution under stability constraints predicted that the substitution process is more disperse than a Poissonian process (Bastolla et al. 1999). This behavior was explained by the fact that the number of neutral neighbors fluctuates across sequence space, and lead to expect a non-Poissonian molecular clock even under neutral evolution.

In a pioneering paper, Chothia and Lesk (1986) showed that substitutions in protein sequences produce exponentially increasing changes in their native structures, measured through the root mean square deviation. This analysis was later extended to measures that better quantify the evolutionary divergence of protein structures as fractional change, and make it comparable with the divergence of protein sequences. It was found that closely related proteins diverge more slowly in structure than in sequence on the average (Illergard et al. 2009; Pascual-Garcia et al. 2010) while distantly related proteins tend to diverge fast in structure. These results naturally raise the question whether and to which extent an approximate molecular clock applies to the divergence of protein structures as it applies to the evolution of protein sequences.

Here, we present a novel test that allows inferring violations of the molecular clock not only for protein sequences but also for protein structures, which is unfeasible for the commonly adopted Tajima test. The test considers the influence of the intrinsic randomness of the substitution process under an overdispersed, non-Poissonian clock and uses all available outgroups to estimate the error of the estimated evolutionary divergences. Moreover, the test minimizes the impact of different functional conformations of the same protein that lead to structure changes in the absence of sequence changes. We evaluate the performances of the test through stability-constrained simulations of protein sequence evolution, which lead to an overdispersed substitution process. Unfortunately, protein structure evolution cannot be currently simulated except for a single mutational step (Echave 2008), but we think that the formal analogy between the measures of structure change and sequence change that we adopt justifies our test for protein structure evolution as well.

We apply our method to three large superfamilies in the CATH database (Orengo et al. 1997): the NADP

enzymes, the P-loop with mainly regulatory functions, and the Globins mainly involved in the storage and transport of oxygen. We measure structure divergence through the contact divergence (Pascual-Garcia et al. 2010) and the template modeling (TM) score (Zhang and Skolnick 2004). For both measures, structure change is slower than sequence change when both are measured in relative terms (fraction of different contacts or fraction of residues that do not superimpose in space, compared with fraction of sequence substituted residues), consistent with the common wisdom that protein structures are more conserved than sequences (Illergard et al. 2009; Pascual-Garcia et al. 2010).

We find that violations of the molecular clock in sequence and in structure are strongly correlated, so that the protein that evolves faster in sequence also tends to evolve faster in structure. However, one third of the significant violations of the molecular clock are significant in structure evolution but not in sequence evolution, indicating that structural information is useful for detecting accelerations of the evolutionary rate that may give insight on the strength of natural selection. Violations of the molecular clock are stronger and more significant for pairs of proteins related by a change of function annotation (FA), as indicated by their Gene Ontology (GO) or InterPro terms.

Materials and Methods

Sequence Divergences

We studied divergence measures based either on multiple sequence alignments (D^{seq}) performed with the program MAFFT (Katoh and Standley 2013) or on multiple structure alignments (D^{str}) performed with the program Mammoth-mult (Lupyan et al. 2005). We found that structure alignments are sensitive to conformation changes and are less reliable at reflecting evolutionary relationships (see Results section). Thus, the sequence identity (SI) presented below is obtained through sequence alignments. On the other hand, the structure divergence measures described below were based on both types of alignments, which are conceptually different and do not need to coincide. Sequence alignments represent homologous residues, while structure alignments represent pairs of residues that are structurally equivalent. Thus, it is legitimate to adopt sequence alignments for assessing evolutionary relationships and structure alignments for quantifying structural divergences. In the main text, we present results obtained with the smallest structural divergences between sequence alignments and structure alignments, but in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.2hs39cg, we show that they are robust with respect to other choices.

From the multiple alignment, we compute the SI between pairs of aligned sequences A and B, and we obtain two measures of sequence divergence,

$$D_{Poiss}(A,B) = -\ln\left(SI(A,B)\right) \quad (1)$$

$$D_{TN}(A,B) = -\ln\left(\frac{SI(A,B)-S_0}{1-S0}\right). \quad (2)$$

The first divergence is the Poisson divergence proposed by Kimura and Ohta (1971) and Dickerson (1971), which estimates the number of substitutions taking into account multiple substitutions at the same site. The second one is the Tajima–Nei (TN) divergence (Tajima and Nei 1984) that also takes into account that two nonhomologous amino acids at an aligned position may converge by chance with probability $S_0 \equiv \sum_a (f_a)^2 = 0.06$, where f_a is the frequency of amino acid a in a large sequence database. The TN divergence cannot be computed if $SI \leq S_0$, and it is unreliable if this threshold is approached, thus we omitted pairs with $SI < 0.10$. We also tested the p-distance $D_p(A,B) = 1 - SI(A,B)$, related with the Hamming distance.

Contact Divergence

We adopt a structure divergence measure based on the contact overlap q, which counts the normalized number of common contacts between a pair of aligned structures. Specifically, the contact matrix is $C_{ij} = 1$ if residues i and j are closer than 4.5 Å and they are not close in sequence ($|i-j| > 4$), while $C_{ij} = 0$ otherwise and the overlap between two aligned contact matrices is defined as

$$q(A,B) = \frac{\sum_{ij} C^{(A)}_{ij} C^{(B)}_{a(i)a(j)}}{\sqrt{\sum_{ij} C^{(A)}_{ij} \sum_{ij} C^{(B)}_{ij}}}, \quad (3)$$

where $a(i)$ is the residue of protein B aligned to residue i in protein A. q takes values between 0 and 1. Note that the computation of the overlap does not require structure superimposition.

From the contact overlap, we obtain the contact divergence measure D_{cont} (Pascual-Garcia et al. 2010), which estimates the evolutionary divergence of contact matrices in analogy with the TN sequence divergence:

$$D_{cont}(A,B) = -\log\left(\frac{q(A,B) - q_\infty(L)}{1 - q_\infty(L)}\right) \quad \text{if } q > \epsilon(L). \quad (4)$$

The parameter $q_\infty(L) \equiv \bar{q}(L) + A\sigma_q(L)$ denotes the limit overlap of distantly related protein pairs, L is the length of the shorter protein, $\bar{q}(L) = 0.386L^{-0.547}$ is the mean for unrelated proteins, $\sigma_q(L) = 1.327L^{-0.673}$ is the standard and $A = 5$. For pairs with $q < \epsilon(L)$ (not considered here), D_{cont} is based on the Z-score of the overlap with respect to pairs of unrelated proteins. The value $\epsilon(L)$ is fixed by imposing that the contact divergence is continuous for $q = \epsilon(L)$. For further details on the parameters see (Pascual-Garcia et al. 2010). D_{cont} is also referred to as CD in the text.

TM Score

We also consider a more fine-grained structure divergence measure based on the template-model (TM) score (Zhang and Skolnick, 2004), which measures the structural similarity between two aligned and superimposed proteins as:

$$\mathrm{TM} = \max\left(\frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right), \tag{5}$$

where the maximum is taken over all rotations, L is the aligned protein length, d_i is the distance between the ith pair of aligned residues, and $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ estimates the average distance between aligned residues of unrelated proteins. The TM score equals one for identical structures and takes approximately the value TM ≈ 0.17 independent of protein length for unrelated pairs. To make this quantity comparable with the CD and the sequence divergence measures, we transformed it into a divergence as $\mathrm{D_{TM}} = -\log(\mathrm{TM})$.

Conformational Changes

Structural divergence measures report both the evolutionary divergence between two proteins and the difference between the conformations observed in the experiments in which the structure was determined. In order to minimize this effect, we define $\mathrm{D_{cont}}(A,B)$ between two proteins A and B as the minimum value of $\mathrm{D_{cont}}$ over all of their experimentally determined conformations a and b and similar for the TM divergence,

$$\mathrm{D_{cont}}(A,B) = \min_{\{a\in A, b\in B\}}\mathrm{D_{cont}}(a,b) \tag{6}$$

Protein Data

We selected from the CATH database (Orengo et al. 1997) structural domains of the three following superfamilies, which are among the largest in CATH: NAD(P)-binding Rossmann-like Domain (accession code 3.40.50.720, 161 domains); P-loop containing nucleotide triphosphate hydrolases (3.40.50.300, 150 domains); Globins (1.10.490.10, 397 domains). The properties of these structural clusters are summarized in the Supplementary Material available on Dryad. They include proteins from both eukaryotic and prokaryotic organisms.

We grouped domains through the hierarchical clustering algorithm Complete Linkage (CL), which computes the divergence between two clusters as the maximum divergence between their elements. We used CL with the $\mathrm{D_{cont}}$ measure and threshold $\mathrm{D_{cont}} < 2.5$, so that all pairs of domains in the same cluster satisfy $\mathrm{D_{cont}} < 2.5$. This condition was imposed because it is not possible to obtain a multiple structure alignment with a common core if the structural domains are too divergent. We selected the largest cluster for each superfamily. To minimize the chances that the proteins are in different conformations, we grouped structures with the same sequence and computed the structure divergence of each pair of proteins as the minimum across all of their structures, Eq. (6).

Identification of Outgroups

For every group of three proteins, we decided which proteins are neighbors and which one is the outgroup depending on the smallest value of $D(A,B) - \frac{1}{n-2}\sum_C \big(D(A,C)+D(B,C)\big)$. In the main text, we assign outgroups based on the TN divergence, but in the Supplementary Material available on Dryad, we show that the results are robust when the structure divergence measures are used as $D(A,B)$.

Triangle Inequality

A distance must fulfil the triangle inequality of metric spaces that states that no intermediate point B can make the walk from A to C shorter than the direct path:

$$\mathrm{D}(A,C) \le \mathrm{D}(A,B) + \mathrm{D}(B,C). \tag{7}$$

However, even if the divergences D are often called distances, they are not distances in the mathematical sense since they can violate the triangle inequality (Felsenstein 2004). In particular, the fraction of different amino acids $1-\mathrm{SI}$ satisfies the triangle inequality if all positions of the multiple sequence alignment are taken into account, since it is proportional to the Hamming distance, which is a mathematical distance. Nevertheless, if the sequence identity is normalized by the length of the shortest sequence, as it is common practice and as we do here, $1-\mathrm{SI}$ may violate the triangle inequality due to indels, and $\mathrm{D_{Poiss}}$ and $\mathrm{D_{TN}}$ may violate it even in the absence of indels, because they are nonlinear functions of the Hamming distance. Similarly, $\mathrm{D_{cont}}$ and $\mathrm{D_{TM}}$ are not distances. Triples that violate the triangle inequality represent instances in which the divergence measures do not reliably estimate the evolutionary divergences, and they lead to overestimate the violations of the molecular clock. Therefore, we assess the validity of the molecular clock only with outgroups that do not violate the triangle inequality.

CV

We quantify the difference between the evolutionary rates of two proteins A and B as the average difference of their divergences with respect to all possible outgroups C, divided by their own divergence, which estimates the divergence time if an approximate molecular clock holds. We call this quantity clock violation CV (the symbol CV should not be confused with CV used in related contexts to designate the coefficient of variation),

$$\text{Clock Violations (CV)}(A,B) = \frac{\frac{1}{n_C}\sum_C \big(D(A,C)-D(B,C)\big)}{D(A,B)}. \tag{8}$$

Here $D(A,B)$ represents the divergence either in sequence or in structure. The triangle inequality implies that $|\mathrm{CV}| \leq 1$. For structure divergences, if the outgroup C has several associated conformations c, we compute CV using as outgroup the structure c that minimizes $|D(a,c)-D(b,c)|$, where a and b are the representative structures of the proteins A and B determined through Eq. 6.

The signifince of CV (SCV) Test

To determine whether a clock violation is significant, we must consider two possible sources of nonvanishing differences $D(A,C)-D(B,C)$. Firstly, they can be due to the intrinsic fluctuations of the evolutionary divergences on the two branches leading to A and to B expected under a molecular clock, which we assume to scale as $D(A,B)^{\alpha}$. Accordingly, we define the α-dependent SCV score (SCV_{α}) as

$$\mathrm{SCV}_{\alpha}(A,B) = \frac{\frac{1}{n_C}\sum_C \big(D(A,C)-D(B,C)\big)}{D(A,B)^{\alpha}}. \qquad (9)$$

Equation (9) generalizes the Poissonian clock for which $\alpha = 0.5$. We studied values of α in the range $[0.5, 2]$, checking that the main conclusions are robust in this range. Note that CV defined above to quantify the differences of evolutionary rates is equal to SCV with exponent $\alpha = 1$, $\mathrm{CV} \equiv \mathrm{SCV}_{\alpha=1}$.

Furthermore, we must consider the error of the inferred evolutionary divergences, which we quantify through the standard error of the mean (S.E.M.) of $D(A,C)-D(B,C)$ over the set of allowed outgroups (see main text). Considering both factors, we evaluate the significance of the violations of the molecular clock through the condition

$$\big(\mathrm{SCV}_{\alpha}(A,B)\big)^2 > (\mathrm{thr})^2 + \left(\frac{\mathrm{SEM}}{D(A,B)^{\alpha}}\right)^2, \qquad (10)$$

where $\mathrm{thr} = 0.185$ is a threshold that we obtained from the simulations of stability-constrained protein evolution (see below). We estimate the standard error of the mean as the standard deviation divided by the square root of the number of independent outgroups, $\mathrm{SEM} = \mathrm{SD}/\sqrt{N_{\mathrm{indep}}}$. When the number of independent outgroups increases, SEM decreases and becomes negligible. N_{indep} is difficult to estimate since outgroups are evolutionarily correlated. We address this problem by counting as independent only the fraction of the outgroup sequence that is different from the outgroups that have been already counted, which is an underestimate of the number of independent outgroups. Denoting the sequence identity by SI, we define $N_{\mathrm{indep}} = \sum_c \big(1 - \max_{c'<c} \mathrm{SI}(c,c')\big)$.

Test of the CV Method through Computer Simulations

In order to test the SCV method for sequence evolution and to estimate the exponent α of the intrinsic fluctuations of the divergence measures, we used the program ProteinEvolver (Arenas et al. 2013, available at https://github.com/MiguelArenas/proteinevolver) to simulate neutral protein sequence evolution with selection on protein folding stability. ProteinEvolver places the PDB sequence at the root of the input phylogenetic tree and evolves it forward in time, proposing mutations under a given substitution model of evolution (Yang 2006; Arenas 2012) such that the number of simulated mutations in each branch follows a Poisson distribution. Mutations are rejected if they reduce the estimated stability against both unfolding and misfolding below a threshold, otherwise they reach fixation. Stability is estimated from the folding free energy computed through the contact interaction matrix derived in Bastolla et al. (2000) computing the free energy of misfolded states as in Minning et al. (2013). The stability threshold is chosen as the stability 5% lower than the one of the sequence in the PDB and all other thermodynamic parameters took default values. These stability-constrained substitutions outperform empirical substitution models for different protein families and its results are very robust with respect to the protein structure (Arenas et al. 2013).

We simulated protein evolution along the phylogenetic tree of the NADP superfamily that we adopted in our empirical analysis (see Supplementary Fig. S3 available on Dryad), reconstructed using the UPGMA method in such a way that the branch lengths obey the molecular clock, that is, the sum of branch lengths from the root to all tips are equal. We then constructed 16 additional trees with the same topology, choosing eight pairs of proteins A and B that are sister in the phylogenetic tree, so that all other proteins are their outgroups. For each pair, we constructed two trees that violate the molecular clock such that the branch connecting the common ancestor to protein A was 50% (1) and 100% (2) longer than the branch from O to B. For each of these 17 trees we simulated 200 realizations of sequence evolution, in which the number of mutations in each branch is a Poissonian variable whose mean is proportional to the branch length and mutations are fixated according to the neutral stability model. We obtained a total of 3400 multiple sequence alignments (MSAs) where each sequence corresponds to a tip node of the tree, upon which we applied the SCV test.

Calibration of the SCV Threshold and the Exponent α

To calibrate the exponent α, we reasoned that the correct exponent has the property that SCV_{α} does not depend on $D(A,B)$ on the average for proteins evolving under the same molecular clock. We applied the SCV test for different values of α to eight pairs of proteins with different divergences $D(A,B)$ simulated under the molecular clock, and for each α we determined

the threshold such that the false positive rate is 0.05, quantified as the fraction of the 200 MSAs with SCV_α above threshold. In this way, we obtained curves that represent the threshold as a function of $D(A,B)$ for different exponents α. As expected, the curves are increasing functions of $D(A,B)$ for small α, while they are decreasing functions for large α. The value $\alpha=0.65$ separates the two behaviors, and it provides the best estimate of the exponent α. For real protein sequences we cannot apply the above method, since we lack a bona fide data set of protein pairs that evolve under the same clock.

Function Annotation (FA)

GO terms (Gene Ontology Consortium 2000) were retrieved from the web page of the Structure integration with function, taxonomy and sequence (SIFTS) initiative at the url http://www.ebi.ac.uk/msd/sifts/ and were used to assign the function of each protein. To avoid wrong assignments of GO, we removed the PDB chains that contained more than one CATH domain. We only used the molecular FA and we only considered GO terms that were manually assigned, that is, we required that the evidence code was one of the following: EXP (Inferred from Experiment), IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), IEP (Inferred from Expression Pattern), or TAS (Traceable Author Statement). All other evidence codes, such as ISS (Inferred from Sequence or Structural Similarity), were discarded. We only retained proteins for which the GO terms relative to molecular function were manually assigned.

For globins, GO terms are not specific enough, so we used InterPro signatures (Hunter et al. 2009), which we retrieved with the SSMap tool (David and Yip 2008) that relates PDB chains with UniProt accessions, including InterPro signatures. We refer to GO and InterPro terms as FA.

Results

The SCV Test of Violations of the Molecular Clock

Under the molecular clock, the divergence of two sister proteins A and B along branches of equal duration is equal apart from stochastic fluctuations. As often implicitly assumed in distance methods for phylogenetic inference, we assume here that our divergence measure $D(A,B)$ is the sum of the additive distance $d(A,B)$ plus the estimation error $\varepsilon(A,B)$, $D(A,B)=d(A,B)+\varepsilon(A,B)$. Ideally, we should quantify violations of the molecular clock through the difference $dr(A,B)\equiv d(A,C)-d(B,C)$. Additivity means that $d(A,B)$ is the sum of the distances traveled along all branches of the phylogenetic tree that connect A with B, and it implies that $d(A,C)-d(B,C)$ only depends on the distance traveled along the branches that connect A and B with their common ancestor and it is independent of C, that is, the violation of the molecular clock does not depend on the outgroup chosen to quantify it.

However, additivity does not hold for the estimated divergence measure $D(A,B)$, and we estimate the violations of the molecular clock through the difference $D(A,C)-D(B,C)$ that does depend on the outgroup C. In our analysis of empirical data, we observed that its sign can change depending on the outgroup. We conclude that using only one outgroup we cannot reliably detect which protein evolves faster. For simulated data with 50% difference in the clock rate, the fraction of outgroups for which the inferred faster protein was the incorrect one ranged from 1% to 17% for very small and large divergences $D(A,B)$, respectively. Thus, it is very important to use the information provided by all possible outgroups, as we show below.

To take into account that the additive property is violated, the Neighbor-Joining method (Saitou and Nei 1987) quantifies the difference of the branch lengths that connect A and B with their common ancestor as the unweighted mean over all possible outgroups, $\frac{1}{n_C}\sum_C \big(D(A,C)-D(B,C)\big)$. We adopt the same estimate in our CV and SCV formulas, Eq. (8) and Eq. (9). This is equivalent to assume that, under the molecular clock, the difference $\varepsilon(A,C)-\varepsilon(B,C)$ is unbiased, that is, it's expected value is zero. Furthermore, the standard error of the mean (S.E.M.) of $D(A,C)-D(B,C)=dr(A,B)+\big(\varepsilon(A,C)-\varepsilon(B,C)\big)$ over outgroups C equals the S.E.M. of $\varepsilon(A,C)-\varepsilon(B,C)$, justifying our formula Eq. (10).

The number of substitutions that happened in the evolution between A and B is clearly an additive distance, which also satisfies the triangle inequality, which is weaker. We estimate it through the Tajima-Nei (TN) divergence, which represents our $D(A,B)$. Similar considerations can be applied to the structure divergence measures. We tested with simulations that the TN divergence provides a good estimate of the number of simulated substitutions when the divergence is not too large. We found that the number of substitutions is proportional to $L\mathrm{D_{TN}}$ for $\mathrm{D_{TN}}<0.65$, corresponding to $\mathrm{SI}>0.55$, that is, up to 45% of the residues differ (see Supplementary Fig. S1 available on Dryad). Above this value the TN divergence grossly underestimates the number of substitutions, which scales approximately as the exponential $(L/0.74)\exp(0.74\mathrm{D_{TN}})$. This result implies that the error with respect to an additive distance is negative and its absolute value scales as $\exp(0.74D)$. This result suggests to performing a weighted average of the outgroups with weight $\exp\big(-\beta[D(A,C)+D(B,C)]\big)$. Nevertheless, our tests showed that the weighted average improves the detection of the molecular clock only very marginally and it introduces an extra parameter β. Therefore, we adopt in the following unweighted means consistent with the NJ method.

If additivity holds, the triangle inequality Eq.(7), a mathematical property of all distances, is automatically satisfied for all triples. In our simulations, we observed that the TN divergence violates the triangle inequality only when just one mutation was present, possibly

because our simulations do not produce indels: In the absence of indels, $1-\mathrm{SI}$ is proportional to the Hamming distance and it obeys the triangle inequality, thus the related TN divergence is more likely to obey the triangle inequality if indels do not occur.

The shortcomings of distance-based methods, arising from the violation of additivity and the dependence of the estimated clock violations on the outgroup, can be avoided by maximum likelihood (ML)-based methods that infer divergences for every branch of the phylogenetic tree and provide more accurate branch-length estimates. However, so far these methods cannot be applied to protein structure divergence. Therefore, we adopt distance-based methods for the SCV test since they can be equally applied to protein sequences and structures allowing their comparison, they are much simpler and computationally less demanding, and they can be corrected when multiple outgroups are available. ML and Bayesian approaches provide also more accurate phylogenetic trees. However, when the number of sites is large distance-based methods are accurate enough for phylogenetic inference, and they have the advantage that they are much simpler and consistent with our SCV test that is based on pairwise distances.

The violations of the molecular clock must be compared with the fluctuations arising from the intrinsic randomness of the substitution process. We assume that the intrinsic fluctuations scale with the mean divergence $D(A,B)$ as $D(A,B)^{\alpha}$, Eq. (9), which lead to Eq. (8) and Eq. (10). For the commonly assumed Poissonian clock (Fitch 1976), the standard deviation scales as $\sqrt{D(A,B)}$, that is, $\alpha=0.5$. Nevertheless, it has been observed both in empirical data (Gillespie 1989) and in simulations of protein evolution with stability constraints (Bastolla et al. 1999) that the number of substitutions is overdispersed, that is, its variance is larger than the mean or, in other words, $\alpha>0.5$. A simple explanation of overdispersion is that mutations are fixed with higher probability if they originate in sequences with a large number of neutral neighbors. Since the number of neutral neighbors is an auto-correlated variable, substitution rates are also auto-correlated in sequence space, which enhances their fluctuations. We studied values of the exponent α in the range $[0.5,2]$. For large divergences, larger α requires larger differences $D(A,C)-D(B,C)$ to decide that a clock violation is significant.

Assessment of the SCV Test through Simulations

To evaluate the SCV test in sequence evolution, we simulated the stability-constrained model of protein sequence evolution along the rooted tree of the NADP superfamily (see Supplementary Fig. S3 available on Dryad) that we studied empirically, with branches determined with the UPGMA method (Sokal and Michener 1958) so that they fulfil the molecular clock, that is, all branch length from root to tip are equal. We simulated evolution under the molecular clock, subject to Poissonian fluctuations in the number of proposed mutations that are accepted if and only if the stability is above a threshold, which is known to lead to an overdispersed substitution process. We considered eight pairs of sister proteins A and B spanning a broad range of divergences (branch lengths) from 0.004 to 0.41, and for every pair we simulated evolution with the average number of proposed mutations from the common ancestor to A 50% and 100% faster than the same number on the branch leading to B, leading to CV Eq. (8) equal to 1/5 and 1/3, respectively. We simulated 200 MSA for each scenario and we applied the SCV test for different values of α, determining for each α and each $D(A,B)$ the threshold such that the false positive rate is 0.05 when the molecular clock holds. As expected from Eq. (9), the threshold increases with $D(A,B)$ for small α and decreases for large α. The threshold shows the weakest dependence on $D(A,B)$ for $\alpha=0.65>0.5$ for the TN divergence and the Poisson divergence and $\alpha=0.55>0.5$ for the p-distance $1-\mathrm{SI}(A,B)$ (see Supplementary Fig. S2 available on Dryad), which confirms that the number of substitutions is overdispersed under selection for protein folding stability.

We compared our method with the Tajima test of the molecular clock (Tajima 1993). We averaged the Tajima parameter over all of the outgroups, which enhances the ability of the test to detect clock violations. Figure 1 reports the fraction of detected violations for the eight divergences and the three values of CV using the balanced value of α and the threshold $t=0.185$ for the TN divergence and $t=4$ for the Tajima test, such that the false positive rate is smaller than 5% in the absence of clock violations for all values of $D(A,B)$. The detected violations show a clear tendency to increase with the divergence $D(A,B)$ and with CV, as expected since these variables increase the number of differences that can be used to detect a clock violation. The SCV method appears to be slightly superior to the Tajima method for small $D(A,B)<0.2$ and slightly inferior for $D(A,B)>0.2$, but differences are small. Therefore, we conclude that the accuracy of our method is comparable to the Tajima method, which is still the state of the art for assessing the molecular clock (Battistuzzi et al. 2011). These results hold qualitatively for all values of α that we tested. Since the Tajima test cannot be easily applied to protein structure evolution, in the following we apply the SCV test to investigate the divergences of both protein sequences and structures.

We also tested the effect of considering only the outgroup that is closest to the common ancestor of A and B and omitting the S.E.M. from Eq. (10). The results presented in the third and fourth columns of Figure 1 show that the performances of both methods experience a decrease, which is smaller for the more stable Tajima's method.

The Influence of the Alignment in the Estimation of Evolutionary Divergences

We now analyze the evolution of three large protein superfamilies: NADP, P-loop, and globins (see

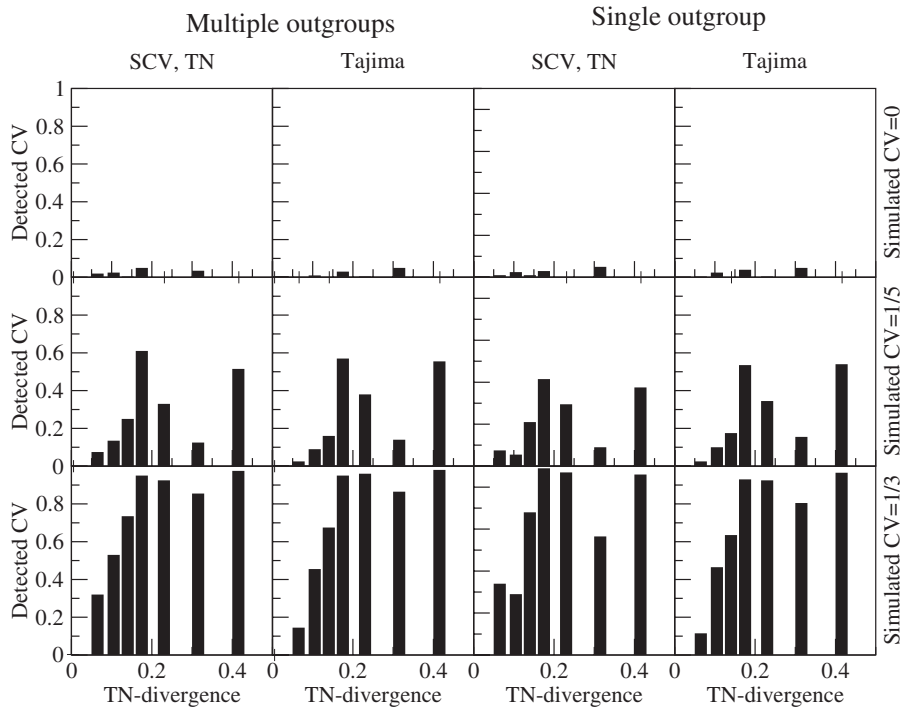

FIGURE 1. Fraction of violations of the molecular clock in simulated protein evolution detected using the Tajima test (leftmost) and the SCV test versus the estimated divergence $D(A,B)$ for $\alpha = 0.65$ that makes the threshold independent of the simulated branch length. The three lines show different simulated clock violations (CV = 0, 0.2, 0.333 corresponding to differences of 0%, 50%, and 100% of the simulated rate on the branches leading to A and to B). The first two columns depict thresholds obtained separately for each divergence value, the third and fourth column depict the more stringent thresholds that are valid for all $D(A,B)$.

Supplementary Figs. S3–S5 available on Dryad). We first compare divergences estimated from sequence and from structure alignments. As expected, sequence identities obtained from structure alignments are always smaller than those obtained from sequence alignments (see Supplementary Fig. S6 available on Dryad). This can be interpreted either as the overfitting of the sequence alignments that match residues at the expense of a poor structural match, or as an artifact of structure alignments that place spurious gaps in order to accommodate conformation changes. To weight the second effect, we examined pairs with SI = 1 according to sequence alignments, which correspond to conformation changes. Their structure alignments present SI as low as 0.77 (P-loop) or 0.92 (NADP), which indicates that spurious gaps have been introduced by the structure alignment program as a consequence of conformation changes. This result speaks against the use of structure alignments for evolutionary studies. Therefore, we analyze violations of the molecular clock adopting sequence alignments for estimating sequence divergences.

On the other hand, for proteins with similar structures ($D_{cont} < 1$) the D_{cont} and TM score measures obtained through sequence alignments are similar to those obtained with structure alignments and there is not any clear bias (see Supplementary Figs. S7 and S8 available on Dryad), which indicates that both kinds of alignment are almost equivalent for structurally similar proteins. Finally, for high structure divergence the sequence alignments heavily overestimate D_{cont} based on the linear trend observed for $D_{cont} < 1$ (see Supplementary Figs. S7 and S8 available on Dryad), evidencing the poor quality of the structure alignments derived from distant sequence alignments, a recognized problem in homology modeling (Abagyan and Batalov 1997). Note that sequence alignments and structure alignments are conceptually different. The first ones infer an evolutionary relationship between residues, and the second ones infer a structural relationship such as same secondary structure and similar buriedness. These relationships coincide in many cases, but they do not need to coincide always, for instance when indels occur in secondary structure elements and, as a result, nonhomologous residues occupy the same position in the secondary structure. Therefore, it is legitimate to infer evolutionary relationships with sequence alignments, as we do here, and structural relationship with structure alignments. To verify that our results are robust, we estimated structural divergences in three different ways: with multiple sequence alignments, with multiple structure alignments, and with the best structural score between the two kinds of alignments. The results presented in the main text are obtained in the last way. Results obtained with other types of alignments are qualitatively equivalent and they are reported in the Supplementary Material available on Dryad.

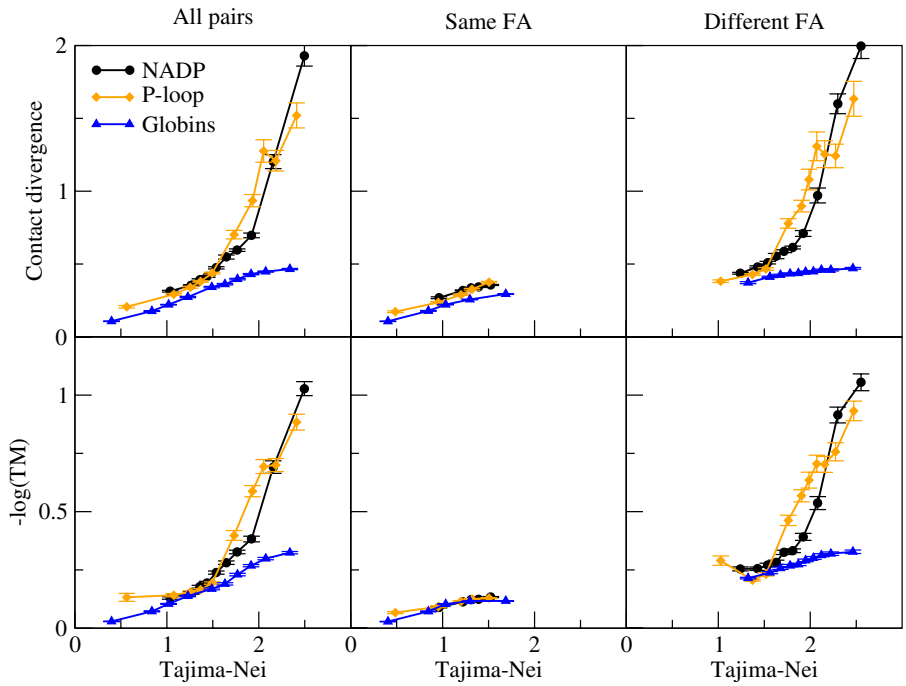

FIGURE 2. Structure divergence (top, contact divergence; bottom, −log(TMscore)) versus sequence divergence (Tajima-Nei) for three superfamilies. Left, all pairs of proteins. Center, only pairs with the same functional annotation (FA). Right, only pairs with different FA.

Structure Evolution is Constrained by Function Conservation

We now discuss the relationship between sequence, structure, and function divergence. The main results were already reported in Pascual-Garcia et al. (2010), but we also report them here since they are important for our discussion. The sequence divergence that correlates strongest with structure divergence is the Tajima-Nei divergence, and we shall use it throughout the article if not otherwise stated. Qualitatively similar results were obtained with the Poisson divergence.

For closely related proteins ($D_{TN} < 1.5$), divergences in sequence and structure are linearly correlated. In this regime, the slopes of the curves in Figure 2 are smaller than one, ranging from 0.22 (globins) to 0.25 (NADP) for contact divergence (CD) and even smaller, from 0.12 (P-loop) to 0.16 (NADP) for TM, see Table 1. Since the three divergence measures are computed in the same way from frequencies of SI, contact identity and identity of superimposed residues (TM score), we can compare them quantitatively and conclude that structural measures (fraction of superimposed structure and fraction of identical contacts) evolve more slowly than the fraction of sequence-identical residues, as previously reported (Illegard et al. 2009; Pascual-Garcia et al. 2010).

Strikingly, structure divergence is more severely limited for protein pairs with the same FA than with different FA, as previously reported (Pascual-Garcia et al. 2010). For pairs with different FA, the average structure divergence grows more than linearly with the sequence divergence and it can reach high values (1 for TM and 2 for CD), while for pairs with the same FA we only observe the linear regime and structure divergences are smaller than 0.25 (see Fig. 2 center and right columns). Furthermore, when the FA is conserved the slope of structure divergence versus sequence divergence is smaller than for all pairs (compare columns b–c and d–e of Table 1). These results are consistent with the observation that for SI between 30% and 70%, orthologous domains that tend to share the same function are more structurally similar than paralogous domains that tend to have different functions (Peterson et al. 2009). This suggests that sequence and structure rates stem from selective constraints on protein function

TABLE 1. Relation between sequence divergence (Tajima-Nei, TN) and structure divergence (CD and TM) for three superfamilies.

Super family	Seq., Struct.	Pairs[a]	Slope TN-CD[b] All	Slope TN-CD[c] Same FA	Slope TN-TM[d] All	Slope TN-TM[e] Same FA
NADP	74, 161	1788	0.25±0.03	0.16±0.02	0.16±0.03	0.079±0.007
P-loop	53, 150	1343	0.24±0.03	0.20±0.02	0.12[f] ±0.03	0.065±0.006
Globins	71, 397	2424	0.22±0.01	0.15±0.01	0.13±0.01	0.073±0.017

[a]Number of sequence pairs.
[b]Slope of contact divergence versus sequence divergence for $D_{TN} < 1.5$.
[c]Same for pairs with the same function annotation (FA, GO terms for P-loop and NADP and InterPro for Globins).
[d]Slope of TM score divergence with respect to sequence divergence in the linear regime.
[e]Same for pairs with the same FA.
[f]For the P-loop superfamily the point with smallest sequence divergence is omitted from the fit since it is heavily influenced by structures with different FA that have larger TM divergence than more closely related pairs.

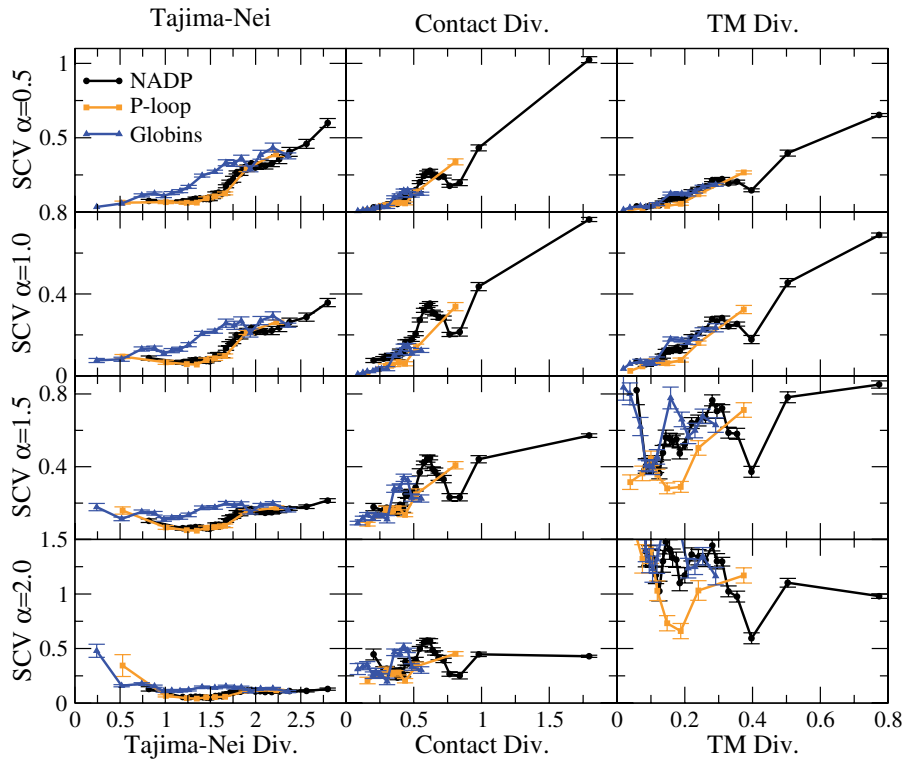

FIGURE 3. SCV versus evolutionary divergence for three divergence measures (left, Tajima-Nei sequence divergence; center, CD; right, TM score divergence) and four exponents (from top to bottom $\alpha = 0.5, 1, 0, 1.5, 2.0$).

more than from how the protein structure responds to sequence changes, since the same sequence changes induce smaller structural changes when the FA is conserved.

CVs Are Consistent in Sequence and Structure Evolution

Now, we examine clock violations in the evolution of the three superfamilies. The total number of examined sequence pairs is 3220 for NADP, 4384 for P-loop and 6656 for Globins.

In Figure 3, we show the significance score SCV_α versus the corresponding divergence measure for four values of the exponent α. Note that CV, which estimates the difference of evolutionary rates, is equivalent to $SCV_{\alpha=1}$ (second line from top). SCV_α increases with the divergence when the differences $|D(A,C) - D(B,C)|$ increase faster than $D(A,B)^\alpha$, otherwise it decreases. For proteins that evolve under the same molecular clock, the dependence of SCV_α on $D(A,B)$ should disappear when α equals the exponent of the fluctuations of the divergence measure. However, in our data sets there are protein pairs that evolve with truly different rates, thus an increasing trend can be explained by the argument that evolutionarily more distant proteins should evolve with more different rates. Thus, Figure 3 cannot determine the exponent of fluctuations α, but it suggests that $\alpha = 0.5$ is too small, consistent with the results of our simulations, since SCV_α increases very rapidly with the divergence, and $\alpha = 2$ is too large, since SCV_α decreases with the divergence, supporting values in the range $1 \leq \alpha \leq 1.5$ for all three divergence measures.

Figure 4 shows the fraction of pairs that satisfy Eq. (10) with the threshold $SCV_{thr} = 0.185$ suggested by simulations as a function of the corresponding divergence measure for three values of $\alpha = 0.7, 1.0, 1.5$. In the rest of the article, we call these pairs "significant pairs." However, since we lack a *bona fide* control set of proteins that evolve in sequence and structure under the molecular clock, we cannot determine a threshold that guarantees a given false positive rate as we did with simulations. Interestingly, for small divergence significant CVs are frequently observed for sequence divergences (Fig. 4 left plots). This observation is consistent with the finding that the substitution rate is larger for pairs of species separated by short time intervals (Ho et al. 2005). In contrast, CVs of structure divergence are seldom significant for small evolutionary divergence (Fig. 4).

We quantify violations of the molecular clock through the estimated difference of the evolutionary rates $CV \equiv SCV_{\alpha=1}$, Eq. (8). This choice of course does not mean that the correct exponent of the fluctuations is $\alpha = 1$. Figure 5 represents the relationship between CV obtained with different divergences, showing that they tend to be consistent and correlated. CV obtained with structure divergences are well correlated with CV obtained with the TN divergence (on the average, $r = 0.68$ for TM and $r = 0.79$ for CD), and even more with each other (on the

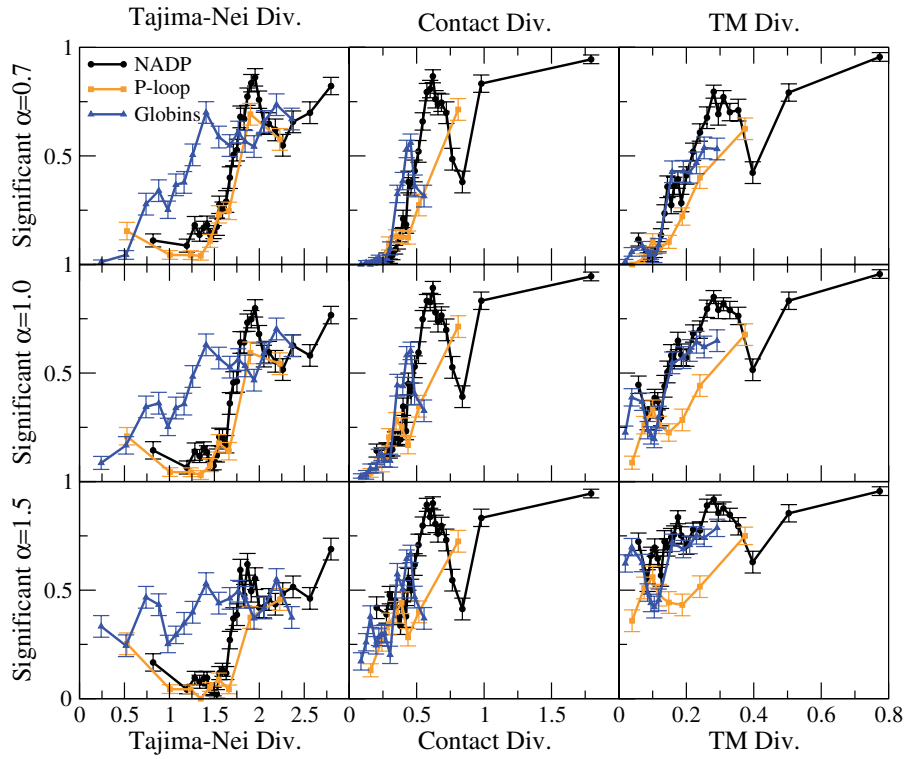

FIGURE 4. Fraction of pairs that satisfy Eq. (10) versus evolutionary divergence for three divergence measures (left, TN sequence divergence; center, contact divergence; right, TM score divergence) and three exponents $\alpha = 0.7, 1.0, 1.5$.

average, $r = 0.86$), see Supplementary Figure S9 available on Dryad. Note that for each outgroup C, the sign of CV is positive if protein A diverged more than protein B with respect to C, otherwise it is negative. Hence, CV is large and significant only if most outgroups consistently support a CV with the same sign. Therefore, these strong correlations indicate that the lineage that evolves faster in sequence tends to evolve faster in structure as well, which supports the interpretation that the violation of the molecular clock is due to faster evolution of a protein with respect to the other one rather than to errors of the inferred evolutionary divergences. The weakest correlations are found for the Globins superfamily.

Violations of the Molecular Clock Are Stronger in Structure Evolution than in Sequence Evolution

The statistical properties of CV are summarized in Figure 6. Because of the normalization, the values of CV depend little on the scale of the evolutionary divergence, so we can meaningfully compare CVs of sequence (TN) and structure divergences (CD and TM). For the NADP and P-loop superfamilies the absolute value of CV tend to be larger for structure divergence (TM and CD) than for sequence divergence (TN), see Figure 6 top plots. An exception is the Globin superfamily, for which the TN divergence yields average values of CV that are intermediate between CD and TM divergence (Fig. 6 top left plot). The other sequence divergence measures present smaller and less significant clock violations than TN (Supplementary Fig. S10 available on Dryad), confirming the result that violations of the molecular clock are stronger in structure evolution. We tested that this result is robust for all possible ways of assigning outgroups, based on the TN, CD, and TM divergence, and for all ways of computing structure divergence, with Sequence and Structure alignments, see Supplementary Figures S11 and S12 available on Dryad. The fraction of "significant" pairs with SCV_α above threshold for $\alpha = 1$ shows the same behavior as above (Fig. 6 top right plot), which is maintained for higher values of α. However, for $\alpha = 0.7$ the TN divergence has the highest fraction of significant pairs, as it can be seen from Figure 4.

For all divergence measures and all superfamilies, there is a noticeable fraction of outgroups, from 13% to 27%, which supports a sign of $D(A,C) - D(B,C)$ different from the majority sign (see Fig. 5 bottom left and Supplementary Fig. S13 available on Dryad), highlighting the large fluctuations from one outgroup to the other one. However, the majority sign of CV tends to be the same for structure and sequence evolution, as shown by Figure 5. The fraction of triples that violate the triangle inequality, which evidence inconsistencies in the estimated divergences, is smaller than 1% for the TN and contact divergence, for which it is smallest, and it is smaller than 2% for the TM divergence for which it is largest (see Fig. 5 bottom right). This suggests that the TM score is less reliably estimated, possibly because it requires a structural superimposition that is not needed

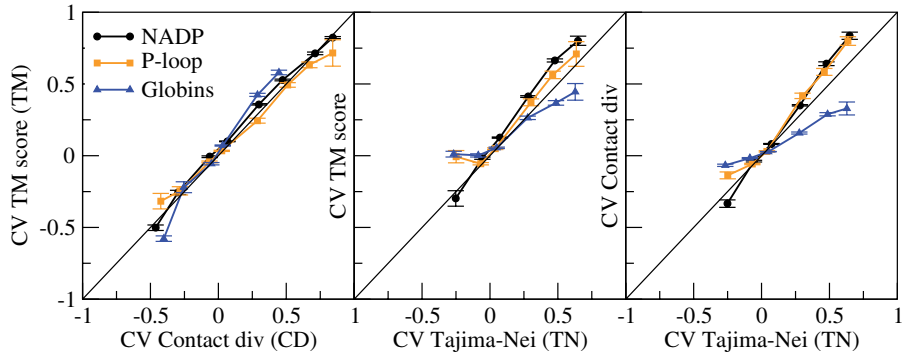

FIGURE 5. Relationships between average CV obtained with different measures.

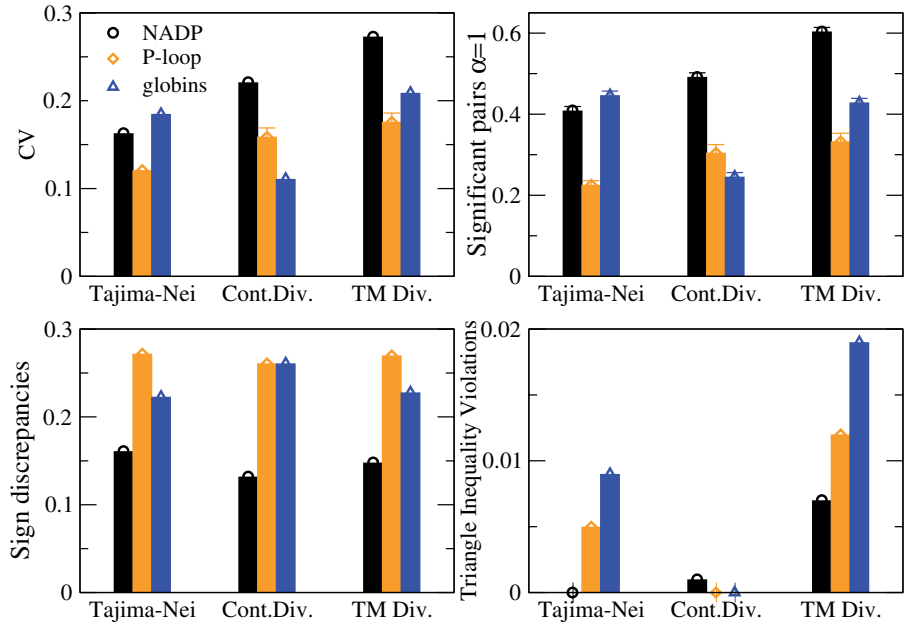

FIGURE 6. Comparison between CV in sequence evolution (Tajima-Nei) and in structure (contact divergence and TM divergence) for the three superfamilies. Top left: average of the absolute value of CV, Eq. (8). Top right: Fraction of pairs with SCV_α above threshold ("significant pairs") computed with $\alpha = 1$. Bottom left: Average fraction of outgroups that show sign discrepancies for any given pair. Bottom right: average fraction of violations of the triangle inequality.

for the other divergence measures. In order to reduce the effect of these inconsistencies, we did not consider triples that violate the triangle inequality, and we considered the variation due to different outgroups in the SCV test.

Fast rates of sequence evolution, such as those detected as a high ratio dN/dS between amino acid substitutions and synonymous substitutions (Nei and Gojobori 1986), are commonly regarded as a signature of positive selection. We may wonder which is the advantage of considering CV in structure, if this is so strongly correlated with CV in sequence. To address this question, we distinguished pairs that show significant SCV_α according to Eq. (10) both in structure and in sequence (StrSeq), in structure but not in sequence (StrNoSeq), in sequence but not in structure (NoStrSeq) and in neither. To favor the significance in sequence, we use $\alpha = 0.7$ for the TN divergence and $\alpha = 1$ for the structural divergences. The results are summarized in Figure 7, whose top left plot shows that a sizeable fraction of pairs, 12–19% depending on the superfamily, present significant SCV in structure but not in sequence, highlighting the advantage to use structure information for assessing accelerated evolution and gathering hints of positive selection. In contrast, a smaller fraction, 3–12%, present significant SCV in sequence but not in structure, and 14–38% present significant SCV both in sequence and in structure. As a result, 26–46% of protein pairs with significant SCV are significant in structure evolution but not in sequence evolution. The figures improve if we use $\alpha = 1$ for all divergence measures, yielding 27–52% of significant pairs that are significant in structure but not in sequence instead of 26–46 (see Supplementary Fig. S14 available on Dryad). The pairs with significant SCV in structure but not in sequence present the smallest divergences both in structure and in sequence (Fig. 7 top right and bottom), indicating

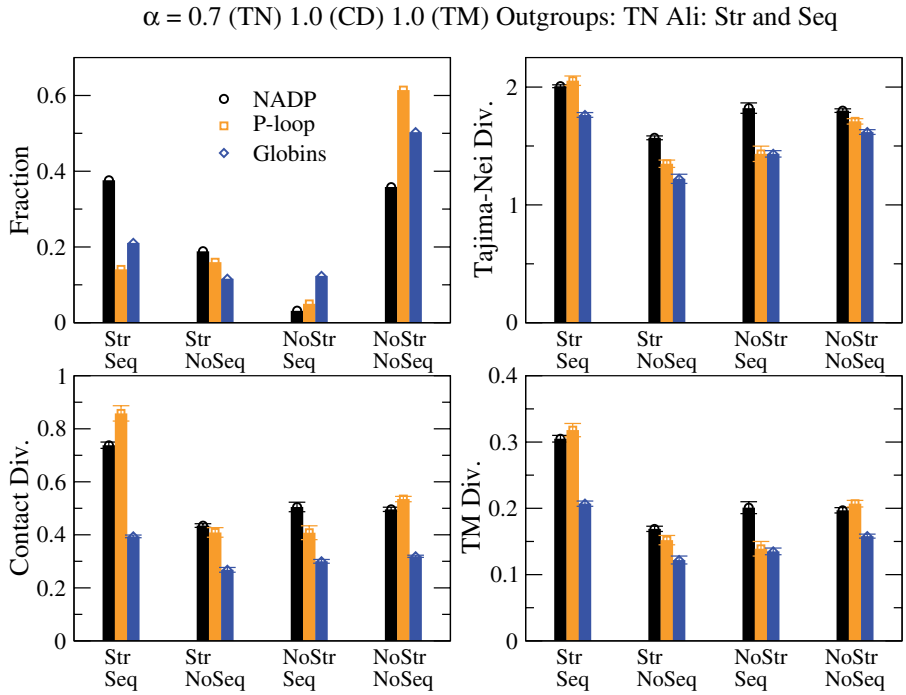

FIGURE 7. Pairs that show significant CV according to Eq. (10) both in structure and in sequence (StrSeq), in structure but not in sequence (StrNoSeq), in sequence but not in structure (NoStrSeq) and in neither. Top left: Fraction of pairs. Other plots: Average divergences of each group. $\alpha = 1.0$ for structure divergence measures and $\alpha = 0.7$ for Tajima-Nei divergence.

that they are difficult cases of closely related proteins for which the structural information is key for detecting accelerated evolution.

Changes of Protein Function Enhance CV

Finally, we investigated whether CVs are systematically influenced by changes in function annotation (FA). To this end, we distinguish pairs of proteins with very similar FA (function similarity ≥ 0.95) and different FA (function similarity ≤ 0.7). The results are plotted in Figure 8. For all divergence measures and all superfamilies, the average value of CV is systematically and significantly larger for pairs with different FA than for pairs with similar FA, and CV is significant for a larger fraction of pairs of proteins. We tested that these results are robust for all possible ways of assigning outgroups, based on the TN, CD, and TM divergence, and for all ways of computing structure divergence, with Sequence and Structure alignments, see Supplementary Figures S15 and S16 available on Dryad. The difference is smaller for the P-loop superfamily than for the Globins and NADP superfamilies, possibly because the concept of function is more difficult to define for this superfamily, many of which members are protein kinases.

The fraction of pairs with sign of CV different from the majority sign is smaller for pairs of proteins with different FA (see Fig. 8 bottom plots), indicating that, for these proteins, the protein that evolves faster is more clearly identified, possibly because it is the protein whose function changed with respect to the common ancestor. Also this result is robust with respect to the way of assigning outgroups and computing structural divergences, see Supplementary Figure S17 available on Dryad.

DISCUSSION

In this work, we introduced the SCV test of the molecular clock hypothesis that can be applied to the evolution of protein sequences and structures quantified through pairwise divergences. We considered a simple measure of sequence divergence, the Tajima-Nei divergence based on SI, and two measures of structure divergence, one discrete and based on the overlap between aligned contact matrices (contact divergence), and the other one continuous and based on the overlap between aligned and superimposed coordinates (TM divergence). Despite their simplicity, these divergences have a strong mathematical analogy that allows to compare them quantitatively.

As other tests of the molecular clock, ideally the SCV test would require an additive distance such as the number of substitutions that take place on each branch of the phylogenetic tree, so that CV is the same irrespective of the outgroup that we use to evaluate it. We tested through simulations that the TN divergence approximates this additive distance well when $D_{TN} < 0.7$, but it severely underestimates the number of substitutions when the divergence is larger. We think that this underestimate is likely due to the assumption

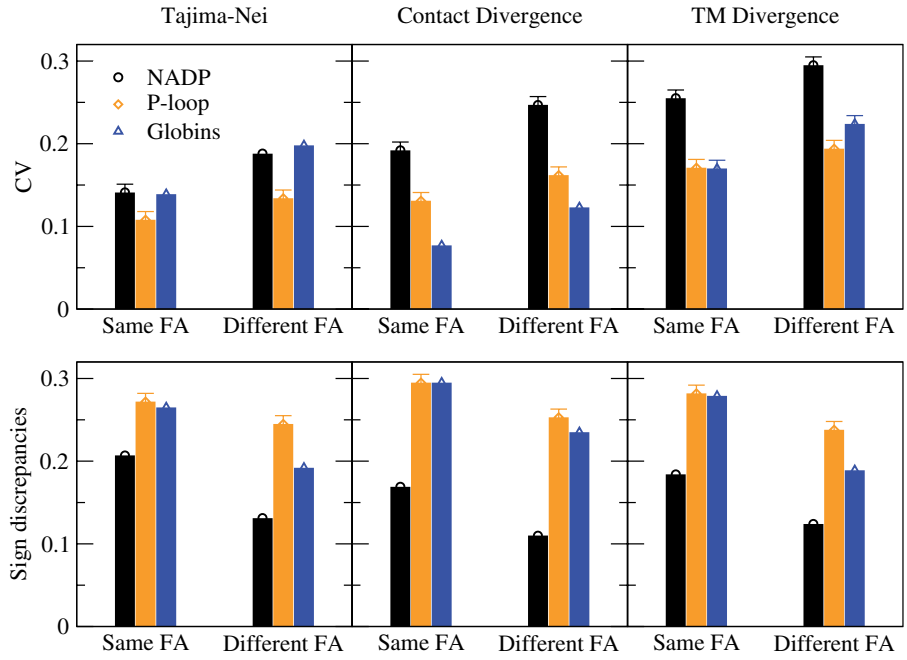

FIGURE 8. Violations of the molecular clock of sequence (left) and structure evolution (center and right) distinguishing protein pairs with the same and different FA. Top: Average value of the CV parameter. Bottom: Fraction of pairs with sign of CV different from the majority sign. One can see that violations of the molecular clock are larger for pairs with different FA.

that all protein sites evolve at the same rate and have the same stationary frequency of amino acids, whereas the evolutionary rates and sequence entropies vary largely across protein sites (Echave et al. 2016). However, site-specificity is seldom considered in models of molecular evolution (Hapern and Bruno 1998; Arenas et al. 2015) and to our knowledge it has not been implemented yet into algorithms that estimate pairwise divergences. Likewise, indels can lead to violations of the triangle inequality but the problem of incorporating indels in models of molecular evolution has not yet been solved (Holmes 2017). Despite these technical problems of the divergence measures, we tested through simulations that the SCV method is as effective as the Tajima's method (Tajima 1993) at detecting violations of the molecular clock in protein sequence evolution for equal false positive rate.

An important property of our SCV test with respect to other traditional tests of the molecular clock such as those by Fitch (1976) and Tajima (1993) is that the difference $D(A,C)-D(B,C)$ is averaged over all suitable outgroups C, producing results equivalent to the difference of the branch lengths estimated through the Neighbor-Joining method. Since $D(A,C)-D(B,C)$ fluctuates significantly from one outgroup to the other due to violations of the additive property, the SCV test also considers the standard error of the mean so that CV is not significant if it changes frequently sign. This increases the power of the test.

We assume that the fluctuations of the divergence scale as $D(A,B)^{\alpha}$, which generalizes the commonly assumed Poisson distribution with $\alpha=0.5$. Simulating protein evolution with stability constraints, we found that the optimal exponent for the Tajima-Nei divergence is $\alpha=0.65$, confirming that it is more dispersed than a Poisson process (Gillespie 1989), consistent with previous simulations (Bastolla et al. 1999). Results for real superfamilies and structural divergences suggest that $0.5<\alpha<2$ for all divergence measures, but we could not determine a precise value of α since we lack a control set that evolves under the same molecular clock.

We considered three large superfamilies with rather distinct properties. P-loop and NADP are among the three superfamilies with largest structural divergence in the CATH database, they have very large functional diversity, as witnessed by their 3602 and 1285 FunFam clusters (Sillitoe et al. 2013), respectively, and they are almost ubiquitously represented in 73,675 and 24,798 species. On the other hand, Globins have more reduced structural divergence, probably because of their lower functional diversity (compare Fig. 2 center and right) and they are present in only 22 FunFam clusters and 6224 species. Despite these differences, they yielded similar properties concerning the molecular clock. Adopting the SCV test, we obtained the following main results:

1. Clock violations in sequence and structure are well correlated (average correlation coefficient 0.74), showing that, when one protein evolves more rapidly than the other one in sequence, it also tends to evolve more rapidly in structure. Moreover, for all divergence measures and all superfamilies, at least 73% of the outgroups identify the same protein of the pair as the one evolving more rapidly (See Supplementary Fig. S13 available

on Dryad). These results strongly suggest that violations of the molecular clock are not due to errors on the evolutionary divergences but to the systematic enhancement of the evolutionary rate of one protein with respect to the other.

2. CV both in structure and in sequence tend to increase with the corresponding divergence. In other words, distantly related proteins tend to differ substantially in their evolutionary rates.

3. Despite the strong correlation between CV in sequence and in structure, there are many pairs whose SCV is not significant in sequence but is significant in structure. This may happen when there is a similar number of sequence changes in the branches that we compare, but they are more biased to maintain the protein structure on one branch than on another one, either due to relaxed negative selection for structure conservation or due to positive selection for structure change. This result suggests that considering structure evolution increases the sensitivity to detect possible accelerations of the evolutionary rate driven by positive selection.

4. CV tends to be larger in protein structure evolution, indicating that this is less clock-like than protein sequence evolution. The exception is the Globin superfamily, which is characterized by a remarkable structural and functional conservation even in case of change of InterPro term.

5. Clock violations are larger and more significant for pairs of proteins that change function, as indicated by their GO or InterPro terms.

Variations of the substitution rate over time (Ayala 1999; Bromham and Penny 2003) can be attributed to multiple processes, including mutational forces (Kvikstad and Duret 2014), relaxation of negative selection associated with decreased population size (Ohta 1976; Moran 1996; Bastolla et al. 2004), or positive selection (Fitch et al. 1991; Franks and Weis 2008; Sironi et al. 2015; Padhi and Parcells 2016). In particular, several methods interpret enhanced rates of amino acid substitutions as evidence of positive selection (e.g., McDonald and Kreitman 1991; Massingham and Goldman 2005; Kosakovsky Pond and Frost 2005), although they have been criticized on the ground that compensatory substitutions can be confounded with positive selection (Dasmeh et al. 2014).

We tested two measures of protein structure divergence with different properties, one based on CD, which is discrete and coarser, and one based on TM, which corresponds to continuous changes of the atomic coordinates. We found that the TM measure presents a slower rate of divergence with respect to sequence divergence and a larger CV than the CD measure. Both observations suggest that the TM measure is subject to stronger selection than the CD measure, possibly because TM enforces the precise conservation of the protein structure, which determines native dynamics. However, the TM measure also presents larger violations of the triangle inequality, indicating that this divergence is less reliably estimated. In fact, different from the CD measure, the TM measure is evaluated after optimal pairwise spatial superimposition, which introduces additional noise and possible inconsistencies among triples, so that the comparison between TM and CD remains an interesting open question without a clear conclusion.

For the largest observed divergences, clock violations tend to be larger for protein structure (both CD and TM score) than for protein sequence evolution for the NADP and P-loop supererfamily but not for the Globin superfamily, see Figure 3 with $\alpha = 1$. We propose that both positive (diversifying) selection and relaxation of negative (purifying) selection contribute to the difference between sequence and structure divergence. Protein structure is under stronger purifying selection than protein sequence, as evidenced by the fact that the fraction of structural changes is smaller than the fraction of sequence changes (Illegard et al. 2009; Pascual-Garcia et al. 2010; see also Fig. 2), especially when the protein function is conserved (Pascual-Garcia et al. 2010, Fig. 2 and Table 1). Therefore, we find more plausible to attribute stronger CVs for protein structure than for protein sequence evolution to stronger positive selection acting on sequence change that change the structure rather than to more relaxed negative selection, although our data are not conclusive. These differences are unlikely due to mutational processes, which affect in a similar way both types of change.

It is natural to expect that large CVs are often associated with function change, since function change is expected to enhance positive selection for improving the new function and to relax negative selection, at least until the new function has been optimized, and it is associated with larger structural changes than function conservation (Pascual-Garcia et al. 2010 and Fig. 2). To test this expectation, we compared protein pairs that conserve the same FA as indicated by their manually annotated GO and InterPro terms, and protein pairs with different FA. We found that, as expected, the latter present larger and more significant CV in structure evolution (Fig. 8; see also Supplementary Figs. S15 and S16 available at Dryad). Our results are consistent with the analysis of two enzyme families performed by Lai et al. (2012), who estimated that the rates of change of the native dynamics predicted from protein structure through elastic network models are faster at branches where protein function diverged. Note, however, that larger CV for pairs with different FA may also be attributed to the fact that they present larger divergences and that CV tends to increase with the divergence.

In addition to the influence of function changes, we also observed significant violations of the molecular clock of structure evolution when the FA is conserved (Fig. 8). These CV may be explained by other sources of variability or by the coarseness of the function

defined through GO and InterPro terms. In particular, coevolution may have a systematic influence on the substitution rate, as assumed by Valencia et al., who proposed methods for inferring protein–protein interactions based on the hypothesis that the substitution rates of interacting proteins are correlated (Pazos et al. 1997; Ochoa et al. 2015). The success of these methods suggests that the substitution rate of a protein changes in response to evolutionary changes of its interaction partners. It would be very interesting to quantify this influence both in sequence and in structure evolution.

Taken together, our results support the view that protein structure is strongly constrained by functional requirements, as supported by the strong conservation of structure for proteins that conserve the function (Pascual-Garcia et al. 2010 and Fig. 2) and by the strong relationship between native structure, dynamics in the native state, and function (Haliloglu and Bahar 2015) among other observations. In contrast, stability is regarded as a more neutral character (Goldstein 2011; Serohijos and Shakhnovich 2014) and it is thought that mutations that decrease stability but conserve structure are neutral or only slightly deleterious and they are often fixed. The interpretation that selection targets protein structure more strongly than protein stability is relevant for modeling the influence of protein structure in evolution. Two classes of such models exist. In stability-constrained models (reviewed in Goldstein 2011; Serohijos and Shakhnovich 2014; Bastolla et al. 2017) selection only targets protein stability, and the possible effects of mutations on the structure do not affect the fitness. In structure constrained models (Echave 2008) selection targets protein structure, and the structural effect of mutations is estimated through the elastic network model (Tirion 1996) under the assumption that the stability does not change. Of course mutations affect both structure and stability, but current models cannot predict both effects at the same time, and each class of models neglects a specific effect. Our results underscore the importance of considering changes in protein structure for estimating the fitness effect of mutations, as structure constrained models do. This is in line with the result of a recent work of our group, which found that stability-constrained models of protein evolution are too tolerant to mutations, probably because they neglect that they can modify the protein structure (Jimenez et al. 2018).

In conclusion, under function conservation protein structures tend to evolve under an approximate molecular clock that is consistent with the approximate molecular clock of protein sequences but is slower in relative terms, reflecting stronger selective constraints on protein structure evolution. The emerging view is that under function conservation sequence mutations are preferentially fixed if they conserve the structure even if they do not strictly conserve stability. In contrast, in a changing molecular environment, either due to function change or, possibly, due to coevolution, positive selection, and relaxed negative selection act on protein sequence and structure and enhance their evolutionary rates causing violations of the molecular clock (Fig. 8), resulting in much larger relative structure changes than those that take place under function conservation (Fig. 2).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.2hs39cg.

Funding

This work has been funded by the Spanish Ministry of Science [BIO2016-79043-P and BFU2012-40020]; the Spanish Government [RYC-2015-18241] and the Xunta de Galiciaž [ED431F 2018/08] to M.A.; the Simons Foundation [542381 to A.P.G., in part].

Acknowledgments

We thank Jeff Thorne and Huw Ogilvie for comments on a previous version of the paper, and we also thank the Editor and the anonymous reviewers for their constructive comments.

References

Abagyan R.A., Batalov S. 1997. Do aligned sequences share the same fold? J. Mol. Biol. 273:355–368.
Arenas M. 2012. Simulation of molecular data under diverse evolutionary scenarios. PLoS Comput. Biol. 8:e1002495.
Arenas M., Dos Santos H.G., Posada D., Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics 29:3020–3028.
Arenas M., Sanchez-Cobos A., Bastolla U. 2015. Maximum likelihood phylogenetic inference with selection on protein folding stability. Mol. Biol. Evol. 32:2195–2207.
Ayala F.J. 1999. Molecular clock mirages. BioEssays 21:71–75.
Bastolla U., Roman H.E., Vendruscolo M. 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J. Theor. Biol. 200:49–64.
Bastolla U., Vendruscolo M., Knapp EW. 2000. A statistical mechanical method to optimize energy functions for protein folding Proc. Natl. Acad. Sci. USA 97:3977–3981.
Bastolla U., Moya A., Viguera E., van Ham RC. 2004. Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J. Mol. Biol. 343:1451–1466.
Bastolla U., Dehouck Y., Echave J. 2017. What evolution tells us about protein physics, and protein physics tells us about evolution. Curr. Opin. Struct. Biol. 42:59–66.
Battistuzzi F.U., Filipski A.J., Kumar S. 2011. Molecular clock: testing. Chichester: John Wiley & Sons Ltd.
Bromham L., Penny D. 2003. The modern molecular clock. Nat. Rev. Genet. 4:216–224.
Chothia C., Lesk A.M. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826.
Dasmeh P., Serohijos A.W., Kepp K.P., Shakhnovich E.I. 2014. The influence of selection for protein stability on dN/dS estimations. Gen. Biol. Evol. 6:2956–2967.
Dos Santos H.G., Klett J., Méndez R., Bastolla U. 2013. Characterizing conformation changes in proteins through the torsional elastic response. Biochim. Biophys. Acta 1834:836–846.

David F.P., Yip Y.L. 2008. SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. BMC Bioinformatics 9:391.
Dickerson R.E. 1971. The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. 1:26–45.
Echave J. 2008. Evolutionary divergence of protein structure: the linearly forced elastic network model. Chem. Phys. Lett. 457:413–416.
Echave J., Spielman S.J., Wilke C.O. 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17:109–121.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Fitch W. 1976. Molecular evolutionary clocks. In: Ayala F.J. editor. Molecular evolution. Sunderland (MA): Sinauer Associates. p. 160–178.
Fitch W.M., Leiter J.M.E., Li X., Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA 88:4270–4274.
Franks S.J., Weis A.E. 2008. A change in climate causes rapid evolution of multiple life-history traits and their interactions in an annual plant. J. Evol. Biol. 21:1321–1334.
Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. Nat. Genet 25:25–29.
Gillespie J.H. 1989. Lineage effects and the index of dispersion of molecular evolution. Mol. Biol. Evol. 6:636–647.
Goldstein R.A. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:1396–1407.
Haliloglu T., Bahar I. 2015. Adaptability of protein structures to enable functional interactions and evolutionary implications. Curr. Opin. Struct. Biol. 35:17–23.
Halpern A., Bruno W.J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.
Ho S.Y.W., Phillips M.J., Cooper A., Drummond A.J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. 22(7):1561–1568.
Holmes I.H. 2017. Solving the master equation for Indels. BMC Bioinformatics 18:255.
Huang T.T., del Valle Marcos M.L., Hwang J.K., Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14:78.
Hunter S., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bork P., Das U., Daugherty L., Duquenne L., Finn R.D., Gough J., Haft D., Hulo N., Kahn D., Kelly E., Laugraud A., Letunic I., Lonsdale D., Lopez R., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Mulder N., Natale D., Orengo C., Quinn A.F., Selengut J.D., Sigrist C.J., Thimma M., Thomas P.D., Valentin F., Wilson D., Wu C.H., Yeats C. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37(Database Issue):D211–D215.
Illergard K., Ardell D.H., 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.
Jimenez M.J., Arenas M., Bastolla U. 2018. Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. Mol. Biol. Evol. 35:743–755.
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Kimura M., Ohta T. 1971. On the rate of molecular evolution. J. Mol. Evol. 1:1–17.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.
Kosakovsky Pond S.L., Frost S.D. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208–1222.
Kvikstad E.M., Duret L. 2014. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. Mol. Biol. Evol. 31:23–36.
Lai J., Jin J., Kubelk J., Liberles D.A. 2012. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. J. Mol. Biol. 422:442–459.
Langley C.H., Fitch W.M. 1973. An estimation of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.
Lupyan D., Leo-Macias A., Ortiz A.R. 2005. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 21:3255–3263.
Massingham T., Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169: 1753–1762.
McDonald J.H., Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.
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.
Moran N.A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA 93: 2873–2878.
Nei M., Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.
Ochoa D., Juan D., Valencia A., Pazos F. 2015. Detection of significant protein coevolution. Bioinformatics 31:2166–2173.
Ohta T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Popul. Biol. 10:254–275.
Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M. 1997. CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108.
Pascual-Garcia A., Abia D., Méndez R., Nido G.S., Bastolla U. 2010. Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78:181–196.
Padhi A., Parcells M.S. 2016. Positive selection drives rapid evolution of the meq oncogene of Mareks disease virus. PLoS One 11(9): e0162180.
Pazos F., Helmer-Citterich M., Ausiello G., Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271:511–523.
Peterson G.I., Masel J. 2009. Quantitative prediction of molecular clock and Ka/Ks at short timescales. Mol. Biol. Evol. 26: 2595–2603.
Peterson M.E., Chen F., Saven J.G., Roos D.S., Babbitt P.C., Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci. 18:1306–1315.
Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.
Serohijos A.W., Shakhnovich E.I. 2014. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr. Opin. Struct. Biol. 26:84–91.
Sillitoe I., Cuff A.L., Dessailly B.H., Dawson N.L., Furnham N., Lee D., Lees J.G., Lewis T.E., Studer R.A., Rentzsch R., Yeats C., Thornton J.M., Orengo C.A. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. 41(Database issue):D490–D498.
Sironi M., Cagliani R., Forni D., Clerici M. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat. Rev. Gen. 16:224–236.
Sokal R, Michener C. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38:1409–1438.
Tajima F, Nei M. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1:269–285.
Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607.
Tirion M.M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 77:1905–1908.
Tokuriki N, Tawfik D.S. 2009. Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. 19:596–604.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.
Zhang Y., Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710.
Zuckerkandl E., Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M., Pullman B., editors. Horizons in biochemistry. Nueva York: Academic Press. p. 189–225.