The molecular clock in the evolution of protein structures

Alberto Pascual-García[1,2,4] Miguel Arenas[1,3,4] and Ugo Bastolla[1]

[1] Centro de Biologia Molecular "Severo Ochoa"

CSIC-UAM Cantoblanco, 28049 Madrid, Spain. E-mail: ubastolla@cbm.csic.es

[2] Dept of Life Sciences, Imperial College London,

Silwood Park Campus, Ascot, United Kingdom.

[3] Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain.

[4] These authors contributed equally to this work.

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 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 GO and InterPro terms are associated with large and significant clock violations in structure evolution. 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 have an effect on stability. These results are consistent with the view that natural selection, both negative and positive, constrains more strongly protein structures than protein sequences. Finally, we suggest that significant clock violations between pairs of proteins that conserve their functional annotation may be influenced by the co-evolution between interacting proteins, which has been proposed to affect the substitution rate. Our codes are freely available at https://ub.cbm.uam.es/software/software.php?menu=long.

Abbreviations: CD, contact divergence; CV, clock violations; FA, functional annotation (GO and

Introduction

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). Although the molecular clock hypothesis is not required by current methods that reconstruct phylogenetic trees (Bouckaert et al. 2014), phylogenetic reconstruction is expected to become unreliable in the presence of erratic evolution that strongly violates the molecular clock (Bromham and Penny, 2003).

In the context of population genetics, the existence of an approximate molecular clock was justified by the neutral theory of Kimura (Kimura, 1983), which predicts that the rate at which neutral mutants are fixed in a population equals the rate with which neutral mutants are produced in the reproductive process, and it is independent of the population size. The neutral theory was later generalized by Ohta to include nearly neutral mutations whose fitness effect is of the order of the inverse of the population size (Ohta, 1976). In this framework, the substitution rate is predicted to be weakly dependent of the population size and therefore it is not strictly constant throughout evolution (Bromham and Penny 2003). Another generalization of the neutral theory postulates that the substitution rate is roughly constant in terms of number of generations instead of millions of years, predicting a generation time effect that has been confirmed by empirical evidence (Britten, 1986). However, corrections to the neutral theory are expected to be weak, and large accelerations of the substitution rate at the amino acid level, assessed as the ratio between nonsynonymous and synonymous substitutions (dN/dS), are often used to evaluate molecular signatures of selection (Kosakovsky Pond and Frost, 2005; Yang 2007; Arenas et al. 2015). This reasoning produced several tests for detecting positive selection based on dN/dS. In particular, the McDonald and Kreitman test compares synonymous and non-synonymous variation within a population (polymorphism) and between species (substitutions) (McDonald and Kreitman, 1991). These tests have been criticized because it is not possible to establish whether accelerated evolution is the result of positive selection that leads to adaptation, of compensatory mutations that compensate a previous fitness loss, or of relaxation of negative selection (Zhai, Nielsen and Slatkin, 2009). Moreover, these tests of neutrality are strongly affected by the across-genome heterogeneity of the mutation rate (Kvikstad and Duret, 2014) and by the recent finding that multi-nucleotide mutations happen more frequently than expected under the assumption that mutations are independent (Venkat et al. 2017).

The first systematic analysis of the molecular clock hypothesis was due to Gillespie (1989), who examined triples of proteins and found pervasive violations of the molecular clock, incompatible with the hypothesis that the substitution process follows a Poissonian process as assumed by Kimura. Several other analysis confirmed broad variations of the molecular clock in protein sequence evolution (Ayala, 1997). Moreover, neutral evolution is predicted to be more disperse than a Poissonian process if the viable neutral mutants

InterPro); GO, gene ontology; TI, triangle inequality; TM, TM-score; TN, Tajima-Nei divergence.

have to satisfy stability constraints that fluctuate across sequence space (Bastolla et al., 1999).

Here we study violations of the molecular clock in the evolution of protein structures and sequences. We consider three large superfamilies in the CATH database (Orengo et al., 1997): The NADP enzymes, the Ploop with mainly regulatory functions and the Globins mainly involved in the storage and transport of oxygen. We adopt the contact divergence (Pascual-Garcia et al. 2010) and the TM-score (Zhang and Skolnick, 2004) as measures of divergence in structure space. Although structure changes can be due to both conformation changes without sequence change and to substitutions in the protein sequence, we developed a methodology to carefully reduce the impact of conformation changes (see Materials and Methods). Our results confirm that structural divergence is well correlated with sequence divergence. If comparable measures of sequence and structure changes are adopted, structure change is found to be slower, consistent with the common wisdom that protein structures are more conserved than sequences (Illergard et al., 2009; Pascual-Garcia et al. 2010). Despite the good correlation, we find that violations of the molecular clock are stronger in structure than in sequence evolution, and that these significant violations of the molecular clock are more frequent for pairs of proteins related by a change of function annotation (FA), as indicated by their GO or InterPro terms.

Materials and Methods

Sequence divergences

The first step for computing the evolutionary divergences consists in performing a multiple alignment. We performed multiple sequence alignments (Seq) with the program MAFFT (Katoh and Standley 2013) and multiple structure alignments (Str) with the program Mammoth-mult (Lupyan, Leo-Macias and Ortiz, 2005).

From both kinds of alignments, we obtained the sequence identity SI between each pair of aligned sequences A and B and from it we computed three types of sequence divergences that estimate the number of substitutions occurred in their evolutionary divergence:

$$\mathrm{D_p}(A,B) = 1 - \mathrm{SI}(A,B) \tag{1}$$

$$\mathrm{D_{Poiss}}(A,B) = -\ln\left(\mathrm{SI}(A,B)\right) \tag{2}$$

$$\mathrm{D_{TN}}(A,B) = -\ln\left(\frac{\mathrm{SI}(A,B) - S_0}{1 - S0}\right) \tag{3}$$

The first divergence is the p-distance that measures the fraction of different amino acids. The second one is the familiar Poisson divergence proposed by Kimura and Ohta (1971) and Dickerson (1971), which takes into account multiple substitutions at the same site. The third one is the Tajima-Nei (TN) divergence (Tajima and Nei 1984) that also takes into account that two non-homologous amino acids at an aligned position may converge

by chance with probability $S_0 = \sum_a (f_a)^2$, where f_a the frequency of amino acid a in a large sequence database. The TN divergence cannot be computed if the sequence identity is smaller than $S_0 = 0.06$, and it is unreliable if this threshold is approached, thus we omitted pairs with SI $< 2S_0 = 0.12$.

Structure divergences

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 4.5Å 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

$$q(A, B) = \frac{\sum_{ij} C^{(A)}_{ij} C^{(B)}_{a(i)a(j)}}{\sqrt{\sum_{ij} C^{(A)}_{ij} \sum_{ij} C^{(B)}_{ij}}} \tag{4}$$

where $a(i)$ represents the residue of protein B aligned to residue i in protein A. q takes values between zero (no shared contacts) and one (perfect identity). Note that the computation of the overlap does not require structure superimposition.

From the contact overlap we obtain the contact divergence measure $\mathrm{D_{cont}}$ (Pascual-Garcia et al. 2010), which estimates the evolutionary relatedness in analogy to the TN sequence divergence:

$$\mathrm{D_{cont}}(A, B) = -\log\left(\frac{q(A, B) - q_\infty(L)}{1 - q_\infty(L)}\right) \quad \text{if } q > \epsilon(L)\,. \tag{5}$$

The parameter $q_\infty(L)$ denotes the limit overlap of distantly diverged protein pairs (here L is the length of the shorter protein), modelled as $q_\infty(L) = \overline{q}(L) + A\sigma_q(L)$, with $\overline{q}(L) = 0.386L^{-0.547}$ (mean of unrelated proteins), $\sigma_q(L) = 1.327L^{-0.673}$ (standard deviation of unrelated proteins) and $A = 5$. We do not consider here pairs with $q < \epsilon(L)$. For these pairs, $\mathrm{D_{cont}}$ is based on the Z-score of the overlap with respect to pairs of unrelated proteins. The cross-over value $\epsilon(L)$ is fixed by imposing that the contact divergence is continuous for $q = \epsilon(L)$. For further details on the determination of parameters see (Pascual-Garcia et al. 2010).

TM score

We considered another structural divergence measure based on the TM-score (Zhang and Skolnick, 2004), which measures the structural similarity between two aligned and superimposed proteins as:

$$\mathrm{TM} = \max\left(\frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right), \tag{6}$$

where L is the aligned protein length, d_i is the distance between the i-th 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 score computed in this way has an approximately constant value TM≈ 0.17 independent of protein length for unrelated pairs, and it equals one for identical structures. 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,

$$\mathrm{D_{cont}}(A, B) = \min_{\{a\in A, b\in B\}}\mathrm{D_{cont}}(a, b) \tag{7}$$

In the main text we present only results based on sequence alignment, but to investigate the influence of the alignment method we studied divergence measures based either on multiple sequence alignments ($\mathrm{D^{seq}}$) or on the multiple structure alignment ($\mathrm{D^{str}}$). We found that structure alignments are sensitive to conformation changes and are less reliable at reflecting evolutionary relationships, which justifies our choice.

Protein data

We selected from the CATH database (Orengo et al. 1997) 161, 150 and 397 structural domains of the NADP, Ploop and Globin superfamilies, respectively. They include proteins from both eukaryotic and prokaryotic organisms.

The domains were selected through the hierarchical clustering algorithm Complete Linkage (CL) algorithm, which defines 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 then selected the largest cluster for each superfamily.

Subsequently, we grouped all structures with the same sequence, and we computed the sequence divergence and $\mathrm{D_{cont}}$ of each pair of sequences as the minimum across all corresponding pairs of structures, Eq.(7), to minimize the chances that the proteins are in different conformations.

Identification of outgroups

For every multiple sequence alignment we inferred a neighbor-joining (NJ) tree (Saitou and Nei 1987) implemented in the program SplitsTree (Huson 1998) with the D_p divergence. The inferred trees were rooted assuming as the root a node that splits the different families of genes involved in the superfamily (Supplementary Figures S1-S3). Next, for every pair of proteins we selected those outgroups observed in the inferred tree with bootstrap higher than 0.7.

Violations of the molecular clock

For any pair of proteins A and B, we consider all possible outgroups C that fulfill the conditions reported above, and we compute the violations of the molecular clock CV as the difference of the divergences of proteins A and B with respect to C, averaged over the n outgroups and suitably rescaled

$$\mathrm{CV}_\alpha(A,B) = \frac{\frac{1}{n}\sum_C \left(D(A,C) - D(B,C)\right)}{D(A,B)^\alpha}\,, \tag{8}$$

Here $D(A,B)$ represents a divergence either in sequence or in structure, and $D(A,B)^\alpha$ estimates the fluctuations of the divergence measure expected under molecular clock. We tested several values of α and determined its optimal value through the simulations described below. For the case of the contact divergence and complementing the condition imposed by Eq. 7, if the outgroup C has several associated conformations c, we compute the violation of the molecular clock of structure evolution using as outgroup the structure c that minimizes $|D(a,c) - D(b,c)|$, where a and b are the representative structures of the pair AB.

Triangle inequality

If the divergence measure represents the number of substitutions or the time that separates two sequences, it must fulfil the triangle inequality (TI) of metric spaces, which states that no intermediate point B can make the walk from A to C shorter than the direct path:

$$\mathrm{D}(A,C) \leq \mathrm{D}(A,B) + \mathrm{D}(B,C) \tag{9}$$

However, even if the D measures are often called distances, they are not distances in the mathematical sense since they can violate TI (Felsenstein 2004). In particular, the fraction of different amino acids, $\mathrm{D_p}$, can be shown to satisfy TI if a multiple sequence alignment is used and all positions 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, the TI may be violated due to indels. $\mathrm{D_{Poiss}}$ and $\mathrm{D_{TN}}$ may violate the TI even in the absence of indels, because they are non-linear functions of the Hamming

distance. Similarly, the reciprocal of the contact overlap is a distance for multiple structure alignments with the same length, but D_{cont} is not.

Triples that violate the TI represent instances in which the divergence measures do not reliably estimate the evolutionary divergences among three proteins and they lead to overestimate the violations of the molecular clock. In fact, if the triple ABC violates TI, the absolute value of CV in Eq.(8) is larger than one for $\alpha = 1$. Therefore, we excluded all triples that violate the TI from the assessment of the molecular clock.

Significance

We require at least 3 outgroups for each pair in order to estimate the significance of CV as the ratio between the mean and the standard error of the mean (S.E.M.) of CV over the set of allowed outgroups,

$$t = \mathrm{CV}/\mathrm{SEM}\,. \tag{10}$$

As a conservative estimate, we consider significant the CV with $|t| > 3$, which is larger than the 5 percent threshold of the t-distribution with the minimum number of degrees of freedom that we considered, corresponding to 3 outgroups.

Here $\mathrm{SEM} = \mathrm{SD}/\sqrt{N_{\mathrm{indep}}}$ is the standard deviation divided by the square root of the number of independent outgroups, which is difficult to estimate since outgroups are evolutionarily correlated. We address this problem by considering as independent only the fraction of the outgroup that is different in sequence from the outgroups that have been already taken into account. Denoting by SI the sequence identity, we define

$$N_{\mathrm{indep}} = \sum_c \left(1 - \max_{c\prime<c}\mathrm{SI}(c, c\prime)\right)\,. \tag{11}$$

Test of the CV formula through simulations

In order to test the mathematical formulas that quantify CV, we simulated protein sequence evolution with selection on protein folding stability using the program ProteinEvolver (Arenas et al. 2013). First, we designed a phylogenetic tree with four outgroup proteins under four different modes for the divergence of the innermost proteins A and B from their common ancestor O: (1) A molecular clock in which the number of mutations from O to A and O to B follow a Poisson distribution with the same mean number of mutations; (2) Violation of the molecular clock where the average number of mutations from O to A is 20% (2a), 50% (2b) or 100% (2c) larger than the number of mutations from O to B. The four trees were simulated both for short evolutionary divergence between A and B ($\mathrm{SI}(A, B) = 0.83$ on the average) and for long evolutionary divergence ($\mathrm{SI}(A, B) = 0.26$ on the average), with all branches rescaled proportionally. For each of these eight trees, we performed 200 realizations of sequence evolution, obtaining a total of 1,600 multiple sequence alignment (MSA) where each sequence corresponds to a tip node of the tree, upon which the CV tests were applied.

We simulated the evolution of the protein kinesin (Uniprot code: KAR3_YEAST, sequence length: 346 amino acids, PDB code: 3KAR), which we studied previously (Arenas et al. 2013). The program ProteinEvolver places the PDB sequence in the root node of the phylogenetic tree and evolves it forward in time, proposing mutations under a given substitution model of evolution (Yang 2006; Arenas 2012) and accepting them if they reduce the estimated protein folding stability not more than 5% with respect to the sequence in the PDB, a criterion that is robust for different protein families and outperforms empirical substitution models because it accounts for stability against both unfolding and misfolding. Stability is estimated with the contact interaction matrix derived in (Bastolla et al. 2000). The other thermodynamic parameters (temperature and configuration entropies for unfolded and misfolded states) are as previously evaluated (Arenas, et al. 2013; Arenas, et al. 2015).

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 function annotation and we only considered GO terms that were manually assigned, i.e. 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). Note that InterPro signatures do not necessarily yield a classification, but we verified that they do in the case of Globins, i.e. having the same InterPro signature is an equivalence relationship for globins. To retrieve these signatures, we used the SSMap tool (David et al. 2008) that relates PDB chains with UniProt accessions, including InterPro signatures. We refer to GO and InterPro terms as function annotation (FA).

Results

Assessment of the molecular clock through simulations

The molecular clock hypothesis assumes that two homologous proteins A and B evolve at the same rate, so that their divergence from any outgroup protein C is equal apart from stochastic fluctuations. We assess clock violations CV for a pair of proteins A and B measuring the difference between $D(A, C)$ and $D(B, C)$, averaged over outgroups and

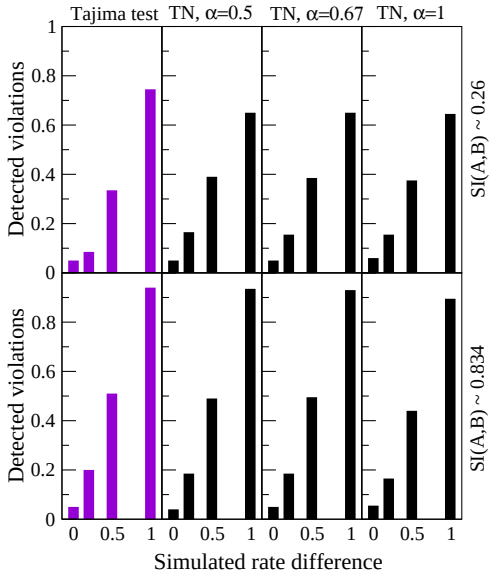

Figure 1: Fraction of violations of the molecular clock in simulated protein evolution detected using the Tajima test (leftmost) and the CV test with different sequence divergence measures and different α parameters. The horizontal axis shows the simulated clock violation (0, 20%, 50% and 100% difference in the simulated rate between lineages A and B). Top: long evolutionary divergence ($SI(A,B) = 0.26$ on the average). Bottom: short evolutionary divergence ($SI(A,B) = 0.83$ on the average).

normalized by the expected difference under a stochastic molecular clock that is estimated as $D(A,B)^{\alpha}$, Eq.(8) in Materials and Methods. Since the two lineages evolve at the same rate until their last common ancestor, a significant CV means that the evolutionary rate of at least one of them changes after their split. For this reason we also call CV "evolutionary accelerations".

For a Poissonian clock, the variance and the mean are equal, so that the standard deviation of $D(A,B)$ scales as $\sqrt{D(A,B)}$ and the value of the exponent is $\alpha = 0.5$. The first method for testing the molecular clock (Fitch 1976) was based on this scaling. Nevertheless, it has been observed both in empirical data (Gillespie 1991) and in simulations of protein evolution with stability constraints (Bastolla et al. 1999) that the number of substitutions is overdispersed, i.e. its variance is significantly larger than the mean, in which case the exponent $\alpha > 0.5$ holds. A simple explanation of overdispersion is that mutations are accepted with higher probability in regions of sequence space where the protein stability is higher, leading to auto-correlated substitution rates. Because of this reason we consider $\alpha > 0.5$, up to $\alpha = 1$, which holds if the evolutionary rates of the two lineages are different.

We discard triples of proteins that violate the triangle inequality (TI) $D(A,C) \leq D(A,B)+D(B,C)$. In fact, the violation of TI would imply that the evolutionary distance from a protein A to a protein C becomes shorter passing through an intermediate protein

B, which is not possible, so it indicates that the divergences do not reliably estimate the evolutionary distances and they overestimate the violation of the molecular clock.

Our test of the molecular clock enforces two criteria: first, the value of CV must be above a threshold; second, the mean CV obtained from different outgroups must be significantly above zero, i.e. the parameter t in Eq.(10) must satisfy $|t| > 3$.

To evaluate the CV test, we simulated protein sequence evolution with selection on protein folding stability using the program ProteinEvolver (Arenas et al. 2013) as described in Methods. We considered one scenario in which the molecular clock holds and three ones in which it is violated by 20, 50 and 100 percent. Each scenario was repeated for short and long branches. For each of the eight scenarios we sampled 200 MSA and we applied the CV tests, measuring the frequency of MSAs that violate the molecular clock with $|t| > 3$ and with a threshold on CV chosen such that the false positive rate is 0.05 for the scenario that obeys the molecular clock. We found that the condition $|t| > 3$ considerably reduces this threshold, enhancing the ability to detect significant violations of the molecular clock. Fig.1 reports the fraction of significant violations for the eight scenarios.

For the same data sets, we also applied the Tajima test of the molecular clock (Tajima, 1993), averaging the Tajima parameter over all of the outgroups, which enhances the ability of the test to detect significant clock violations. We confirmed that the Tajima test is superior to our CV test for all values of α, consistent with recent analysis (Battistuzzi et al. 2011). However, the advantage of the Tajima test is not substantial (see Fig.1), whereas the test is more complicated than ours since it involves performing computations on triples of proteins. In particular, for detecting clock violations in protein structure evolution, the Tajima test should be modified to consider contacts or superimpositions between residues instead of amino acid identities. More importantly, the software to evaluate structural divergence between pairs of proteins should be modified in order to be applied to triples of proteins. Therefore, in the following we apply the CV test, which is almost as powerful as the Tajima test and is more easily applicable to structural variation.

For each exponent α, we measured two threshold values of CV such that the false positive rate is 5% for small evolutionary distances and for large evolutionary distances. The two threshold values cross at an exponent α such that the threshold is almost independent of the divergence $D(A, B)$. This value is $\alpha = 0.67$ when the divergence is the Tajima-Nei divergence (TN) and $\alpha = 0.55$ when the divergence is the Poisson divergence (see Supplementary Fig. S4). The fact that the optimal α is larger than 0.5 confirms that the number of substitutions is overdispersed under selection for protein folding stability. When not otherwise stated, we use the TN divergence with $\alpha = 0.67$ and $CV = 0.14$ as threshold. The same exponent and threshold are used for structure divergences, which we cannot simulate, because of their mathematical analogy with the TN divergence.

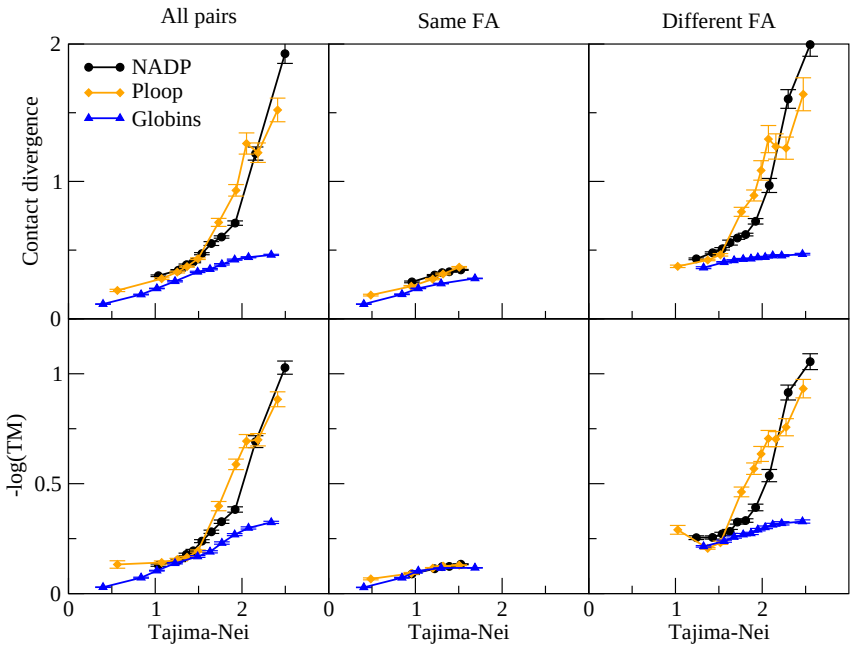

Figure 2: Structural divergence (top, Contact Divergence; bottom, $-\log(\mathrm{TM}-\mathrm{score})$) versus sequence divergence (Tajima-Nei) for three superfamilies. Left, all pairs of proteins. Center, only pairs with the same functional annotation (FA). Right, only pairs with different FA.

Structure evolution is constrained by function conservation

We now turn to analyse the molecular clock in the evolution of real protein superfamilies. We consider three large superfamilies, NADP, Ploop and globins.

We first compared divergences estimated from sequence and from structure alignments. As expected, sequence identities obtained from structural alignments are always smaller than those obtained from sequence alignments (see Supplementary Figure S5). This can be interpreted either as an overfitting of the sequence alignments that match residues at the expense of a poor structural match, or as an artefact of structure alignments that place spurious gaps in order to accomodate structural rearrangements due to conformation changes. To weight the second effect, we examined pairs having SI = 1 according to sequence alignments, which correspond to conformation changes. They present structure alignments with SI as low as 0.77 (Ploop) or 0.92 (NADP), which indicates that conformation changes can substantially affect the structure alignment, speaking against the use of structure alignments for evolutionary studies. Therefore, we analyse violations of the molecular clock adopting sequence alignments. On the other hand, for $\mathrm{D_{cont}} < 1$ the $\mathrm{D_{cont}}$ measures obtained through sequence alignments are similar to those obtained with structure alignments and there is not any clear bias, while for high divergence sequence alignments heavily overestimate $\mathrm{D_{cont}}$, evidencing the poor quality of the structure alignments derived from distant sequence alignments, a recognized problem in homology

Superfamily	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
Ploop	53, 150	1343	0.24 ± 0.03	0.20 ± 0.02	$0.12^{f} \pm 0.03$	0.065 ± 0.006
Globins	71, 397	2424	0.22 ± 0.01	0.15 ± 0.01	0.13 ± 0.01	0.073 ± 0.017

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

modelling (Abagyan and Batalov, 1997).

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 report them here as well since they are useful for our discussion. The sequence divergence that correlates strongest with structure divergence measures is the TN divergence, and we shall use it throughout the paper if not otherwise stated. Qualitatively similar results were obtained with the other two studied sequence divergence measures.

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 Fig.2 are smaller than one, ranging from 0.22 (globins) to 0.25 (NADP) for the contact divergence and even smaller, from 0.12 (Ploop) to 0.16 (NADP) for the TM divergence, see Table 1. Since the three divergence measures are computed in the same way from frequencies of sequence identity, 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 strongly limited for protein pairs with the same function annotation (FA) than with different FA, as previously reported (Pascual-Garcia et al., 2010). On the average, when the FA varies, the 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 the structure divergence versus the sequence divergence is smaller than when we consider all pairs (compare columns b-c and d-e of Table 1). This suggests that the rate of change of sequence with structure stems from selective

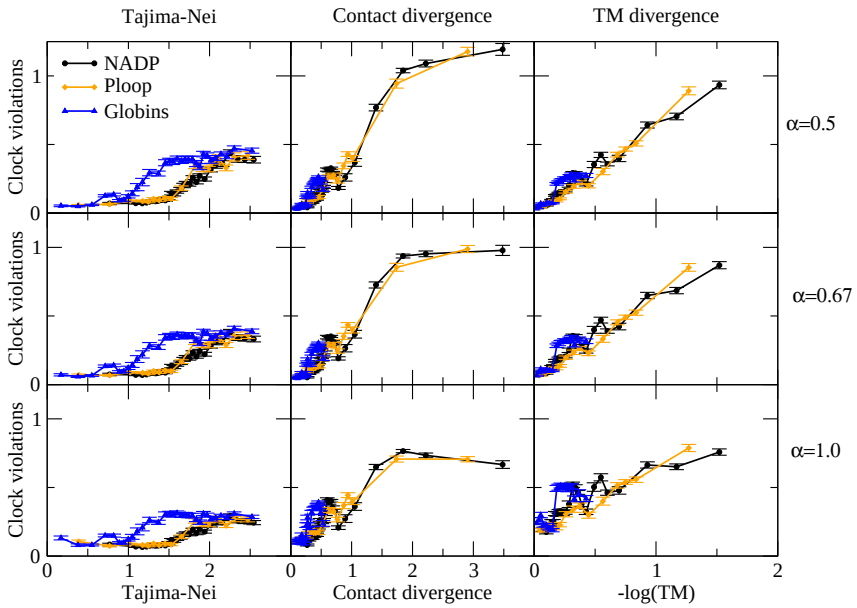

Figure 3: CV versus evolutionary divergence for three divergence measures (left, TN sequence divergence; center, CD; right, TM score divergence) and three exponents (top $\alpha = 0.5$; center $\alpha = 0.67$; bottom $\alpha = 1$).

constraints on protein function 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 (CV) in the evolution of the three studied superfamilies. We start showing in Fig.3 CV versus the corresponding divergence measure for three values of the exponent α. One can see that CV tends to increase with the divergence, i.e. lineages that are evolutionarily close tend to evolve with the same rate while distant lineages tend to evolve with markedly different rates. For the CD measure, CV reaches a plateau at large divergence for $\alpha = 0.67$, while it tends to increase for $\alpha = 0.50$ and to decrease for $\alpha = 1$, supporting $\alpha = 0.67$ as the optimal scaling for CD. Thus, when not otherwise stated, we assess CV through Eq.(8) with exponent $\alpha = 0.67$, which also turned out to be almost independent of the TN divergence with simulated data. We verified that all results are robust with respect to variations of α between 0.5 and 1

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. CV will be large and significant only if most outgroups consistently predict a CV with the same sign. CV measures are also consistent between different divergence measures. Despite the correlation between sequence and structure divergence can be as low as 0.60 (between TN and CD for the Ploop

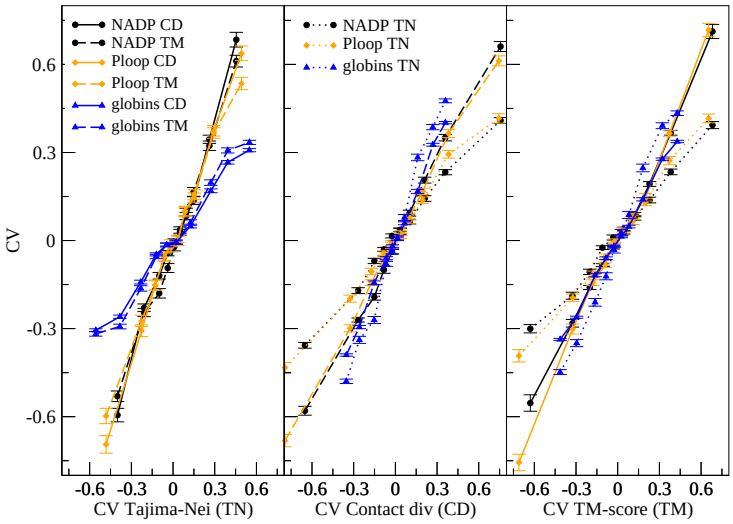

Figure 4: Relationships between average CV obtained with different measures. In all cases $\alpha = 0.67$ was considered, although other exponents do not change the results qualitatively.

superfamily, see Supplementary Fig.S6 top row), the CV measures based on sequence and structure divergence are strongly correlated, with correlation coefficient larger than 0.85 (Supplementary Fig.S6). The strong relationship between CV obtained with different divergences is represented in Fig.4, which shows that they tend to have the same sign and that the two structural measures of CV (CD and TM) are almost equal. 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 random errors in the estimate of the evolutionary divergences.

The molecular clock is more violated in structure evolution than in sequence evolution

Because of the normalization, the definition of CV depends little on the scale of the evolutionary divergence, therefore we can meaningfully compare its value for sequence (TN) and structure (CD and TM) divergences. For the NADP and Ploop superfamilies the absolute value of CV tends to be larger in structure than in sequence, while the opposite holds for the Globin superfamily (see Fig.4). The same conclusion can be derived from Fig.3. This discrepancy between the Globin superfamily and the rest disappears when we consider the significance of the CV. The fraction of significant pairs is shown in Fig.5 top row as a function of the corresponding divergence measure. One can see that it is much larger for the TM divergence than for the CD and the smallest for the TN sequence divergence, for all three superfamilies. The other sequence divergence measures are reported in Supplementary Figure S7. They presented smaller and less significant

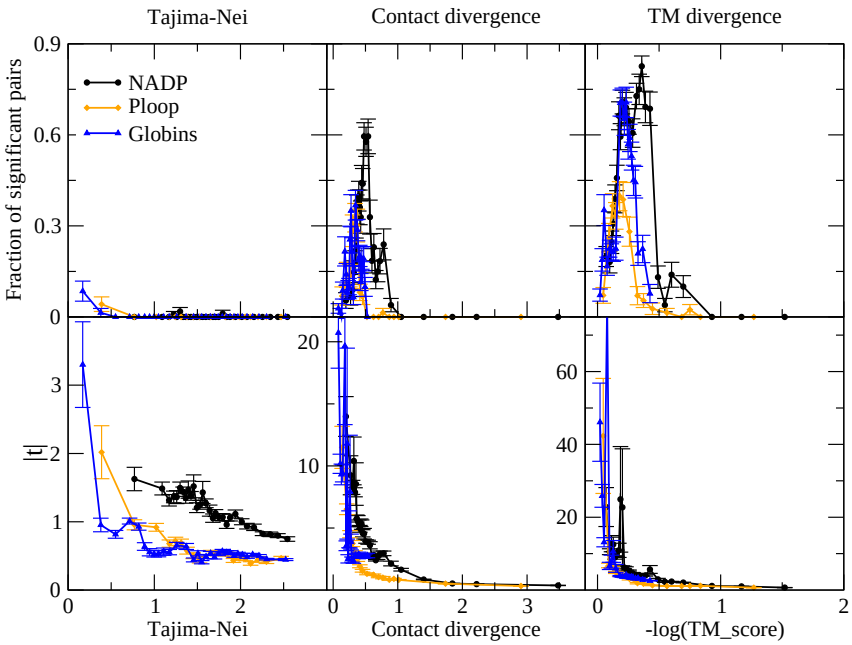

Figure 5: Significance of CV (top, fraction of significant pairs) versus evolutionary divergence for three divergence measures (left, TN sequence divergence; center, contact divergence; right, TM score divergence) and the exponent $\alpha = 0.67$.

clock violations than TN, and they confirm the result that CV is stronger and more significant in structure evolution.

Interestingly, while for sequence divergence significant pairs are only observed for the smallest divergences, for structure divergence they peak at intermediate values (from 0.2 to 0.4 for TM divergence and from 0.3 to 0.6 for contact divergence). The large fraction of significant pairs that we observe for small sequence divergence 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, structure evolution does not present increased CV for small evolutionary divergence (Fig.5), which we interpret as an indication of stronger purifying selection at the structural level (see below).

The average value of the CV significance parameter $|t|$ is also larger for structure divergences (Fig.5 bottom line), and it tends to decrease with the divergence, both in sequence and in structure. This result is not surprising. A significant and positive CV between protein A and protein B means that protein A tends to evolve faster than protein B over all of their evolutionary divergence. When the divergence is very long, however, it is reasonable that there will be periods in which the faster protein is A and periods in which it is B, leading to non significant CV.

The statistical properties of CV in sequence and in structure evolution are summarized in Fig.6. As already observed, the average of the absolute value of CV (top left) is larger for structure than for sequence evolution, except for globins. All superfamilies,

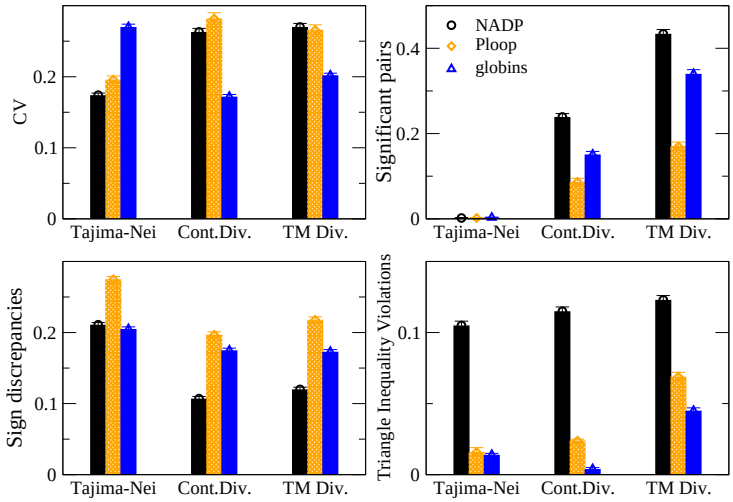

Figure 6: Comparison between CV in sequence evolution (Tajima-Nei) and in structure (Contact divergence and TM divergence) for the three superfamilies. Top left: average of the absolute value of CV. Top right: Fraction of significant pairs. Bottom left: Average fraction of outgroups that show sign discrepancies for any given pair. Bottom right: average fraction of violations of the triangle inequality TI (triples violating TI are omitted from the computations).

including globins, present a larger fraction of significant CV for structure than for sequence evolution, in particular it is larger for TM divergence than for contact divergence (top right). The average fraction of outgroups that show sign discrepancies in CV for any given pair is smaller for structure evolution than for sequence evolution for all superfamilies (bottom left), indicating that the sign of CV is more consistent in structure evolution. Nevertheless, the strong correlations observed in Fig.4 indicate that the dominant sign tends to be the same in both cases. These results concur in indicating that structure evolution is more episodic than sequence evolution.

Finally, the average fraction of triples that violate the triangle inequality (TI) is almost the same between TN divergence and contact divergence but it is larger for the TM divergence, suggesting that the TM score is less reliably estimated (we remind that we limit the influence of this factor by omitting triangles that violate TI and evidence inconsistencies in the estimated divergences).

Changes of protein function enhance CV

Finally, we investigated whether the 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 with different FA. The results are plotted in Fig.7. For all divergence measures and all superfamilies, the average value of CV is systematically

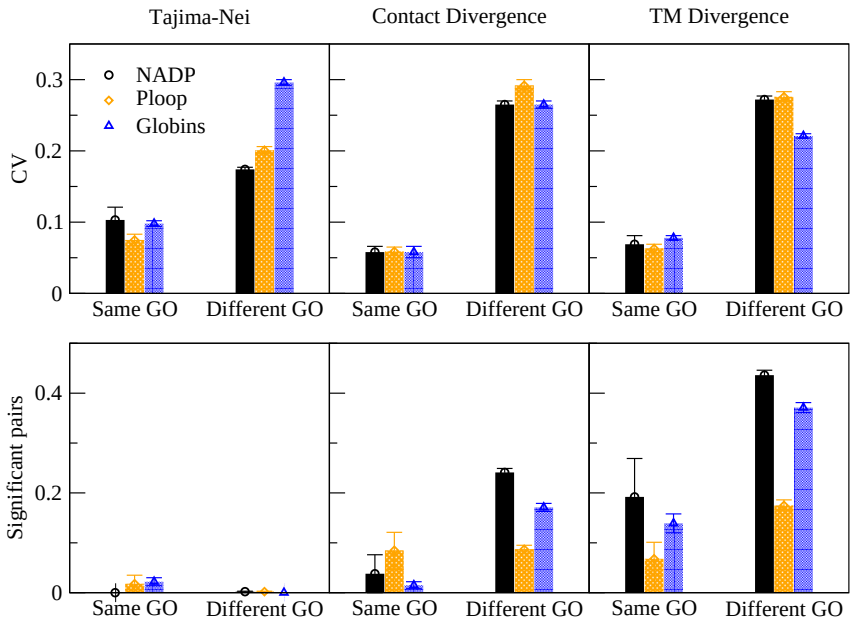

Figure 7: Violations of the molecular clock of sequence (left) and structure evolution (center and right) distinguishing protein pairs with the same and different FA. Top: Average value of the CV parameter. Bottom: Fraction of significant pairs. One can see that violations of the molecular clock are larger for pairs with different FA.

larger for pairs with different FA than for pairs with similar FA.

The same result holds for the fraction of pairs with significant CV in structure evolution, with the exception of the Ploop superfamily with the CD divergence, which shows no significant difference between same and different FA. In contrast, for the sequence divergence measure (TN) the few pairs that significantly violate the molecular clock have more frequently the same FA than different FA for both Ploop and Globin superfamilies while no difference is seen for NADP. This result is consistent with the observation that, for the TN divergence, significant CV are only found in closely related proteins.

Despite globally different FA induce stronger and more significant violations of the molecular clock, the values of CV are large also for pairs with the same FA, especially for sequence evolution.

Discussion

In this work, we tested the molecular clock hypothesis in the evolution of protein sequences and structures. An important property of our CV test with respect to other traditional tests of the molecular clock such as those by Fitch (1976) and Tajima (1993) is that all suitable outgroups are used. This increases the power of the CV test. A pair of proteins significantly violates the molecular clock only if the different outgroups yield consistent

estimates, as assessed through a t-test, whereas the violation is not significant if CV changes frequently sign for different outgroups.

Another characteristic of our method is that we do not assume that the evolutionary divergence follows a Poisson distribution, but we parameterize their expected fluctuations as $D(A,B)^{\alpha}$. Simulating protein evolution with stability constraints, we found $\alpha = 0.67$ as the optimal exponent for the TN divergence, confirming that the evolutionary divergence is more dispersed than a Poisson process for which $\alpha = 0.5$ (Gillespie 1991), as previously observed in simulations (Bastolla et al. 1999). Results for real superfamilies and CD divergence, which cannot be simulated, also suggest $\alpha = 0.67$ as the optimal exponent. Qualitative results are robust in the range $\alpha \in [0.5, 1]$.

We considered three large superfamilies with rather distinct properties. Ploop 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 3,602 and 1,285 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, apparently because of their lower functional diversity (compare Fig.2 center and right) and they are present in only 22 FunFam clusters and 6,224 species. Despite these differences, they yielded similar results under the point of view of the molecular clock. Adopting the CV test, we obtained the following main results:

1. For a large fraction of pairs, violations of the molecular clock in structure evolution are not due to random fluctuations of evolutionary divergence measures but to the systematic enhancement of the evolutionary rate of one protein with respect of the other one. This conclusion is supported by the fact that more than 80 percent of the outgroups for sequence divergence and more than 85 percent for structure divergence identify the same protein of the pair as the one that evolves more rapidly and the significance parameter is $|t| > 3$ for a large fraction of pairs.

2. Clock violations in sequence and structure are strongly correlated (correlation coefficients > 0.86 for all three superfamilies), showing that, when one protein evolves more rapidly than the other one in structure, it also tends to evolve more rapidly in sequence. This suggests that the mechanisms that produce an increase in the evolutionary rate, either selection or mutation or both, act similarly on sequence and structure.

3. Clock violations both in structure and in sequence tend to increase with the evolutionary divergence, even if they are normalized through a power of the divergence. In other words, two proteins with large evolutionary divergence tend to differ substantially in their evolutionary rates, in particular for structure divergence. However, the fraction of pairs with significant CV in structure evolution peaks at intermediate values of the divergence, since for large divergence CV are mostly not significant. This is not surprising: significant CV between two proteins is observed when one

protein evolves more rapidly than the other one during all of their evolutionary divergence, but if the protein that evolves more rapidly changes throughout evolution the CV is not significant for long divergence.

4. Despite their strong correlation, CV in sequence and structure present different trends. For structure evolution CV is not significant at small divergence and becomes significant at intermediate divergence, whereas in sequence evolution CV is only significant at small divergence. These observations can be interpreted as a consequence of relaxed selection at short evolutionary separation for protein sequence but not for protein structure evolution (see below).

5. Protein structure evolution is less clock-like than protein sequence evolution: CV tends to be larger and more significant. While CVs are more significant in structure than in sequence for all superfamilies, they tend to be larger in sequence than in structure for the Globin superfamily. This superfamily is characterized by a remarkable structural and functional conservation even in case of change of InterPro term, which may explain why their CVs are smaller.

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

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; van Ham et al. 2003) 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 an evidence of positive selection (e.g., McDonald and Kreitman 1991, Massingham and Goldman 2005; Kosakovsy 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).

In the present work, the observed differences between sequence evolution and structure evolution cannot be attributed to mutational or demographic processes, which should act in a similar way on protein sequence and structure, and may provide interesting insight on the selective forces that mould proteins.

The observation that the molecular clock in sequence evolution is strongly and significantly violated at small sequence separation is consistent with the previous observation that the substitution rate is larger for pairs of species separated by short time intervals (Ho et al., 2005), which has been attributed to segregating neutral and slightly deleterious popymorphisms after a speciation event (Peterson and Masel 2009). This trend is not observed in structure evolution, consistent with the view that protein structure is more purged by negative selection than protein sequence. We interpret these enhanced substitutions as variations that are neutral or mildly deleterious in sequence, probably

at the level of protein stability, but do not significantly modify the protein structure. It would be interesting to test this hypothesis either experimentally or computationally.

On the other hand, when the evolutionary divergence is not small, clock violations are larger for protein structure than for protein sequence evolution. We propose that positive selection is a natural explanation for this difference, although our data is not conclusive. In fact, protein structure is under stronger 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), in particular 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 protein structures rather than to relaxed negative selection. This is consistent with our observation that the CV is not significant during long evolutionary periods of time, i.e. for pairs of distantly related proteins the faster evolving protein changes over time, consistent with the expectation that positive selection is episodic.

Note that the TM measure presents a slower rate of divergence with respect to sequence divergence and a larger CV than the CD measure. These observations suggest that the TM measure, which responds to continuous changes of the atomic coordinates, is subjected to stronger selection than the coarser CD measure, which responds to changes of binary contacts. 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 clear conclusions.

If significant CV in protein structure evolution is attributable to positive selection, it is natural to expect that these CVs are systematically associated to function change, since function change is expected to be favoured by positive selection that improves the new protein functionality 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 function annotation (FA) as indicated by their manually annotated GO and InterPro terms, and protein pairs with different FA. We found that the latter present larger and more significant CV in structure evolution, as expected (Fig.7). In contrast, in sequence evolution pairs with different FA are associated with larger but not more significant CV. 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.

In summary, our results support the view that natural selection acts more on protein structure than on protein sequence, both in the form of positive selection that enhances structure divergence when the function varies (Fig.7) and in the form of negative selection that constraints the structure more than the sequence, in particular when the function is

conserved (Fig.2). Consistent with this view, the enhancement of the substitution rate at short evolutionary divergence (Ho et al., 2005) that has been attributed to neutral and slightly deleterious variation (Peterson and Masel 2009) is observed in protein sequence but not protein structure evolution (Fig.3). Thus, under functional conservation sequence mutations are preferentially fixed when they conserve the structure, although they may change other properties such as stability, provided this is maintained above a sufficient limit. On the contrary, under changes of function the structure experiences larger changes than the sequence.

Perspectives

In addition to the influence of function changes on CV, we also observed significant violations of the molecular clock of structure evolution when the FA is conserved (Fig.7). 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. In this view, the function embodied in the GO and InterPro terms is only one determinant of the molecular environment of the protein, which is constituted by all the molecules that interact with it.

This influence of coevolution on evolutionary rates is at the basis of the methods proposed by Valencia, Pazos and coworkers for inferring protein-protein interactions (Pazos et al., 1997; 2008; Ochoa et al. 2015). Their MirrorTree method and its sequels are based on assuming that the normalized substitution rates of interacting proteins are correlated (Ochoa et al. 2015). The success of these methods suggests the substitution rate of a protein changes in response to evolutionary changes of its interaction partners, and it would be very interesting to quantitatively assess this influence both in sequence and in structure evolution.

The interpretation that selection targets protein strcture 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 2017) selection targets protein stability, and the effect of mutations on stability is estimated through empirical free energy funtions under the assumption that the structure is approximately conserved. In structure constrained models (Echave 2008) selection targets protein structure, and the structural effect of mutations is estimated through the elastic network model (ENM, 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 support the importance to consider 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 (Jimenez, Arenas and Bastolla, submitted), which studied the site-specific substitution rates predicted through stability constrained models, finding that they are too tolerant at amphiphilic sites that can be occupied by polar of hydrophobic amino acids under

stability requirements, probably because they neglect that mutations at these sites can modify the protein structure, while the rates predicted through structurally constrained protein evolution models are in better agreement with observed data (Huang et al. 2014).

Stability constrained models assume a sigmoidal relationship between folding free energy ΔG and fitness, $f \approx 1/(1+\exp(\Delta G))$, thus when ΔG is strongly negative even relatively large changes $\Delta\Delta G$ only produce moderate to small changes in fitness, implying that stability evolution is almost neutral and clock-like, at least as far as the native structure does not change. However, even small changes in the native structure can modify the native dynamics, which can be predicted based only on protein function through the ENM, and which is a target of natural selection (Haliloglu and Bahar 2015; Dos Santos et al. 2013), having significant fitness consequences.

In conclusion, our results support the following view of protein sequence and structure evolution. In a constant molecular environment the 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 by the strong relationship between native structure, dynamics in the native state and function (Haliloglu and Bahar 2015) among other observations. Conversely, stability is a more neutral character (Goldstein 2011; Serohijos and Shakhnovich 2014) so that neutral and slightly deleterious alleles that decrease stability but conserve structure are often fixed. In contrast, in a changing molecular environment, either due to coevolution or due to function change, positive selection strongly acts on protein structure and enhances its evolutionary rate. Structure evolution drives accelerated sequence evolution as well, both because sequence changes are necessary for structure changes and because structure changes tend to be accompanied by decrease in stability that drive compensatory mutations at the sequence level (Tokuriki and Tawfik 2009).

Acknowledgements

This work has been financed by the Spanish Ministry of Economy, grants BIO2016-79043 and BFU2012-40020. MA was supported by the Ramón y Cajal Grant RYC-2015-18241 from the Spanish Government. We thank the Editor and three Reviewers for their insightful comments that helped us to improve the paper.

References

[1] Abagyan RA, Batalov S. 1997. Do aligned sequences share the same fold? J Mol Biol. 273:355-68.

[2] Arenas M. 2012. Simulation of Molecular Data under Diverse Evolutionary Scenarios. PLoS Comput Biol 8:e1002495.

[3] Arenas M, Dos Santos HG, Posada D, Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics 29:3020-3028.

[4] Arenas, M., Lopes, J.S., Beaumont, M.A., Posada, D., 2015. CodABC: A Computational Framework to Coestimate Recombination, Substitution, and Molecular Adaptation Rates by Approximate Bayesian Computation. Mol Biol Evol 32, 1109-1112.

[5] Arenas M, Sanchez-Cobos A, Bastolla U. 2015. Maximum likelihood phylogenetic inference with selection on protein folding stability. Molecular Biology and Evolution 32:2195-2207.

[6] Ayala, F.J. 1997. Vagaries of the molecular clock *Proc. Natl. Acad. Sci. USA* **94**: 7776-7783.

[7] Ayala, F.J. 1999. Molecular clock mirages. BioEssays 21:7175.

[8] Bastolla U, Roman HE, Vendruscolo M. 1999. Neutral evolution of model proteins: Diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**: 49-64.

[9] Bastolla U, Vendruscolo M, Knapp EW. 2000. A statistical mechanical method to optimize energy functions for protein folding Proc. Nat. Acad. Sci. USA 97: 3977-3981.

[10] 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.

[11] Battistuzzi, F.U., Filipski, A.J. and Kumar, S. 2011. Molecular Clock: Testing. eLS. John Wiley & Sons Ltd, Chichester.

[12] Bouckaert, R., Heled, J., Khnert, D., Vaughan, T., Wu, C-H., Xie, D., Suchard, MA., Rambaut, A. and Drummond, A. J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comp Biol 10: e1003537.

[13] Britten RJ. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**: 1393-1398.

[14] Bromham L, Penny D. 2003. The modern molecular clock. Nat Rev Genet. 4:216-24.

[15] Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. 2014. The influence of selection for protein stability on dN/dS estimations. Gen Biol Evol. 6:2956-67.

[16] Dos Santos HG, Klett J, Méndez R, Bastolla U. 2013. Characterizing conformation changes in proteins through the torsional elastic reponse. Biochim Biophys Acta 1834:836846.

[17] David FP, Yip YL. 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.

[18] Dickerson RE. 1971. The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26-45.

[19] Echave J. 2008. Evolutionary divergence of protein structure: the linearly forced elastic network model. Chem Phys Lett 457:413416.

[20] Felsenstein J. 2004. Inferring Phylogenies Sinauer Associates: Sunderland, MA.

[21] Fitch W. 1976. Molecular evolutionary clocks. In: Ayala FJ (ed.) Molecular Evolution, pp. 160178. Sunderland, MA: Sinauer Associates.

[22] Fitch WM, Leiter JME, Li X, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. Proc Nat Ac Sci USA 88: 4270-4274.

[23] Franks SJ, Weis AE. 2008. A change in climate causes rapid evolution of multiple life-history traits and their interactions in an annual plant. J Evol Biol. 21:1321-34.

[24] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet 2000;25:2529.

[25] Gillespie, J.H. (1989) Lineage effects and the index of dispersion of molecular evolution, *Mol. Biol. Evol.* **6**, 636-647.

[26] Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:13961407.

[27] Haliloglu T, Bahar I. 2015. Adaptability of protein structures to enable functional interactions and evolutionary implications. Curr Opin Struct Biol. 35:17-23.

[28] Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol. 22(7):15611568

[29] Huang TT, del Valle Marcos ML, Hwang JK, 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.

[30] Hunter S, Apweiler R, Attwood TK, Bairoch A, et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res 37 (Database Issue):D211D215.

bibitemHuson1998 Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14:68-73.

[31] 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.

[32] Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772-780.

[33] Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624-626.

[34] Kimura M, Ohta T. 1971. On the rate of molecular evolution. J Mol Evol 1: 1-17.

[35] Kimura M. 1983. The neutral theory of molecular evolution (Cambridge University Press).

[36] Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino Acid sites under selection. Mol Biol Evol 22:1208-1222.

[37] Kvikstad EM, 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.

[38] Lai J, Jin J, Kubelk J, Liberles DA. 2012. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. J Mol Biol. 422: 442459.

[39] Langley CH, Fitch WM. 1973. An estimation of the constancy of the rate of molecular evolution. J Mol Evol. 3: 161-177.

[40] Lupyan D, Leo-Macias A, Ortiz AR. 2005. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 21:3255-63.

[41] Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753-62.

[42] McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-4.

[43] Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA 93:287378

[44] Ochoa D, Juan D, Valencia A, Pazos F. 2015. Detection of significant protein coevolution. Bioinformatics. 2015 31:2166-73.

[45] Ohta T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism, *Theor. Pop. Biol.* **10**, 254-275.

[46] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:10931108.

[47] Pascual-Garcia A, Abia D, Méndez 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.

[48] Padhi A, Parcells MS. 2016. Positive selection drives rapid evolution of the meq oncogene of Mareks disease virus. PLoS ONE 11(9): e0162180.

[49] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 271:511-23.

[50] Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A. 2008. Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol Biol. 484:523-35.

[51] Peterson GI, Masel J. 2009. Quantitative Prediction of Molecular Clock and Ka/Ks at Short Timescales. Mol. Biol. Evol. 26:25952603.

[52] Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

[53] Serohijos AW, Shakhnovich EI. 2014. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr Opin Struct Biol 2014, 26:8491.

[54] Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucl Ac Res. 41(Database issue):D490-8.

[55] Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Gen 16, 224236.

[56] Tajima F, Nei M. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269-85.

[57] Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607.

[58] Tirion MM. 1996. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett. 77:1905-1908.

[59] Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol. 19:596-604.

[60] van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, et al. 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. USA 100:58186.

[61] Venkat A, Hahn MW, Thornton JW. 2017. Multinucleotide mutations cause false inferences of positive selection. Preprint at bioRxiv. doi: https://doi.org/10.1101/165969

[62] Yang Z. 2006. Computational Molecular Evolution. Oxford, England.: Oxford University Press.

[63] Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591.

[64] Zhai W, Nielsen R, Slatkin M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol. 26:273-83.

[65] Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57, 702-710.

[66] Zuckerkandl E, Pauling L. 1962. in *Horizons in Biochemistry*, eds. M. Kasha and B. Pullman (Academic Press, New York).