PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation

Alberto Pascual-García, David Abia, Raúl Méndez, Gonzalo S. Nido, and Ugo Bastolla*

Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), Cantoblanco, Madrid 28049, Spain

ABSTRACT

The molecular clock hypothesis, stating that protein sequences diverge in evolution by accumulating amino acid substitutions at an almost constant rate, played a major role in the development of molecular evolution and boosted quantitative theories of evolutionary change. These studies were extended to protein structures by the seminal paper by Chothia and Lesk, which established the approximate proportionality between structure and sequence divergence. Here we analyse how function influences the relationship between sequence and structure divergence, studying four large superfamilies of evolutionarily related proteins: globins, aldolases, P-loop and NADP-binding. We introduce the contact divergence, which is more consistent with sequence divergence than previously used structure divergence measures. Our main findings are: (1) Small structure and sequence divergences are proportional, consistent with the molecular clock. Approximate validity of the clock is also supported by the analysis of the clustering coefficient of structure similarity networks. (2) Functional constraints strongly limit the structure divergence of proteins performing the same function and may allow to identify incomplete or wrong functional annotations. (3) The rate of structure versus sequence divergence is larger for proteins performing different functions than for proteins performing the same function. We conjecture that this acceleration is due to positive selection for new functions. Accelerations in structure divergence are also suggested by the analysis of the clustering coefficient. (4) For low sequence identity, structural diversity explodes. We conjecture that this explosion is related to functional diversification. (5) Large indels are almost always associated with function changes.

Proteins 2009; 00:000–000.

Key words: protein structure evoluation; molecular clock; protein function; protein structure classification.

INTRODUCTION

The molecular clock hypothesis[1] played a fundamental role in the early days of molecular evolution studies after Zuckerkandl and Pauling recognized that protein sequences accumulate amino acid substitutions almost linearly in time, with a rate that varies with the protein family but is almost constant in different lineages.[2] The neutral theory, proposed almost simultaneously by Kimura[3] and King and Jukes,[4] interprets the constancy of the evolutionary rates as the result of neutral substitutions,[5,6] i.e., substitutions that have very little effect on fitness and are fixed in natural populations through random genetic drift instead of positive selection. This theory was subsequently generalized by Ohta to include nearly neutral substitutions for which either the selective effect or the effective population size is small.[7] The nearly neutral theory can be derived from standard population genetics models[8] and it is formally equivalent to equilibrium statistical mechanics, since molecular properties arise from a balance between mutational entropy in sequence space and fitness, where population size plays the role of inverse temperature.[9] In particular, protein folding stabilities in bacterial genomes are predicted to be smaller for bacteria with low effective population size.[10]

Though controversial in a first time,[11] the neutral and nearly neutral hypothesis had the great merit to give theoretical support to the molecular clock hypothesis, which is still of fundamental importance for methods that reconstruct evolutionary trees from molecular data.[12] Moreover, the neutral hypothesis also predicts that we should find violations of the molecular clock in the interesting cases when adaptive evolution takes place, for instance when new molecular functions emerge.

The quantitative study of the rate of protein structure evolution has received comparatively less attention. A milestone was the 1980 paper by Chothia and Lesk, who showed that the Root Mean Square Deviation (RMSD) between different globins diverges regularly with the number of amino acid substitutions,

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: Spanish Ministery of Science and Innovation (Ramón y Cajal fellowship), Grant numbers: BIO2008-04384, CSD200623.
*Correspondence to: Dr. Ugo Bastolla, Centro de Biologia Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, Madrid 28049, Spain. E-mail: ubastolla@cbm.uam.es.
Received 8 April 2009; Revised 9 September 2009; Accepted 10 September 2009
Published online 22 September 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22616

up to a limit of low sequence identity where the RMSD explodes.[13] Although this result suggests a generalization of the molecular clock hypothesis to the evolution of protein structure, it is limited by the fact that the RMSD can be used as a measure of structure divergence only for aligned residues that have a good spatial superimposition. We will propose here a measure of structure divergence based on evolutionary considerations, which is more suitable for such a quantification.

Together with the clock-like divergence of protein structures, the results by Chothia and Lesk also suggested that protein evolution conserves the fold, an equivalence class of protein structures defined as a spatial arrangement with "the same major number and direction of secondary structures with a same connectivity".[14] This view strongly influenced the classification of protein structures in databases such as SCOP[14] and CATH,[15] where proteins recognized as evolutionarily related (i.e., homologous) are classified in the same structural fold. However, the accumulation of protein structure data has revealed that "fold change" is relatively frequent in the evolution of proteins[16–18] and that folds or topologies as defined in SCOP and CATH fail to pass tests of consistency with respect to structure similarity measures.[19]

Protein classification and molecular clocks are intimately related. The very possibility to objectively classify protein structures requires that the structure similarity measure is transitive, i.e., similarity between *a* and *b* and between *b* and *c* must imply similarity between *a* and *c*. This property is guaranteed by the phylogenetic trees underlying the gene duplication process. Therefore, if protein sequences or structures diverge regularly in evolution (the molecular clock hypothesis), their divergence can be used for objective and consistent classification. However, if divergence is accelerated for instance through positive selection, function diversification or large insertions and deletions (which are not mutually exclusive processes), we expect that the transitive property is violated and consistent classification is not possible. We will test here the molecular clock through a quantitative study of structural and functional divergence in the evolution of four large superfamilies: Globins, Aldolases, P-loop containing nucleotide triphosphate hydrolases, and NADP-binding Rossmann-like domains.

The relationship between protein function on one hand, and sequence and structure on the other hand, has been subject to intense investigation. For instance, Devos and Valencia[20] and Wilson *et al.*[21] independently concluded that protein function, assessed through the Enzyme Commission (EC) classification, is generally conserved above 40 percent sequence identity. Using the CATH classification of proteins, Todd, Orengo and Thornton[22] found that function divergence is common in homologous superfamilies, although the extent of this divergence varies from one superfamily to the other. Lecomte *et al.*[23] studied the divergence of protein sequences, structures and functions in the globins superfamily, and Sangar *et al.*[24] found that, for proteins with more than 50% sequence identity, function assigned through homology is correct in 94% of the cases. It has been found through these studies that structure similarity at the fold level is compatible with a multiplicity of functions. It has been proposed that these multiple functions originated from divergent evolution followed by structure and function diversification,[25] a view that we adopt in this analysis, examining proteins in the same superfamily that are believed to share a common ancestor. This multiplicity of functions makes function prediction from sequence and structure a difficult problem, because homologous proteins often have different functions.[22,26] And yet it is a more and more urgent problem, due to the accumulation of huge sequence data waiting for annotation.[27] Despite the ambiguity of the structure-function relationship, it has been found that structural information provides added value for function prediction with respect to plain sequence information.[28,29] We can shed light on the structure-function uncertainty[30] using evolutionary information, since phylogenetic, structural and functional distance are correlated.[31] These considerations motivated us to undertake a study of how function change and function conservation influence the evolutionary divergence of protein structures.

RESULTS

Contact divergence: a new measure of structure divergence

In their seminal paper, Chothia and Lesk quantified protein structure divergence through the RMSD. However, this measure can be computed only for aligned residues that are well superimposed in space. In practice, it is necessary to fix a cut-off distance that specifies which residues are well superimposed, and the RMSD increases with the cut-off. A more robust measure of structure similarity is the number of superimposed residues within this cut-off, called percentage of structure identity (PSI). This and other measures of structure similarity have to be normalized in such a way that the comparison between two unrelated proteins is not trivially correlated with their size. To achieve this normalization, one typically uses the mean and standard deviation of the similarity of unrelated proteins of similar length, assuming either Gaussian statistics (the *Z* score) as in the Dali program,[32] or extreme value statistics, as in the significance score of the program Mammoth.[33] However, this normalization has the drawback that the similarity of related proteins becomes strongly dependent on their length. For instance, the *Z* score of 100 percent PSI increases as a power law of protein length. Therefore, this significance can not measure the evolutionary divergence. A possible solution to this problem is a new type of normalization,

such as the TM-score proposed by Zhang and Skolnick.[34] Proteins with 100 percent structure identity have TM score equal to one and unrelated proteins have TM score that uncorrelated with their length.

We have recently observed that the contact overlap (see Materials and Methods) performs better than the number of superimposed residues for the sake of classifying protein structures based on their similarity.[19] There are two reasons for this: (1) The contact overlap weights more the aligned residues in the core of the protein, where the number of contacts is large or, equivalently, it penalizes less the non-superimposed residues with few contacts, such as those in loops; (2) Relative motions of two subdomains, such as hinge motions, are much less penalized by the contact overlap since intra-subdomain contacts are conserved in the two conformations. Therefore, we look here for a way to normalize the contact overlap making it independent of protein length both for related and unrelated proteins.

To this aim, we will use the analogy with an evolutionarily motivated measure of protein sequence divergence. Consider two proteins related by gene duplication that diverged during t years. We assume that the probability that no substitution happens in the time t decays exponentially with rate $1/\tau$ as $exp(-t/\tau)$. The conditional probability that two amino acids are equal given that at least one change happened at their common position is $p = \sum_a f(a)^2$, where $f(a)$ is the frequency of amino acid a. Using the frequencies $f(a)$ measured by Jones *et al.*[35] on the SwissProt database, we get $p = 0.058$. Therefore, the probability to observe the same amino acid at an aligned position i in two proteins that diverged for a time t is

$$P\{A_i^1 = A_i^2\} = e^{-t/\tau} + p(1 - e^{-t/\tau}). \quad (1)$$

We can estimate the probability that two amino acids are equal as the sequence identity between the two proteins, SI. This estimate is only rigorous if the substitution process is independent at each protein position, which is clearly not true, but this is an almost unavoidable assumption. Using $P\{A_i^1 = A_i^2\} \approx$ SI, we can solve Eq. (1) finding the evolutionary divergence time t as

$$t/\tau = -\log\left(\frac{SI - p}{1 - p}\right). \quad (2)$$

(from here on, log indicates the Neperian logarithm). When SI $\gg p$, this formula coincides with the standard Poisson formula used to estimate evolutionary distances.[36] Equation (2) is also in fair agreement with simulations of protein sequence evolution subject to the global constraint of folding stability,[37] provided that SI is not close to p, in which limit the evolutionary information is wiped out. The formula is not defined if SI $\leq p$.

We generalize Eq. (2) to the evolutionary divergence of the inter-residue contacts in a protein structure. Given two proteins with contact overlap q (see Materials and Methods), we define their contact divergence as

$$D_{cont}(q, L) = \begin{cases} -\log\left(\frac{q - q_\infty(L)}{1 - q_\infty(L)}\right) & \text{if } q > \epsilon(L) \\ D_0 - (q - \overline{q}(L))/\sigma_q(L) & \text{otherwise} \end{cases} \quad (3)$$

The upper line of the aforementioned equation defines the contact divergence of related proteins, in analogy to how sequence identity is tranformed to estimate evolutionary divergence in Eq. (2), so that $D_{cont} = 0$ for proteins having identical contact matrices and $D_{cont} \to \infty$ for $q \to q_\infty(L)$. Therefore, the parameter $q_\infty(L)$, which is the analogous of p for protein structures, represents the asymptotic limit of the contact overlap after a very long evolutionary time. For $q \leq q_\infty(L)$ the logarithm in the upper line is not defined, and we define in the lower line the contact divergence of unrelated and distantly related proteins. The cross-over takes place at $q \leq \epsilon(L) > q_\infty(L)$, and after this point contact divergence is given by a linear function of the Z score of the overlap, $Z = (q - \overline{q})/\sigma_q$. We have tested in previous work that the Z score is a convenient similarity measure for unrelated proteins. As for other structure similarity measures, the mean and standard deviation of the overlap of unrelated proteins, $\overline{q}(L)$ and $\sigma_q(L)$, depend on protein length. To simplify this dependence, we parameterize the size of the protein pair as the geometric mean of the length of the two proteins, $L = \sqrt{L_1 L_2}$, and we measure $\overline{q}(L)$ and $\sigma_q(L)$ for unrelated protein pairs of length L in the representative set of structural domains ASTRAL40 (see Materials and Methods).

The formula (3) depends on the parameters $q_\infty(L)$ (asymptotic overlap), $\epsilon(L)$ (threshold overlap) and D_0. To reduce the number of free parameters, we make the following assumptions. First, we assume that the asymptotic overlap $q_\infty(L)$ is a linear function of the mean and standard deviation of the overlap of unrelated proteins:

$$q_\infty(L) = \overline{q}(L) + A\sigma_q(L). \quad (4)$$

Since $\overline{q}(L)$ and $\sigma_q(L)$ depend on length, so does $q_\infty(L)$ as well. A is a free parameter whose positive value means that the asymptotic overlap of homologous proteins separated by a very long evolutionary distance is larger than the mean overlap of unrelated proteins, i.e., the memory of the relatedness is never lost. Second, we fix the parameter $\epsilon(L)$ by imposing continuity of Eq. (3) at $q = \epsilon(L)$

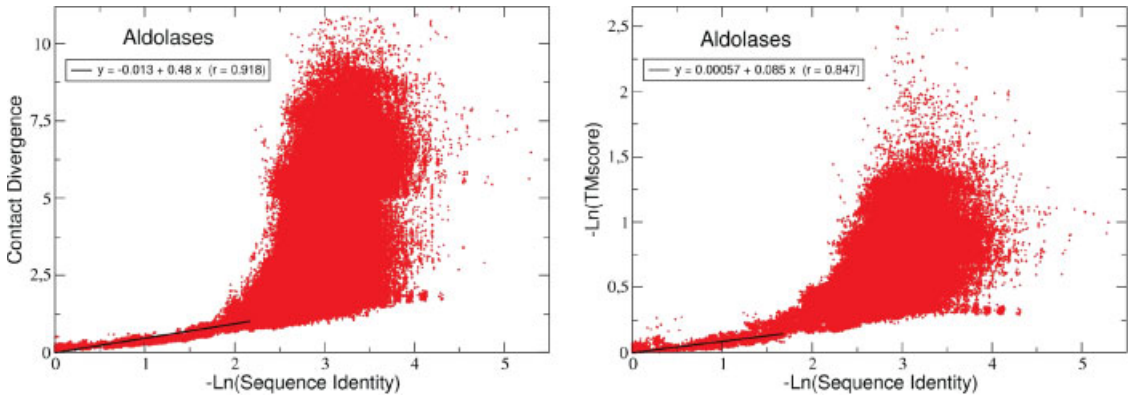

Figure 1

Structure divergence versus sequence divergence for proteins in the aldolase superfamily. Left plot: Contact divergence. Right plot: natural logarithm of the TM score. The linear fits are restricted to the largest region in which the intercept of the fit does not differ significantly from zero. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

(see Materials and Methods). This continuity condition can be imposed only if the parameter D_0, which is independent of length, is large enough. We therefore decided to take the smallest value of D_0 such that the continuity condition is met for all protein pairs in our representative data set of single domain proteins (see Materials and Methods). With these choices, the only free parameter in the definition of the contact divergence is the parameter A in Eq. (4). We chose A by testing the consistency between the new measure and evolutionarily grounded classifications of protein structures. We clustered 2890 nonredundant protein structures with less than 40 percent pairwise sequence identity using the average linkage algorithm applied to different similarity scores, and compared the corresponding classifications with the SCOP and CATH classifications at superfamily level using the weighted kappa measure.[38] This level was chosen because superfamily relationships reflect common evolutionary origin, and because SCOP and CATH agree with each other at the superfamily level much more than at the fold level.[19] We also compared our structural clusters to the ones obtained using as similarity the sequence identity after optimal structure alignment (see Supporting Information Fig. 1). Notice that, since protein structure is used for the alignment, this measure is much more reliable than the sequence identity obtained through sequence alignment. At large identity, corresponding to the initial steps of the clustering algorithm, sequence identity is believed to yield reliable phylogenetic trees. Therefore, this comparison tests the ability of the structural score to yield trees that are consistent with the process of evolutionary divergence for closely related proteins, whereas the superfamily comparison addresses farther evolutionary relationships. For each comparison, we selected the maximum weighted kappa for all threshold structure similarities.

The results of these tests (see Table I) show that the contact divergence score outperforms both the Z score of the contact overlap and the TM score regarding its consistency with evolutionary based classifications, such as SCOP superfamilies, CATH superfamilies, and sequence identity based trees. All three evolutionary classifications give very similar rankings, despite the sequence identity measure has a low agreement with the superfamily classifications. This is not surprising, since most pairs have sequence identities below 25 percent (the so-called twilight zone) that would not be significant in the absence of structure information, which is used for superfamily assignment in both CATH and SCOP. We found the worst agreement with sequence identity using the Z score of the overlap. The latter measure reduces as much as possible the length dependence for unrelated protein pairs but it is strongly length dependent for closely

Table I

Consistency Between the Clusters Obtained Through Different Similarity Measures and Evolutionary Based Classifications

Score	Parameter	WK SCOP S.F.	WK CATH S.F.	WK Seq. Id.
Seq. Id.	—	0.48	0.48	—
Z-Score	—	0.63	0.61	0.562
TM-Score	—	0.59	0.58	0.720
Cont. Divergence	$A = 0$	0.56	0.58	0.723
Cont. Divergence	$A = 2$	0.58	0.58	0.745
Cont. Divergence	$A = 3$	0.62	0.60	0.749
Cont. Divergence	$A = 4$	0.64	0.62	0.753
Cont. Divergence	$A = 5$	0.66	0.64	0.754
Cont. Divergence	$A = 6$	0.64	0.62	0.750
Cont. Divergence	$A = 8$	0.63	0.61	0.692

As a test set, we used a consensus set of 2890 nonredundant domains classified in 779 SCOP superfamilies and 885 CATH superfamilies. Consistency was assessed through the maximum weighted kappa measure[38] obtained for all threshold similarities. We did not perform computations for $A = 1$ since, interpolating results with $A = 0$ and $A = 2$, it is clear that this value is suboptimal. The same holds for $A = 7$.

related proteins. For instance for $q = 1$, corresponding to $D_{cont} = 0$, we have $Z = (1 - \overline{q}(L))/\sigma_q(L)$. The worst agreement with the superfamily classifications was found for the TM score, confirming that scores based on the number of superimposed residues perform worse than scores based on contacts for detecting distant evolutionary relationships. The best consistency with all evolutionary classifications was found for the contact divergence measure with $A = 5$. In the following, when we mention contact divergence we will mean this choice of parameters.

Molecular clock for structure divergence

We now analyse four large superfamilies, each containing more than thousand crystallized structures: Globins, Aldolases, P-loop containing nucleoside triphosphate hydrolases and NADP-binding Rossman-fold. The list of domains and their definition were taken from the CATH database.[15] The list of the corresponding SCOP domains is very similar, but their definition is somewhat different, since SCOP domains are typically larger than CATH. We eliminated NMR structures, chains with more than one domain, for which function assignment is problematic, and redundant domains almost identical both in sequence and in structure. Identical sequences with slightly diverged structures were retained in order to have a glimpse at conformation changes. For each pair of domains in the same superfamily we measured pairwise dissimilarities in structure, sequence, function and length (see Materials and Methods). In particular, structure divergence was measured through the contact divergence score D_{cont} defined earlier, sequence divergence was measured as $-\log(SI)$, where SI is the sequence identity obtained through structure alignment, and function similarity was defined to be one if all GO terms[39] of the two proteins coincide, zero otherwise. For globins we also used InterPro signatures[40] to complement GO terms.

First, we examined the relationship between sequence and structure divergence. One can see from Figure 1 that structure divergence increases almost linearly with sequence divergence when this is not too large. If the sequence diverges in a clock-like manner, this result is consistent with the extended molecular clock hypothesis that structure divergence accumulates linearly with time. Figure 1 represents the Aldolase superfamily. In the left plot we measure structure divergence through the contact divergence measure. We linearly fitted contact divergence versus sequence divergence up to the point where the intercept of the fit differs significantly from zero (i.e., where the intercept becomes larger than its standard error). This point corresponds to SI = 0.115, and the correlation coefficient of the fit is $r = 0.918$. We repeated the same procedure using the TM score, measuring TM score divergence as $-\log(\text{TM score})$. Also in this case the molecular clock hypothesis holds, but its range of validity is narrower (it is SI $\geq$ 0.187) and the correlation coefficient is smaller, $r = 0.847$. If we assume that the sequence divergence $-\log(SI)$ evolves approximately clockwise, the fact that contact divergence is approximately linearly related to sequence divergence over a broader range suggests that this measure evolves more clockwise than the TM score and it is more convenient for quantifying the evolutionary divergence of protein structures.

The other superfamilies yielded similar results, except for the Globin superfamily for which several proteins with conformation changes and unchanged sequences are present. In this case, the intercept of the linear fit is significantly different from zero also for very small divergence, and we could not apply the aforementioned method to determine the range of validity of the molecular clock. These results confirm that contact divergence is a convenient measure for quantifying protein structure change in evolution.

Structure diversity explosion

For small sequence identity (large divergence), the approximate proportionality between structure divergence and sequence divergence disappears and one can see an explosion of structure diversity. One possible explanation to this spectacular explosion, observed for all structure divergence measures and in all four superfamilies with very similar characteristics, is the attenuation of functional constraints, since almost all of the strongly diverged pairs have different function (see below). Strongly diverged pairs also tend to have large insertions and deletions, which may be responsible for the increased structure divergence. As we will see in the following, our analysis supports both interpretations. However, an even simpler interpretation is also possible.

As expressed in Eq. 1, after a very long divergence time multiple substitutions have occurred at most sites, and the sequence identity of two homologous proteins reaches an asymptotic distribution where aligned residues may become identical by chance rather than by common origin and all evolutionary information is lost. This situation can be studied by simulating protein sequence divergence through random mutations that are fixed if they do not appreciably modify the stability of the target protein structure, assessed through an effective free energy function.[37] Using these simulations, the mean number of attempted mutations, which is related with the evolutionary divergence time, may be represented versus $-\log(SI)$ as in Figure 2. We can see from this plot that the sequence divergence $-\log(SI)$ is a reliable estimate of the divergence time only for large enough sequence identity (small divergence), whereas large sequence divergences tend to strongly underestimate the divergence time. After a very long time all evolutionary information is lost, and sequence identity reaches an asymptoyic distribution centered around the small value

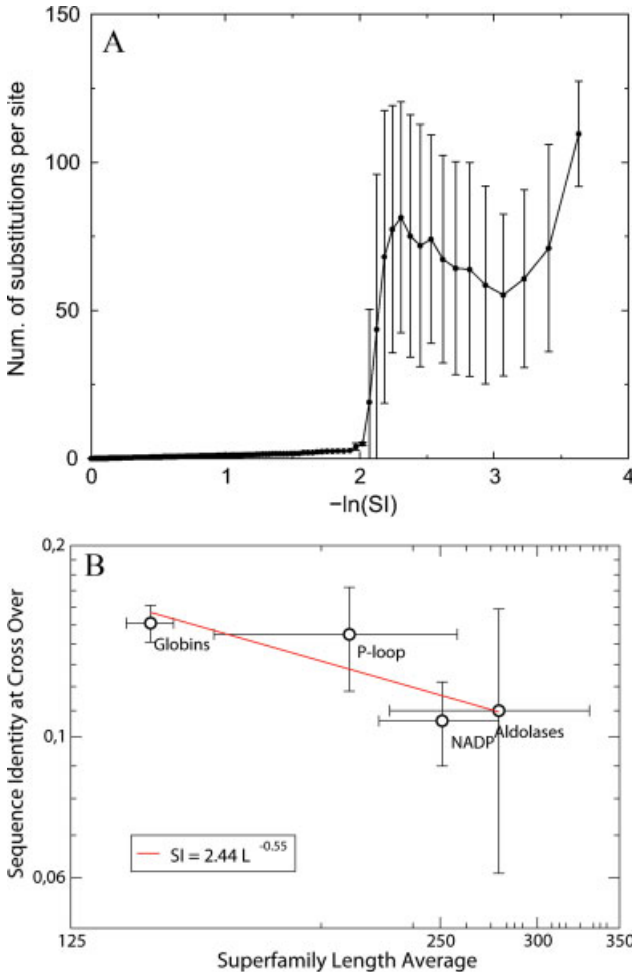

Figure 2

(**A**) Results from a simulation of protein sequence evolution with conservation of the folding stability of the target structure. The mean number of substitutions, measuring the divergence time, is plotted versus the sequence divergence −log(SI). Data modified from.[37] (**B**) For the four superfamilies studied, we plot the sequence similarity at which the structural explosion occurs versus the average length of the superfamily. The error bars indicate the uncertainty on the cross-over point and the standard deviation of protein length. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

SI $= p \approx 0.058$. The largest sequence identity found with non negligible probability in this asymptotic ensemble determine the cross-over, since smaller identities do not allow to estimate the divergence time. Both probabilistic arguments and simulations[37] suggest that the sequence identity at the cross-over decreases with protein length L approximately as $L^{-1/2}$. For the four superfamilies studied in Figure 3, we estimated the sequence identity at the cross-over by plotting the standard deviation of contact divergence versus sequence identity. This quantity makes a jump at the cross-over that allows to identify it with reasonable accuracy (data not shown). For the Aldolases and NADP superfamilies, which do not present important conformation changes, the cross-over estimated in this way is in very good agreement with the limit of validity of the molecular clock estimated in the previous section through the condition that the intercept of the linear fit should be zero. We found that the cross-over identity decreases as $L^{-0.55}$ when the mean length L of the superfamily increases (see Fig. 2, right plot), consistent with the aforementioned interpretation. Therefore, the apparent explosion of structure divergence at the cross-over might be a simple consequence of the fact that sequence identity below the cross-over strongly underestimates the divergence time, coupled with the relaxation of functional constraints on protein structure that will be discussed below.

Functional constraints on protein evolution

We represent in Figure 3 structure divergence versus sequence identity, distinguishing protein pairs that perform the same function (i.e., all their GO terms regarding the Molecular function are equal) from those with different functions. We only consider in this analysis proteins whose GO terms have been manually curated, as indicated by their evidence code. As one can see from this figure, proteins sharing the same function are more conserved in sequence and in particular in structure with respect to pairs with different functions. This result is expected, since protein function is known to constraint sequence and structure. Nevertheless, the strength of these constraints is surprising, since we found very few pairs with different functions having contact divergence larger than 2, and almost all of them can be attributed to conformation changes rather than evolutionary divergence (see below). Structure divergence is very limited even for pairs with sequence identity lower than the cross-over of structural explosion, for which the evolutionary divergence time may be very large. Moreover, as we will see later, several pairs with very large structure divergence have been electronically annotated as having the same GO terms. Both our results and the InterPro annotations suggest that these electronic annotations may be incomplete. Therefore, using the knowledge of the strength of functional constraints on protein structure would have avoided these incomplete annotations.

Conversely, we observed many pairs of proteins with different function and very similar structure, confirming the known fact that even small structure divergence may be sufficient to change protein function. In other words, structure divergence is a very strong indication of functional change, but structure conservation does not always imply function conservation.

Electronic annotations

We now plot in Figure 4 structure divergence versus sequence identity also for proteins that have been electronically annotated, according to the evidence codes in

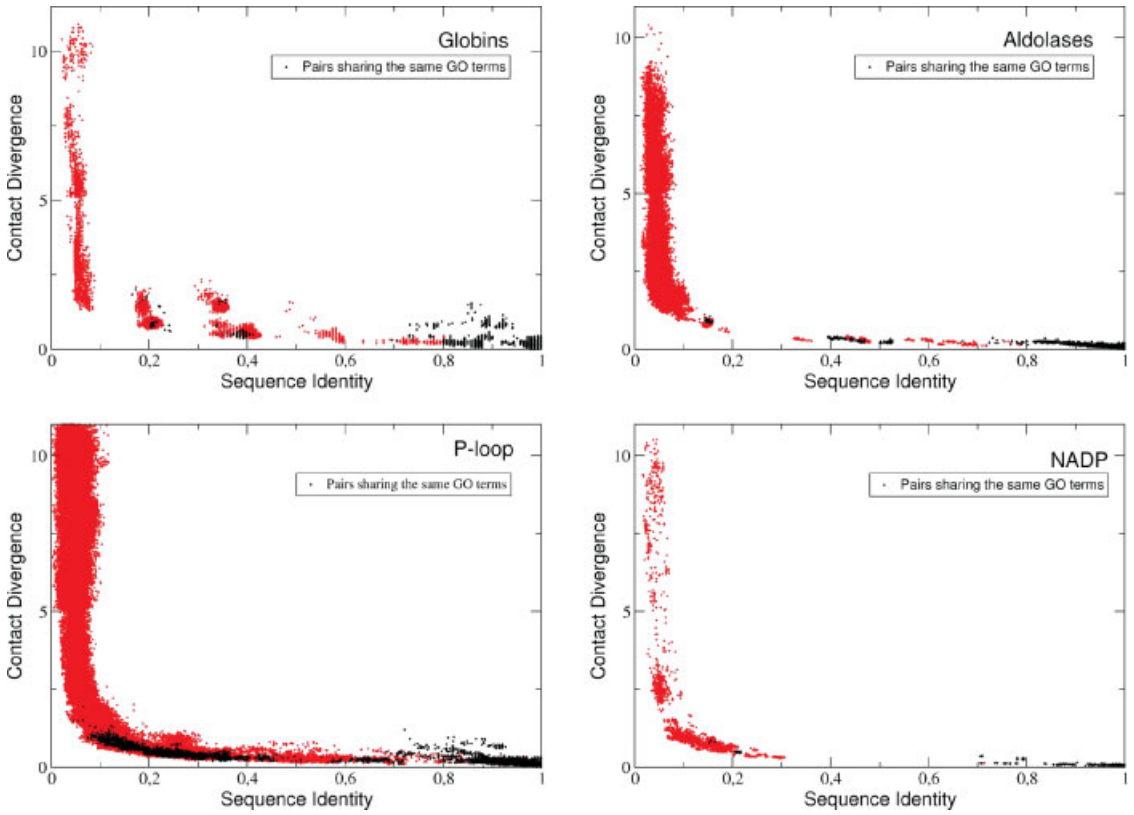

Figure 3

For each of the four superfamilies we plot contact divergence versus sequence identity, distinguishing protein pairs performing the same function according to all of their GO terms (dark points). Only proteins with manually annotated GO terms are represented. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the GO. For the Aldolases and the NADP superfamilies the plots are very similar to Figure 3, and they are not shown. However, for the P-loop and in particular the Globins superfamilies, we found a very large number of pairs annotated as having the same function but with very large contact divergence. Most globins that are electronically annotated are classified as having the heme binding, oxygen binding, and oxygen transporter function. We then adopted the InterPro classification (see Materials and Methods), which distinguishes different types of Hemoglobin chains (alpha, beta, zeta, pi), and lamprey and annelid globins. Although all of them are involved in oxygen transport, they may have rather different affinity for oxygen and regulation mechanisms[41] and may perform secondary functions,[42] which makes it reductive to classify all of them under the same functional class. Besides, paralog genes are believed to perform different functions in order to be retained in evolution, so that classifying all hemoglobin types under the same function is likely to reduce too much the resolution at which we can look at protein function. We found the surprising results that no protein pair with the same InterPro signature has contact divergence larger than 2 (see Fig. 4), except for a pair involved in a large conformation change, in perfect agreement with the result that we obtained for manually annotated GO terms. This result suggests that proteins with contact divergence larger than 2 with respect to manually annotated proteins with the same function may be incompletely or wrongly annotated. For the P-loop superfamily all such outliers, i.e., the dark points in Figure 4 with contact divergence larger than 2, are explained by only 5 domains (PDB codes 1xjcA, 1gvnB, 1gvnD, 1y63A and 1ghhA), for the NADP superfamily we identified two proteins that may be incompletely annotated (1jax and 1jay), and no one for the Aldolases superfamily, whereas most globins are insufficiently annotated electronically, as discussed earlier.

Global structure conservation and function prediction

As expected, the results presented in the previous section show that large sequence and structure divergence are strong predictors of function change, and sequence and structure conservation are (weaker) predictors of function conservation. To quantitatively assess the performances of

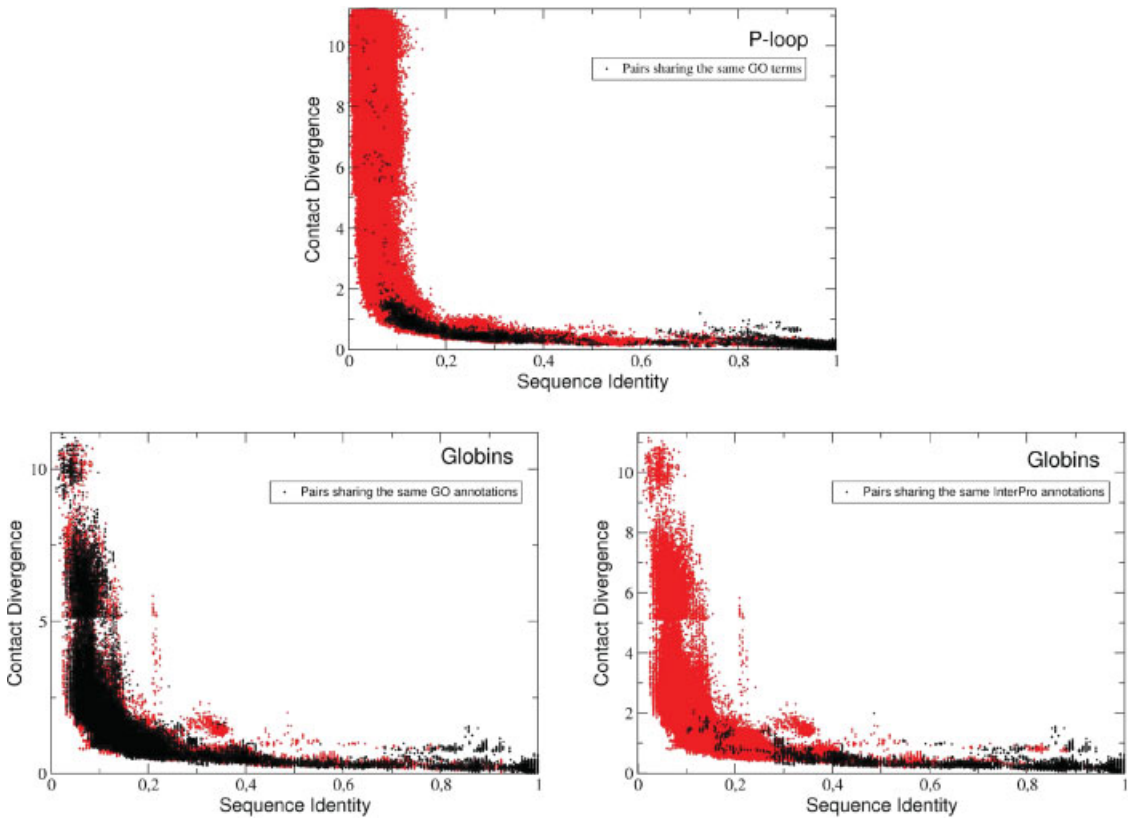

Figure 4

For the Globins and P-loop superfamilies we plot contact divergence versus sequence identity, distinguishing protein pairs sharing the same function according to the GO terms. Also electronically annotated proteins are considered in these plots. For the Globins superfamily we represent on the lower right panel the same plot where we identify proteins with the same function as those with the same InterPro signatures. Notice that we find several pairs with the same electronically annotated function but very large contact divergence. Such pairs are not found in case of manual annotations and InterPro signatures. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

contact divergence and other sequence and structure similarity measures under this respect, we measured the sensitivity and selectivity (see Materials and Methods) for predicting function conservation using different thresholds on structure similarity. The corresponding ROC plots show almost perfect Area Under the Curve (AUC), tabulated in Table II (see Supporting Information Fig. 2). AUC of one means perfect prediction, 0.5 indicates a random prediction. All scores perform very similarly but, surprisingly, the sequence identity score is an even better predictor than actual structure similarity measures. Notice, however, that we measure sequence identity after optimal structure alignment, so that the performances of this measure would not be possible if we did not dispose of structural information.

Structure evolution is accelerated upon function change

To characterize more quantitatively the effect of function on structure divergence, we quantified the relationship between sequence and structure divergence. For sequence identities above the cross-over, we can estimate the divergence time either through sequence divergence as $t \approx -\log(\mathrm{SI})$ or through structure divergence as $t \approx D_{\mathrm{cont}}$. These two estimates are proportional, which means that the molecular clock based on sequence and the one based on structure are consistent.

However, a closer look shows that the two molecular clocks present discrepancies when functional changes occur. Through a linear fit, we computed the slope of D_{cont} versus $-\log(\mathrm{SI})$ before the cross-over, distinguishing protein pairs with the same function (see Table III). One can see that all of these slopes are smaller than one, confirming that protein structure diverges more slowly than sequence, and they are all in a relatively limited range, from 0.25 for P-loops to 0.48 for Aldolases.

For all four families, protein pairs with different functions present significantly larger slopes (in the range 0.29 to 0.48) than those with the same function (from 0.25 to 0.37). Although not unexpected, this is a rather interesting result, since it demonstrates a quantitative

Table II
AUC (Area Under the Curve) of the ROC Plots of Function Conservation Predictions Using Different Structure Similarity Measures

Superfamily	Seq. Id.	Cont.Div.	*Z-Score*	TM-Score
Aldolases	0.988	0.980	0.979	0.988
Globins	0.977	0.984	0.979	0.982
P-loop	0.978	0.973	0.977	0.974
NADP	0.840	0.812	0.809	0.833

influence of protein function on the sequence to structure relationship. Moreover, it suggests possible improvements to protein function prediction. In fact, it is known that very small changes in sequence and structure are sufficient to modify protein function, so that sequence and structure conservation are not a sufficient indication of function conservation. Our observation that function change modifies quantitatively the sequence to structure relationship suggests that this information could be used in order to predict function conservation more reliably.

This influence of function change on the sequence to structure relationship can be interpreted either as due to the fact that function change relaxes the constraints on protein structure (negative selection) or due to positive selection for modified structure to perform the new function. We think that the latter mechanism is more relevant. In fact, we observed this behavior while structure divergence is still within the range $D_{\text{cont}} < 2$ typical for proteins with the same function, so that structural constraints imposed by function conservation would be fulfilled.

Evolutionary rates and clustering

We have seen earlier that, for sequence divergence below the cross-over, sequence divergence and structure divergence are approximately proportional. This implies that the divergence times estimated through sequence divergence and through structure divergence are proportional, so that the molecular clock based on sequence and the one based on structure are consistent. This approximate clockwise evolution of protein structures has an important implication for protein structure classification. In fact, if the molecular clock approximately holds, structure divergence is expected to be able to reconstruct the phylogenetic tree underlying protein evolution, similar to how this is done with sequence divergence. Given a phylogenetic tree, the time distance from the leaves of the tree passing through their closest common ancestor is ultrametric,[43] i.e., if C is the outgroup of the triple A,B,C it holds $t_{AC} = t_{BC} > t_{AB}$. This relationship is valid for all triples, and it guarantees that the transitive property holds for all distance thresholds, i.e., if A and B are related and B and C are related also A and C must be related. Therefore, relatedness along the tree induces an equivalence relationship whose equivalence classes are the phylogenetic groups. If the molecular clock approximately holds, the divergence $D_{AB} \approx kt_{AB}$ can be used to estimate the divergence time and to reconstruct the tree.

Table III
Slope of Contact Divergence Versus Sequence Divergence for Protein Pairs Sharing the Same Function and for all Possible Protein Pairs

Superfamily	Slope (all pairs)	Slope (same function)
Aldolases	0.4830 ± 0.0007	0.3733 ± 0.0006
Globins	0.3912 ± 0.0003	0.3572 ± 0.0007
P-loop	0.2888 ± 0.0008	0.2529 ± 0.0011
NADP	0.3914 ± 0.0005	0.3245 ± 0.0007

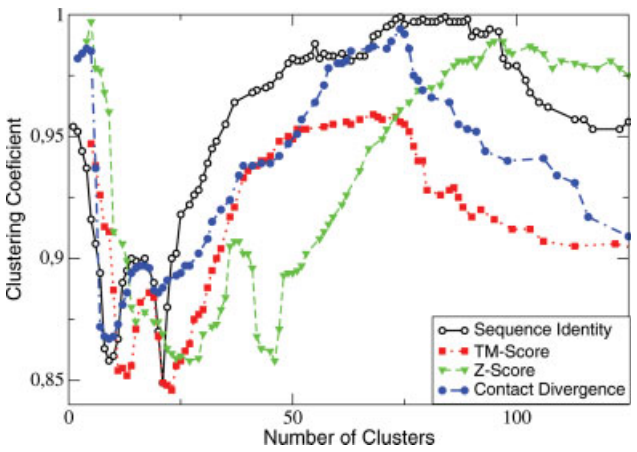

Figure 5
For the NADP superfamily and for each different divergence measures and distance thresholds, we constructed a network by joining all pairs of proteins closer than the threshold. For each network, we plot the clustering coefficient versus the number of connected components, i.e., the number of clusters obtained with single linkage, which decreases with increasing distance threshold. One can see that there is a range of optimal similarity threshold such that the clustering coefficients are close to one, as expected from the molecular clock hypothesis. Also notice that the clustering coefficients present dips that suggests that the evolutionary rate is not constant in this region.

To test the clustering properties of the divergence measures studied here, we measured the clustering coefficient (see Materials and Methods) of the networks constructed by joining together proteins with D_{AB} smaller than some threshold. If the clustering coefficient is one, all related proteins share all their neighbors and transitivity exactly holds. Ultrametricity implies clustering coefficient equal to one for all thresholds. The validity of the molecular clock hypothesis therefore implies that the clustering coefficient is close to one for all thresholds.

Figure 5 shows the clustering coefficient of the networks obtained with a given distance measure and given threshold versus the number of connected components of the same network (i.e., the number of clusters obtained with single linkage clustering). The larger this number, the smaller the distance threshold. There is a range of distance thresholds such that the clustering coefficient is close to one, consistent with the molecular clock hypothesis. The clustering coefficient is larger using

sequence identity measured with the optimal structure alignment than using structure similarity measures. This suggests that protein sequences evolve in a more clock-like fashion than protein structures. Among structure similarity measures, the contact divergence yields the best clustering coefficient for NADP and P-loop whereas the TM score is the most clock-like for globins and aldolases.

Notice that in Figure 5 the clustering coefficients present dips suggesting that the evolutionary rate is not constant for some values of the thresholds, corresponding to some typical evolutionary distance. We conjecture that this phenomenon is related to the rate acceleration when the protein function changes. This interpretation is supported by the fact that we measured a significantly larger rate of structure divergence for proteins with different function. We will analyse this issue in future work.

Conformation changes

Figures 3 and 4 show some outliers, i.e., protein pairs with large sequence identity whose structures diverge much more than expected. These are often examples of conformation changes, i.e., proteins that change conformations while performing their biological activity. We discuss in this section the examples that produce the most severe outliers in Figure 4, i.e., pairs whose structure divergence is much larger than expected based on their sequence identity. The conformation changes discussed below are represented in Supporting Information Figure 4.

Globins

Many proteins in this superfamily have been crystallized with different co-factors (mostly oxydized and reduced Heme) that give raise to small scale conformation change. Engineered mutants as well are associated to small conformation changes. The strongest conformation change involves hemoglobin crystallized together with the alpha-haemoglobin-stabilizing protein (AHSP), which inhibits its capacity to react with oxygen (PDB code 1z8u). "The structure of AHSP bound to ferrous alpha-HB is thought to represent a transitional complex through which alpha-Hb is converted to a nonreactive, hexacoordinate ferric form (...) The structure of the complex shows significant conformational changes involving translocation of main chain atoms by as much as 10 Å".[44] This structure is responsible of the most serious outliers in Figure 3, in particular an almost vertical line of outliers at sequence identity ≈ 0.22, a large blob of outliers at sequence identities between 0.30 and 0.40, and a long horizontal line of outliers with SI > 0.4.

Aldolase

The most relevant conformation change in this superfamily involves a mutant (Y24F) of the protein glicolate oxidase, PDB code 1gylB. This mutant could not be crystallized with its natural cofactor FMN. According to the authors, "the absence of the cofactor FMN and differences in packing of the subunits give raise to much larger differences in the structure than the mutation per se".[45] This structure is involved in most outliers with sequence identity larger than 0.5.

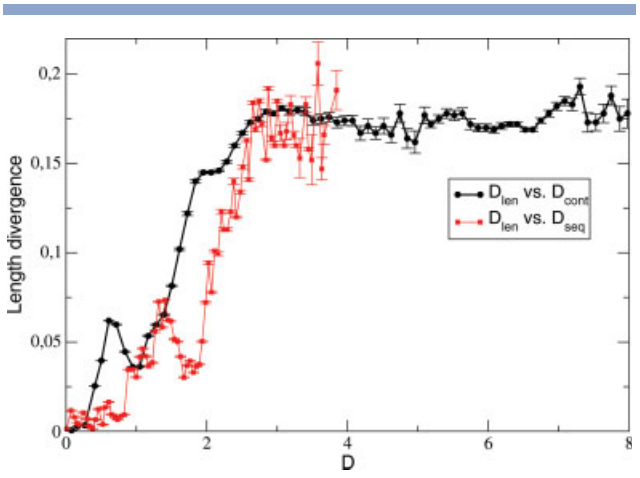

Figure 6

For the case of globins, we show the effect of sequence and structure divergence on the average length divergence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

NADP

In this superfamily, the largest conformation change involves the protein Abeta-binding alcohol dehydrogenase (ABAD), PDB code 1so8A which displays substantial distortion of the NAD-binding pocket and the catalytic triad.[46] Other smaller conformational changes involve the Enoyl-acyl carrier reductase of *Plasmodium falciparum* crystallized with different ligands.[47]

Ploop

In this superfamily, the most severe outliers (in particular, those with SI > 0.6) involve three structures of the protein p21(H-ras) studied at different time points along the GTPase reaction with the synchrotron Laue method[48] (PDB codes 1plj, 1plk and 1pll).

Relationship between length divergence and structure divergence

Finally, we studied how length divergence (defined in Materials and Methods) influences sequence and structure divergence and is influenced by it. Large length differences between two proteins are an indication that large insertions and deletions have occurred in their evolution. However the contrary does not hold, i.e., even proteins of the same length may undergo multiple insertions and dele-

tions in their evolution. We show in Figure 6 results for the case of Globins. Other superfamilies yield qualitatively similar pictures. As expected, length divergence increases more or less gradually with sequence and structure divergence. It then reaches a plateau, more or less corresponding to the cross-over of the structural divergence explosion. Beyond the cross-over, most protein pairs differ significantly in length. Interestingly, large length divergence strongly predicts function change (see Supporting Information Fig. 3). On the other hand, similarity in length is a weak predictor of sequence, structure and function conservation (data not shown). In this case, the four superfamilies yield different pictures: whereas for Globins proteins with similar length tend to be similar in sequence, structure and function, this is less true for the other superfamilies.

DISCUSSION

In this study, we examined how protein function change and protein function conservation quantitatively influence the relationship between sequence and structure divergence in evolution. We quantified structure divergence through a novel measure, the contact divergence, which is based on the similarity of contact matrices. This measure is evolutionarily motivated, since it is constructed in analogy with a sequence divergence measure grounded in molecular evolution studies, and it is properly normalized both for related and for unrelated pairs, in such a way that it is suitable both for evolutionary analysis and for protein structure classification. We tested that this measure is more consistent with evolution based classifications than other previously proposed measures of structure divergence, and that it allows to better represent the molecular clock of protein structure divergence.

Our first qualitative conclusion confirms that, for small divergence, structure divergence and sequence divergence are proportional, as previously shown by Chothia and Lesk using as structure divergence measure the RMSD.[13] This implies that the molecular clock hypothesis approximately holds also for protein structure divergence if it holds for sequence divergence. The approximate validity of the molecular clock is also supported by our finding that networks constructed using structure similarity have clustering coefficient close to one, so that they are consistent with phylogenetic trees. Therefore, we can use structure similarity to reconstruct evolutionary trees for protein structures.

Secondly, our results show that proteins that perform exactly the same molecular function are limited in their sequence and, even more, structure divergence. Although this result is expected as a consequence of functional constraints on protein structure, the strength of these constraints is somewhat surprising. Notice that conservation seems to act on global structure similarity measures, not only on the active site. This is at first surprising, but it is consistent with the idea that allosteric effects at the level of the whole structure are important for protein function. This finding may have important consequences both for protein structure and for protein function prediction. Concerning structure prediction, if two proteins perform the same function they will have very similar structures even if their sequence identity is below the twilight zone, and the known structure of one of them will be a good template for homology or threading based modeling of the other one even at very low identity. Concerning function prediction, we have seen that structure divergence larger than a threshold is an almost certain indication of some (possibly subtle) function change. We have also shown that this observation can be used to identify electronically annotated functions that are likely to be incomplete or wrong. This observation can be therefore used to improve automatic annotation methods. The complete linkage clustering method, which forbids to join in the same cluster any two proteins with divergence larger than a threshold, should be the natural way to exploit structural information for automatic function prediction. We found that the ROC plot for predicting protein function from structure similarity have an area under the curve very close to one, meaning that it is possible to achieve very good prediction accuracy if the structure is known, or if it can be predicted through threading methods.

Third, we have found that the rate of structure versus sequence divergence is larger for proteins performing different functions than for proteins performing the same function. This acceleration may be attributed either to positive selection for new function or to relaxation of negative selection for structure conservation upon function change. We prefer the first interpretation, since the acceleration is also observed for low structure divergence, which is compatible with function conservation. The accelerations of the rate of structure divergence are also supported by the observation that the clustering coefficient of networks constructed with measures of sequence and structure divergence present significant dips, indicating violations of the molecular clock hypothesis, and by the finding that protein sequence evolution is more clock-like than structure evolution, also based on the analysis of the clustering coefficient. We conjecture that this is due to the acceleration of the rate of structural evolution in the presence of positive selection for functional changes. We will test this hypothesis in future work. In any case, this finding also suggests a way to improve protein function prediction when structure information is available. In fact, sequence and structure conservation is not sufficient to unambiguously decide that two proteins perform exactly the same function. Complementing structure similarity with a test of the constancy of the evolutionary rate may improve the accuracy of function prediction.

Fourth, we have observed that, below a cross-over value of sequence identity, there is an explosion of structural di-

versity, which may increase much faster than linearly with sequence divergence for proteins with different functions. This finding extends the previous finding of Chothia and Lesk based on the RMSD as divergence measure. The simplest explanation for such an explosion is that, below the cross-over, sequence identity does not allow to estimate the evolutionary divergence time, so that protein pairs with identity below the cross-over may have diverged for a time much longer than what is inferred from their sequence identity, allowing them to reach very different conformations. Despite this simple explanation is supported by the observed relationship between the cross-over values and the protein length, it is interesting that a qualitatively similar explosion of structural diversity has been found in a recent study of protein sequence design.[49] In this study, protein sequences were designed by optimizing the folding stability of a target structure. It was found that, when the target structure and the reference structure in the PDB are very similar, the designed sequence has a rather large identity with the reference sequence. However, when the target and the reference structure become more different, as it would be in case of selection for new function, designed and reference sequence only share very low identity, of the order of twenty percent, i.e., slightly more than the average identity of unrelated protein pairs. This phenomenon has the appearance of a cross-over in the relationship between sequence divergence and structure divergence, very much reminiscent of the one that we observed, and it may provide an alternative explanation for it: When two proteins perform the same function, natural selection targets very similar structures, determining sequence and structure conservation, whereas for proteins with significantly different function natural selection targets different structures, whose typical sequence identities are below the cross-over region. This interpretation is consistent with the findings, here reported, that protein function influences evolution by limiting the extent of sequence and structure divergence in case of function conservation, and by accelerating structure divergence with respect to sequence divergence in case of function change.

Finally, we observed that large length divergence, which is an indication of insertions and deletions, are almost always associated with functional changes (see Supporting Information Fig. 3), but length conservation is not an indicator of functional conservation. In other words, large differences of length of homologous proteins are a strong hint of functional change, i.e., large length differences are hardly neutral under a functional point of view.

MATERIALS AND METHODS

Protein sets

In this work, we used five protein domain sets. (1) A representative set of protein domains having less than 40 percent sequence identity, which are decomposed almost identically in the CATH and SCOP database (consensus set available at the URL: http://ub.cbm.uam.es/research/ProtNet.php). (2) Four superfamilies: Globins, Aldolases (TIM barrel fold), P-loop containing nucleoside triphosphate hydrolases and NADP-binding Rossmann-fold domains. The list and the definition of domains in each superfamily were taken from the CATH database.[15] From all sets, we eliminated NMR structures, domains extracted from multi-domain chains, for which the function assignment is problematic, and domains with both very high structure and sequence identity (the product of sequence identity times contact overlap must be smaller than 0.98). The sequence identity and contact overlap, respectively, took values in the ranges (0.01, 1.00) and (0.13, 1.00) for Globins, (0.01, 1.00) and (0.08, 1.00) for Aldolases, (0.00, 1.00) and (0.04, 1.00) for P-loop, (0.00, 1.00) and (0.00, 1.00) for NADP.

Function characterization

We retrieved Gene Ontology (GO)[39] terms for PDB chains from the web page of the Structure integration with function, taxonomy and sequence (SIFTS) initiative (http://www.ebi.ac.uk/msd/sifts/). To avoid wrong assignments of GO terms to CATH domains, we removed those cases were more than one CATH domain correspond to the same PDB chain. From GO terms, we used only the molecular function annotation and we removed annotations contained in paths already assigned to the same PDB chain.

For globins, GO terms were not specific enough, so we also used InterPro Signatures.[40] Notice that InterPro signatures do not necessarily yield a classification, but we verified that they do in the case of Globins, i.e., that in this set having the same InterPro signature is an equivalence relationship. To retrieve these signatures, we used the SSMap tool[50] that relating PDB chains with UniProt accessions, which also include InterPro Signatures.

We considered GO terms to be manually assigned if their evidence code was 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 for instance ISS (Inferred from Sequence or Structural Similarity), were attributed to computational analysis. The number of manually annotated domains is dramatically reduced: 92 over 676 for NADP, 533 over 1209 for P-loop, 272 over 1341 for Aldolases, 702 over 1313 for Globins.

Divergence measures

For each pair of domains in the same superfamily structure alignments were computed using a new version

of the program Mammoth[33] which was improved performing the same two steps dynamic programming procedure implemented in the multiple alignment version Mammoth-mult,[51] and optimizing the corresponding parameters (UB, APG, Florian Teichert and Markus Porto, unpublished). We measured pairwise dissimilarities in structure, sequence, function and length.

1. Sequence divergence D_{seq} was computed from the sequence identity SI $\in [0, 1]$ measured from the optimal structure alignment as

$$D_{seq} = -\log(\mathrm{SI}). \tag{5}$$

   Here and in the following, log indicates Neperian logarithms.
2. Function similarity was based on GO terms[52] or on InterPro signatures[40] in the case of Globins. Two proteins where considered to perform the same function ($D_{fun} = 0$) if all of their GO terms or InterPro signatures coicided, otherwise they were regarded as performing different functions ($D_{fun} = 1$).
3. For two proteins A and B, we measure their length difference and define the dimension-less variable $d_{len}(A, B)$ as

$$d_{len}(A, B) = \frac{|L_A - L_B|}{\sqrt{L_A L_B}} \tag{6}$$

   We observed that $d_{len} < 1$ for all pairs of proteins in the superfamilies that we examined. We therefore defined the corresponding length divergence as $D_{len} = -\log(1 - d_{len})$, in analogy with sequence or structure divergence (notice that this variable is not defined if $d_{len} \geq 1$, i.e., if $L_A/L_B > 2.6$).
4. The contact overlap is a convenient measure of protein structure similarity, which counts the fraction of contacts in common between two aligned protein structures A and B. The contact matrix of protein A, $C_{ij}^{(A)}$, is defined such that $C_{ij}^{(A)}$ equals one if two heavy atoms of residues i and j are closer than 4.5 Å and $|i - j| \geq 5$, and zero otherwise, so that we do not consider short range contacts. As the same short range contacts are formed with higher probability in unrelated structures, eliminating them has the effect to reduce the mean overlap of unrelated structures. We expect in this way to increase the signal to noise ratio of the contact overlap. Denoting by $a(i)$ the residue in structure B aligned with residue i in structure A, the contact overlap can be written as

$$q_{AB} = \frac{\sum_{ij} C_{ij}^{(A)} C_{a(i)a(j)}^{(B)}}{\sqrt{\sum_{ij} C_{ij}^{(A)} \sum_{ij} C_{ij}^{(B)}}}. \tag{7}$$

   where summation runs over all pairs of residues in protein A.
5. The contact overlap of unrelated proteins depends on their length. We characterize the length of the protein pair as the geometric mean of the two lengths,

$$L = \sqrt{L_A L_B}. \tag{8}$$

   The mean $\overline{q}(L)$ and standard deviation $\sigma_q(L)$ were computed by performing pairwise alignments for the ASTRAL40 set of domains having less than 40% sequence identity, using the program MAMMOTH[33] and considering only pairs in different SCOP folds. In this case, only short regions of the two proteins superimpose in space, typically consisting of one or few secondary structure elements. For each length in the range 40 to 800 residues, $\overline{q}(L)$ and $\sigma_q(L)$ were well fitted by the power laws

$$\overline{q}(L) = 0.386 L^{-0.547} \tag{9}$$

$$\sigma_q(L) = 1.327 L^{-0.673} \tag{10}$$

   To eliminate the length dependence, we used the Z score of the overlap, subtracting the average value of the overlap of unrelated protein pairs with the same length, $\overline{q}(L)$, and dividing times the corresponding standard deviation, $\sigma_q(L)$, to obtain

$$Z = \frac{(q - \overline{q}(L))}{\sigma_q(L)} \tag{11}$$

6. As explained in the main text, the overlap q was transformed to obtain a measure of contact divergence D_{cont}, defined as

$$D_{cont}(q, L) = \begin{cases} -\log\left(\frac{q - q_\infty(L)}{1 - q_\infty(L)}\right) & \text{if } q > \epsilon(L) \\ D_0 - (q - \overline{q}(L))/\sigma_q(L) & \text{otherwise} \end{cases} \tag{12}$$

   The upper line of the aforementioned equation defines the contact divergence of related proteins, in analogy to how sequence identity is tranformed to estimate evolutionary divergence, Eq. (2). It is such that $D_{cont} = 0$ for proteins having identical contact matrices and $D_{cont} \to \infty$ for $q \to q_\infty(L)$. The lower line defines the contact divergence of unrelated or distantly related proteins as $D_{cont} = D_0 - Z$, where Z is defined in Eq. (11). The aforementioned equation depends on three parameters, the asymptotic overlap $q_\infty(L)$, the cross-over overlap $\epsilon(L)$ and the parameter D_0. They are fixed as follows. First, we make the ansatz

$$q_\infty(L) = \overline{q}(L) + A\sigma_q(L), \tag{13}$$

which means that the asymptotic overlap of distantly related proteins is larger than the mean overlap $\overline{q}(L)$ of unrelated proteins Since both $\overline{q}(L)$ and $\sigma_q(L)$ depend on protein length, so does $q_\infty(L)$. The parameter A in the aforementioned equation was fixed to the value $A = 5$ by assessing the contact divergence measure through the clustering experiments described in the main text.

The cross-over $\epsilon(L)$ is fixed imposing that Eq. (12) is continuous for $q = \epsilon(L)$. To this end, we introduce the variable $z = (\epsilon(L) - q_\infty(L))/\sigma_q(L)$. The continuity condition reads

$$z - \log(z) = D_0 - A + \log \sigma_q(L) - \log(1 - q_\infty(L)), \tag{14}$$

The function $z - \log(z)$ takes values between one, for $z = 1$, and infinite, for z tending to zero and to infinite. Therefore, two solutions of the aforementioned equations exist if and only if the right hand side is larger than one, i.e., $D_0 - A + \log \sigma_q(L) - log(1 - q_\infty(L)) > 1$. We decided to take the smallest value of D_0 such that solutions exist for all protein domains contained in our set, i.e.,

$$D_0 = 1 + A - \log \sigma_q(L_{\max}) + \log(1 - q_\infty(L_{\max})) = 10.2 \tag{15}$$

where $L_{\max} = 880$ is the length of the longest domain in all sets that we used and $A = 5$. We then numerically solved Eq. (14) for each L, taking the solution with $z < 1$, which corresponds to an ϵ with small Z score, and we obtained $\epsilon(L) = q_\infty(L) + \sigma_q(L)z(L)$. In this way, the only free parameter in the definition of the contact divergence is the parameter A that expresses the extent to which homologous proteins keep memory of their evolutionary relatedness. This parameter was fixed to the value $A = 5$ by performing the clustering tests described in the main text.

Classification analysis

We assessed the agreement of two classifications through the weighted kappa measure,[38] which uses as reference the expected agreement for two independent classifications with the same number of relationships. We define N_A (N_B) the number of related pairs in classification A (B) of the same N objects, with $N_{tot} = N(N-1)/2$ pairs in total. If A and B are independent, the number of pairs that are either related or unrelated in both A and B is given by

$$N_e = \frac{N_A N_B + (N_{tot} - N_A)(N_{tot} - N_B)}{N_{tot}} \tag{16}$$

We compare this number to the observed number of pairs that agree,

$$N_o = N_{AB} + (N_{tot} - N_A - N_B + N_{AB}), \tag{17}$$

where N_{AB} is the number of pairs that are related in both classifications From this number, the weighted kappa is computed as

$$\kappa = \frac{N_o - N_e}{N_{tot} - N_e}. \tag{18}$$

A value of zero means that two classifications are as related as independent classifications, one means that the two classifications coincide.

Clustering coefficient

The clustering coefficient of node i in a network is defined as the fraction of pairs of its neighbors j and k that are neighbors between each other, and the clustering coefficient of the network is the average clustering coefficient of its nodes. Formally, this is defined as

$$\text{Clustering coefficient} = \frac{1}{N}\sum_i \frac{2\sum_{j<k} A_{ij}A_{ik}A_{jk}}{n_i(n_i-1)} \tag{19}$$

where N is the number of nodes, A_{ij} is the adjacency matrix (one if nodes i and j are joined, zero otherwise), $n_i = \sum_j A_{ij}$ is the number of neighbors or degree of node i If the clustering coefficient is one for all nodes, connections on the network define an equivalence relationship.

We have computed the clustering coefficient for the network obtained by joining domains with similarity $S_{ij} > S_0$, for various values of S_0. To compare different similarity measures, we have plotted the clustering coefficient versus the number of disjoint components found in the network.

ROC plots

Given a binary classifier (predictor) assigning positive and negative values, and a test set of examples whose positive and negative values are considered true, the receiver operating characteristic (ROC) plots the sensitivity, or true positive rate, defined as sensitivity = TP/P versus the false positive rate or 1-specificity, FPR = FP/N, for different thresholds used for classification. The performance of the classifier is evaluated through the area under the curve (AUC) of the ROC plot, which is 0.5 for ran-

dom classifiers and 1 for a perfect classifier having sensitivity one for all thresholds.

Conditional averages

For studying the relationship between different types of divergence measures, we measured the conditional average of one variable conditioned to values of the other variable in a given interval.

ACKNOWLEDGMENTS

The authors is gratefully dedicate this article to Ángel Ramirez Ortiz, who guided their studies of protein structure evolution and classification and has been a great mentor, colleague, and friend. The authors acknowledge interesting discussions with Julián Echave, Florian Teichert, and Markus Porto. This work has been financed through a Ramón y Cajal fellowship to UB and through projects BIO2008-04384 and CSD200623 of the Spanish Ministery of Science and Innovation.

REFERENCES

1. Bromham L, Penny D. The modern molecular clock. Nature Reviews Genetics 2003;4:216–224.
2. Zuckerkandl E, Pauling L. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press; 1962.
3. Kimura M. Evolutionary rate at the molecular level. Nature 1968; 217:624–626.
4. King J-L, Jukes TH. Non-Darwinian evolution. Science 1969;164: 788–798.
5. Ohta T, Kimura M. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1971;1:18–25.
6. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
7. Ohta T. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Pop Biol 1976;10:254–275.
8. Durrett R. Probability models for DNA sequence evolution. New York: Springer; 2002.
9. Sella G, Hirsch AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 2005;102:9541–9546.
10. Bastolla U, Moya A, Viguera E, van Ham RCHJ. Genomic determinants of protein folding thermodynamics. J Mol Biol 2004;343: 1451–1466.
11. Gillespie JH. The causes of molecular evolution. Oxford: Oxford University Press; 1991.
12. Graur D, Li WH. Fundamentals of molecular evolution. Sinauer, Sunderland: Vagaries of the molecular clock; 2000.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
14. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
15. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
16. Grishin NV. Fold change in evolution of protein structures. J Struct Biol 2001;134:167–185.
17. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. Bioinformatics 2005;21:1308–1310.
18. Viksna J, Gilbert D. Assessment of the probabilities for evolutionary structural changes in protein folds Bioinformatics 2007;23:832–841.
19. Pascual-García A, Abia D, Ortiz AR, Bastolla U. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. PLoS Comput Biol 2009;5:e1000331.
20. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
21. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
22. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001; 307:1113–1143.
23. Lecomte JT, Vuletich DA, Lesk AM. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struct Biol 2005;15:290–301.
24. Sangar V, Blankenberg DJ, Altman N, Lesk AM. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 2007;8:294.
25. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. J Mol Biol 2003;326:1–9.
26. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36:307340.
27. Friedberg I. Automated protein function predictionthe genomic challenge. Brief Bioinf 2006;7:225–242.
28. Ponomarenko JV, Bourne PE, Shindyalov IN. Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology. Proteins. 2005;58:855–865.
29. Wang K, Horst JA, Cheng G, Nickle DC, Samudrala1 R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput Biol 2008;4:e1000181.
30. Shakhnovich BE, Max Harvey J. Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. J Mol Biol 2004;337: 933–949.
31. Shakhnovich BE. Improving the precision of the structure-function relationship by considering phylogenetic context. PLoS Comput Biol 2005;1:e9.
32. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
33. Ortiz AR, Strauss C, Olmea O. MAMMOTH (Matching Molecular Models Obtained from Theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
34. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
35. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS 1992;8:275–282.
36. Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford: Oxford University Press; 2000.
37. Bastolla U, Porto M, Roman HE, Vendruscolo M. Statistical properties of neutral evolution. J Mol Evol 2003;57 (Suppl 1):S103–S119.
38. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–220.
39. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet 2000;25:25–29.
40. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, 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 AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res 2009;37 (Database Issue):D211–D215.

41. Clegg JB, Gagnon J. Structure of the zeta chain of human embryonic hemoglobin. Proc Natl Acad Sci USA 1981;78:6076–6080.
42. Giardina B, Messana I, Scatena R, Castagnola M. The multiple functions of hemoglobin. Crit Rev Biochem Mol Biol 1995;30: 165–196.
43. Rammal R, Toulouse G, Virasoro MA. Ultrametricity for physicists Rev Mod Phys 1986;58:765–788.
44. Feng L, Zhou S, Gu L, Gell DA, Mackay JP, Weiss MJ, Gow AJ, Shi Y. Structure of oxidized alpha-haemoglobin bound to AHSP reveals a protective mechanism for haem. Nature 2005;435:697–701.
45. Stenberg K, Clausen T, Lindqvist Y, Macheroux P. Involvement of Tyr24 and Trp108 in substrate binding and substrate specificity of glycolate oxidase. Eur J Biochem 1995;228:408–416.
46. Lustbader JW, Cirilli M, Lin C, Xu HW, Takuma K, Wang N, Caspersen C, Chen X, Pollak S, Chaney M, Trinchese F, Liu S, Gunn-Moore F, Lue LF, Walker DG, Kuppusamy P, Zewier ZL, Arancio O, Stern D, Yan SS, Wu H. ABAD directly links Abeta to mitochondrial toxicity in Alzheimer's disease. Science 2004;304: 448–452.
47. Perozzo R, Kuo M, Sidhu AS, Valiyaveettil JT, Bittman R, Jacobs WR Jr, Fidock DA, Sacchettini JC. Structural elucidation of the specificity of the antibacterial agent triclosan for malarial enoyl acyl carrier protein reductase. J Biol Chem 2002;277:13106–13114.
48. Scheidig AJ, Sanchez-Llorente A, Lautwein A, Pai EF, Corrie JE, Reid GP, Wittinghofer A, Goody RS. Crystallographic studies on p21(H-ras) using the synchrotron Laue method: improvement of crystal quality and monitoring of the GTPase reaction at different time points. Acta Cryst 1994;D50:512–520.
49. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comp Biol 2006;2:e85.
50. David FP, Yip YL. SSMap: A new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/ Swiss-Prot Knowledgebase. BMC Bioinformatics 2008;9:391.
51. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algoithm for multiple structure alignment. Bioinformatics 2005;21: 3255–3263.
52. Wang JD, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007; 23:1274–1281.