J Mol Evol (2006) 63:448–457
DOI: 10.1007/s00239-005-0209-x

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Conservation of Protein Backbone Flexibility

Sandra Maguid,[1] Sebastián Fernández-Alberti,[1] Gustavo Parisi,[1] Julián Echave[1,2]

[1] Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes, Saenz Peña 180, 1876 Bernal, Buenos Aires, Argentina
[2] Instituto Nacional de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Universidad Nacional de La Plata, Suc.4 C.C. 16, 1900 La Plata, Buenos Aires, Argentina

Received: 1 September 2005 / Accepted: 25 May 2006 [*Reviewing Editor:* Dr. David Pollock]

Abstract. Internal protein dynamics is essential for biological function. During evolution, protein divergence is functionally constrained: properties more relevant for function vary more slowly than less important properties. Thus, if protein dynamics is relevant for function, it should be evolutionary conserved. In contrast with the well-studied evolution of protein structure, the evolutionary divergence of protein dynamics has not been addressed systematically before, apart from a few case studies. X-Ray diffraction analysis gives information not only on protein structure but also on B-factors, which characterize the flexibility that results from protein dynamics. Here we study the evolutionary divergence of protein backbone dynamics by comparing the C_α flexibility (B-factor) profiles for a large dataset of homologous proteins classified into families and superfamilies. We show that C_α flexibility profiles diverge slowly, so that they are conserved at family and superfamily levels, even for pairs of proteins with nonsignificant sequence similarity. We also analyze and discuss the correlations among the divergences of flexibility, sequence, and structure.

Key words: Flexibility profiles — Protein dynamics — Protein evolution

Introduction

Evolutionary divergence is functionally constrained, so that properties more relevant for function diverge more slowly (Kimura 1983). Such functional constraints would explain the evolutionary conservation of protein structure, reported almost 20 years ago in the pioneering work of Chothia and Lesk (1986) and confirmed by several other studies (Chothia and Gerstein 1997; Russell et al. 1997; Rost 1997; Wood and Pearson 1999). This work has had an enormous impact on structural biology, being the basis of diverse active research areas such as threading, homology modeling, and structural classifications of proteins. Moreover, structure conservation constrains sequence divergence in predictable ways, which is the basis of recent successful structure-based models of protein evolution, which account for observed amino acid substitution patterns using simple models of structurally constrained protein evolution (see Porto et al. [2005] and Parisi and Echave [2005] and references therein).

Not only the static native structure, but also the internal dynamics is essential for the biological function of proteins (Daniel et al. 2003). A well-known example is myoglobin (Frauenfelder et al. 2003), for which already the first X-ray studies demonstrated that there is no obvious pathway for O_2 to enter Mb (Perutz and Mathews 1966): motions are necessary to open entry and exit channels (Case and Karplus 1979). The relationship between protein function and dynamics makes it relevant to study whether dynamics is evolutionary conserved. In contrast with the many systematic studies on the

Correspondence to: Julián Echave; *email:* jechave@inifta.unlp.edu.ar

evolution of protein structures on large databases of homologous proteins (Chothia and Lesk 1986; Chothia and Gerstein 1997; Russell et al. 1997; Rost 1997; Wood and Pearson 1999), the evolutionary divergence of protein dynamics has not been systematically studied before. The aim of the present work is to address this issue.

There are novel experimental techniques that can be used to probe protein dynamics (Schotte et al. 2003; Bourgeois et al. 2003; Lindorff-Larsen et al. 2005). Such techniques have been applied only to a few cases, which prevents their use for a systematic study of the evolution of protein dynamics as the one intended in the present work. However, X-ray diffraction provides not only the mean positions of the protein's atoms, but also their Debye Waller factors (B-factors). For each nonhydrogen atom j, its B-factor B_j is proportional to the mean square displacement of the atom from its equilibrium position: $B_j = 8\pi^2\langle x_j^2\rangle$. There are different contributions to $\langle x_j^2\rangle$: $\langle x_j^2\rangle = \langle x_j^2\rangle_{TLS} + \langle x_j^2\rangle_V + \langle x_j^2\rangle_{CS}$, where $\langle x_j^2\rangle_{TLS}$ is the "translation, rotation, screwing" contribution, which may come either from motions of the whole molecule or from static displacements of the molecule in the crystal lattice, $\langle x_j^2\rangle_V$ is essentially caused by vibrations faster than 0.1 ps, and $\langle x_j^2\rangle_{CS}$ is due to different conformational substates (Parak 2003). The relative $\langle x_j^2\rangle$ values are mainly determined by the intramolecular contributions $\langle x_j^2\rangle_V + \langle x_j^2\rangle_{CS}$ (Ringe and Petsko 1985). Therefore, the B-factors are an important source of information about protein internal dynamics, providing a map of the flexibility of the ground-state protein conformation (Sternberg et al. 1979; Artymiuk et al. 1979; Frauenfelder et al. 1979; Debrunner and Frauenfelder 1982; Ringe and Petsko 1985). B-Factor flexibility profiles depend on the equilibrium distribution of protein conformations and can be considered fingerprints of protein dynamics. The limitation must be kept in mind, however, that even though B-factors depend on intramolecular dynamics, they do not characterize it fully, because they contain no time-scale information.

Here we report a descriptive statistical analysis of the evolutionary divergence of backbone flexibility profiles on a dataset of protein "ranges" classified into families and superfamilies. As explained below, a "range" is not necessarily the whole protein but that part which is homologous to other ranges of the same family. Ranges are usually single domains, but multidomain families are also included in the dataset. For the sake of clarity, we use "proteins" throughout this paper, but it must be kept in mind that this means ranges.

We compared C_α B-factor profiles, which characterize the backbone flexibility that results from protein dynamics. We found that such flexibility profiles diverge slowly, being conserved at family and superfamily levels, even for pairs of proteins with nonsignificant sequence similarity. We also performed a multivariate analysis of the correlations among the divergences of flexibility, structure, and sequence. We found small but significant correlations, where the contributions of sequence similarity (*Id%*) and structure similarity (*RMSd*) to the variation of p^+ cannot be disentangled. We discuss the possible origin of such correlations and the impossibility, at this stage, of separating physicochemical cause-effect relationships from correlations due to correlations arising from covariation of the parameters considered with divergence time.

Materials and Methods

Dataset of Homologous Pairs

We used the database of structurally aligned homologous proteins HOMSTRAD (Mizuguchi et al. 1998; Stebbings and Mizuguchi 2004). This database consists of 1032 families. It provides combined protein sequence and structure information extracted from the Protein Data Bank. It defines a family as a group of proteins that show clear evidence of common ancestry, as judged by a combination of automatic methods and eye inspection. Rather than full proteins, entries are "ranges." A range is a region of the PDB file that belongs to a family. A range may be composed by noncontiguous fragments and is composed by one or more domains. Each entry within a family is a representative of a number of other PDB chains and is selected on the basis of several criteria, primarily the resolution of the structure. To avoid problems related to redundancy, representatives are chosen so that no two HOMSTRAD entries within a family have an identity higher than 90%.

We grouped HOMSTRAD families into superfamilies using the CAMPASS database of superfamilies (Sowdhamini et al. 1998). If any member of a HOMSTRAD family belonged to a given CAMPASS superfamily, we assigned that CAMPASS superfamily code to all other members of the family. Those families which had no member in CAMPASS could not be assigned to a superfamily and were not used. In this way, we obtained 2834 proteins (ranges) grouped into 866 HOMSTRAD families and into 585 CAMPASS superfamilies.

From the previous dataset, we removed (i) proteins determined by NMR, (ii) proteins whose pdb files had no information on B-factors, and (iii) proteins whose B-factor profiles were too smooth for a significant evaluation of the flexibility similarity measure (see below). Then all pairs of proteins (ranges) with the same superfamily codes were structurally aligned. And those protein pairs which could not be structurally aligned (e.g., because they were too different or because there were missing coordinates in the pdb files) were removed from the dataset.

We were left with a total of 2087 proteins with assigned family and superfamily memberships, aligned into 11,580 protein pairs, classified into 624 families (sets of pairs of proteins of the same family) and 160 superfamilies. Note that here we define a superfamily of pairs as a set of protein pairs that belong to the same superfamily but to different families. In this way any given pair is classified into either a family or a superfamily. Supplementary Tables 1 and 2 include, respectively, the list of families and superfamilies and the number of protein pairs for each case.

Dataset of Nonhomologous Pairs

For statistical assessments, we obtained a reference dataset of 13934 structurally aligned pairs of nonhomologous proteins. Each pair in the dataset was obtained by randomly picking 2 proteins from the set of 2087 proteins described previously and keeping it only if the proteins belonged to different superfamilies (and thus families) and could be structurally aligned by the alignment program.

Structural Alignment

Pair structural alignments were obtained using the program MAMMOTH (Ortiz et al. 2002). For proteins that have in their pdb files more than one conformation, the first conformation was used. To quantify structure dissimilarity, we calculated the root mean square deviation (*RMSd*) between matched C_α.

Sequence Similarity

To quantify sequence similarity, we calculated the percentage identity (*Id%*) between aligned residues (not considering gaps). Usually, the statistical significance of *Id%* is calculated taking as a reference the distribution corresponding to the sequence alignment of random pairs of proteins. This is not appropriate in the present case, because our *Id%* values correspond to structurally aligned, rather than sequence aligned, proteins. Thus, an appropriate reference is the distribution of *Id%* obtained from the dataset of nonhomologous structurally aligned pairs described previously. This distribution has an average of 6.7 and a 99th percentile of 13.7.

Flexibility Similarity

As is usually done, the flexibility of the protein backbone was characterized by the profile of B-factors of α carbons included together with the structure in the PDB files. This is very similar to a profile obtained by averaging the B-factors over the backbone atoms (Smith et al. 2003). A full comparison of the side chains is impossible since we are comparing proteins with different sequences, which prevents the superposition of most side-chain atoms. However, side-chain degrees of freedom are controlled by the C_α positions (Micheletti et al. 2004).

To measure the similarity between the flexibility profiles of two aligned proteins, we calculated the Spearman rank-order correlation coefficient ρ_B between the corresponding Cα B-factor profiles (Halle 2002). Since ρ_B depends on ranks, it is determined by the relative $\langle x_j^2 \rangle$ values, which depend mainly on the protein's internal dynamics (Ringe and Petsko 1985). The range of flexibility similarity is $-1 \leq \rho_B \leq 1$, where perfectly correlated profiles would give $\rho_B = 1$ and uncorrelated profiles $\rho_B \sim 0$. Since B-factors are correlated along the sequence (autocorrelation), the effective sample size in a statistical assessment of ρ_B is smaller than the actual number of sites compared (Bayley and Hammersley 1946). As a result, two profiles cannot be meaningfully compared if the correlation length is larger than the alignment length. For this reason, such cases were removed from the dataset (see above).

Other authors have preferred the Spearman nonparametric rank correlation rather than the Pearson linear correlation coefficient to quantify the similarity between B-factor profiles (Halle 2002; Micheletti et al. 2004). The main reason is that the usual statistical assessment of the Pearson correlation is based on the assumption that the data are normally distributed, whereas the Spearman correlation does not depend on the distribution of B-factors, which is known to be a complex multicomponent distribution (Wampler 1997). Moreover, the rank-order correlation usually provides a more stringent and robust measure than the linear correlation (Micheletti et al. 2004). However, information is lost when data are transformed into ranks and, also, the evolutionary divergence of a similarity measure based on ranks might be less well behaved than a measure based on the actual values. Therefore, we also calculated the Pearson correlation coefficient r_B to measure the similarity of C_α flexibility profiles.

B-Factors are expected to depend on crystal packing. It has been shown that the correlation between B-factors predicted by simple models correlate better with experimental ones when residues involved in crystal contacts are eliminated from the calculation of correlation coefficients (Kundu et al. 2002). Thus, we also calculated ρ_B and r_B after removing sites whose C_α was closer than 7 å from any C_α of a neighbor molecule in the crystal.

Conservation of Flexibility in Sets of Proteins

We calculated the distribution of ρ_B values for the dataset of nonhomologous proteins. Let ρ_B^0 be the median of this distribution. Then for a set of protein pairs (e.g., a family or superfamily) a measure of conservation of flexibility profiles is the fraction of pairs with $\rho_B > \rho_B^0$, which we denote p^+, and is an estimator of $P(\rho_B > \rho_B^0)$ in the set considered. For a set of nonhomologous proteins, by construction we expect $p^+ = 0.5$. Thus, if for a given set of protein pairs $p^+ > 0.5$, we can say that flexibility profiles are conserved. To assess whether $p^+ > 0.5$ and its statistical significance, we used the binomial test. Also, we followed (Vollset 1993) to calculate the binomial confidence intervals. The limits p_*^+ of the confidence interval of level α are the solutions of the quadratic equation $\frac{(p^+ - p_*^+)^2}{p_*^+(1-p_*^+)} = z_{\alpha/2}^2$, where $z_{\alpha/2}$ is the $\alpha/2$-level normal deviate. It has been shown that this interval works better in almost all circumstances than exact intervals, even for the smallest sample sizes (Agresti and Coul 1998).

The analysis described in the previous paragraph was repeated using the Pearson correlation coefficient r_B to quantify the similarity of C_α flexibility profiles.

Correlation Analysis

To analyse the correlation among ρ_B (flexibility similarity), *Id%* (sequence similarity), and *RMSd* (structure dissimilarity), we calculated correlation coefficients $r_{x,y} = \frac{<xy> - <x><y>}{\sqrt{\sigma_x \sigma_y}}$ and part correlation coefficients $r_{x,y(z)} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{yz}^2}}$. Using these, we separated the variance of ρ_B into four contributions (Cohen and Cohen 1983): (i) variance accounted inseparably by *Id%* and *RMSd*, (ii) independent contribution of *Id%*, (iii) independent contribution of *RMSd*, and (iv) fraction of the variance unaccounted by either *RMSd* or *Id%*. We also performed this correlation analysis using the Pearson correlation coefficient r_B instead of ρ_B.

Results and Discussion

We compared the backbone flexibility, as characterized by the C_α B-factor profiles, of a dataset of 2087 proteins classified into homologous families and superfamilies. All pairs of proteins with the same superfamily codes were structurally aligned, leading to 11,580 pair alignments. For each pair alignment we calculated (i) the root mean-square deviation

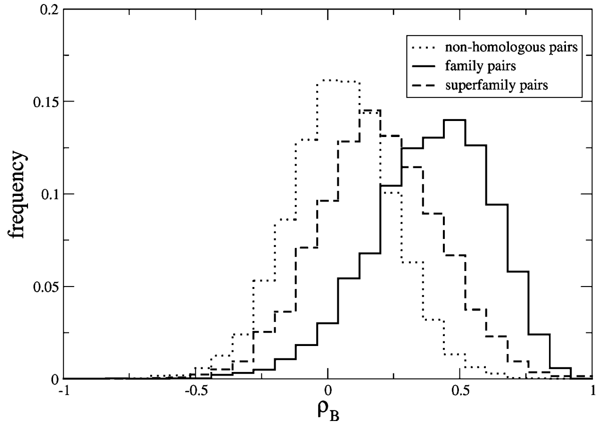

Fig. 1. Distributions of backbone flexibility similarity ρ_B. The three histograms show the frequency distributions of ρ_B for family pairs, superfamily pairs, and a reference set of nonhomologous pairs. There is significant conservation at the family and superfamily levels. Flexibility is more conserved in families than in superfamilies.

between aligned C_α, *RMSd* (structure dissimilarity), (ii) the percentage identical residues, *Id%* (sequence similarity), and (iii) the Spearman rank-order correlation coefficient between the profile of C_α B-factors, ρ_B (backbone flexibility similarity). We divided the 11,580 pairs into two sets according to whether the two proteins of a pair belong to the same family (5362 pairs) or to the same superfamily but different families (6218 pairs).

Distributions of ρ_B*: Homologous vs. Nonhomologous Protein Pairs*

In Fig. 1 we compare the ρ_B distributions obtained for families, superfamilies, and nonhomologous structurally aligned pairs of proteins. The means (standard deviations of the means) of the reference distributions displayed in the figure are as follows: nonhomologous pairs, 0.051 (0.002); superfamily pairs, 0.204 (0.003); and family pairs, 0.390 (0.003). A *t*-test shows that $\langle\rho_B\rangle_{\text{non-homologous}} < \langle\rho_B\rangle_{\text{superfamily}} < \langle\rho_B\rangle_{\text{family}}$ with $P < < 10^{-2}$. Thus, there is significant flexibility conservation in families and superfamilies with respect to the reference dataset of pairs of nonhomologous pairs, and moreover, there is less conservation for homologous pairs of different families (same superfamily) than for the same family, which shows that flexibility diverges.

It is interesting to note that even though $\langle\rho_B\rangle_{\text{non-homologous}} = 0.051$ is small, $\langle\rho_B\rangle_{\text{non-homologous}} > 0$ with significance $p < 10^{-2}$. Thus, the (Spearman) correlation coefficient between the C_a B-factor profiles of the structural alignment of two nonhomologous proteins is expected to be small but positive, in contrast with what is expected from uncorrelated data. This is due to the fact that the structurally aligned sites tend to be structurally similar and that local structure correlates with backbone flexibility (Halle 2002).

Table 1. Conservation of protein flexibility profiles

Set[a]	N^b_{pairs}	p^{+c}	$(p_a^+, p_b^+)^d$
All homologous protein pairs	11,580	0.82	(0.81, 0.83)
Same family	5,362	0.92	(0.91, 0.93)
Same superfamily, different families	6,218	0.74	(0.72, 0.75)
Immunoglobulin superfamily	1,586	0.69	(0.66, 0.72)
Immunoglobulin/V set/ heavy-chain family	190	0.98	(0.93, 0.99)
Immunoglobulin/V set/ light-chain family	253	0.94	(0.89, 0.97)
Globin-like superfamily	468	0.67	(0.62, 0.73)
Globin family	741	0.83	(0.80, 0.87)
Phycocyanin family	66	0.76	(0.60, 0.87)
Similar sequences (*Id%* > 13.7)	6,249	0.92	(0.91, 0.93)
Nonsimilar sequences (*Id%* < 13.7)	5,331	0.71	(0.69, 0.72)

[a] Set of protein pairs.

[b] Total number of protein pairs in the set.

[c] Fraction of pairs with $\rho_B > 0.052$ (the median of the distribution of ρ_B for nonhomologous protein pairs).

[d] Ninety-nine percent binomial confidence interval.

Flexibility Conservation at the Family and Superfamily Levels

The conservation of any group of protein pairs can be characterized by p^+, the fraction of pairs with $\rho_B > \rho_B^0$, where ρ_B^0 is the median of the distribution of ρ_B of nonhomologous pairs. If flexibility profiles are conserved within a group, $p^+ > 0.5$, which can be assessed using a binomial test. The median of the reference dataset is $\rho_B^0 = 0.052$, very close to its mean, since the distribution is nearly symmetric (see Fig. 1).

In Table 1, we show the values of p^+ for different groupings of the protein pairs studied, together with the corresponding 99% binomial confidence intervals. In the first row in Table 1 we find that for the whole dataset of homologous proteins (families + superfamilies), $p^+_{\text{homologous}} = 0.82$: 82% of all the pairs of homologous proteins have ρ_B larger than the median of the reference distribution. A binomial test shows that $p^+_{\text{homologous}} > 0.5$ with significance $P < < 10^{-2}$, which means that protein flexibility profiles of homologous proteins are significantly conserved. Rows 2 and 3 in Table 1 show that $p^+_{\text{family}} = 0.92$ and $p^+_{\text{superfamily}} = 0.74$. A binomial test shows that $p^+_{\text{family}} > p^+_{\text{superfamily}} > 0.5$ with significance $P < < 10^{-2}$. Therefore, in agreement with the analysis based on the means of the ρ_B distributions, C_α flexibility profiles diverge, but depite this divergence, there is a remarkable conservation at both the family and the superfamily levels.

Flexibility Conservation in Individual Families and Superfamilies

To see whether flexibility profiles are conserved in all families and superfamilies or only in some of them, we divided the dataset into 624 families (pairs of proteins of the same family) and 160 superfamilies (pairs of the same superfamily but different families) and we calculated p^+ for each set. Some examples are shown in Table 1, and results for all families and superfamilies are listed in Supplementary Tables 1 and 2. The 99% binomial confidence intervals of p^+ are also reported in these tables. From Table 1 we see, for example, that $\rho_B > \rho_B^0$ for 69% of protein pairs in the Immunoglobulins superfamily, in 98% of pairs in the Immunoglobulin/V set/heavy-chain family, and in 94% of pairs in the Immunoglobulin/V set/light-chain family, the three values significantly larger than 50% with $P < < 10^{-2}$ (binomial test).

When all families and superfamilies are considered, the fact that the number of protein pairs available for each case varies has to be taken into account, since the binomial confidence interval of p^+ for a given group depends on its size. In Fig. 2, we show the frequency distributions of the number of protein pairs in families and superfamilies. It can be seen that there are many families with a few pairs and a few families with many pairs. When all pairs are grouped together into "family" and "superfamily" sets, as in the previous analysis, the concern remains whether the most populated families may be skewing the analysis. Supplementary Tables 1 and 2 show that the p^+ confidence intervals for the families with small numbers of pairs are too wide for a significant assessment of whether $p^+ > 0.5$. Thus, for example, for families which include only one pair of proteins, p^+ is either 0 or 1, and in either case the 99% confidence interval includes $p^+ = 0.5$, so that the null hypothesis of no conservation cannot be discarded at $P \leq 10^{-2}$ levels. However, it is also true that most of the families with one member pair have $p^+ = 1$ rather than 0. A more complete picture can be obtained from Fig. 3, where we show the frequency distribution of p^+ for families and superfamilies together with the fraction of families and superfamilies with p^+ larger than any given cutoff. From this figure we see clearly that most families (95%) and superfamilies (81%) have $p^+ > 0.5$, which confirms the previous findings: backbone flexibility diverges slowly, so that it is conserved at the family and superfamily levels.

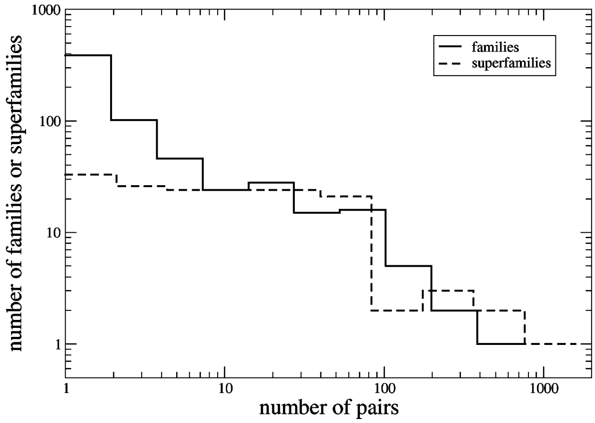

Fig. 2. Distribution of the size of families and superfamilies. This figure shows the number of families and superfamilies as a function of their size (number of pairs). Note that there are in general many families with a few pairs and a few families with many pairs.

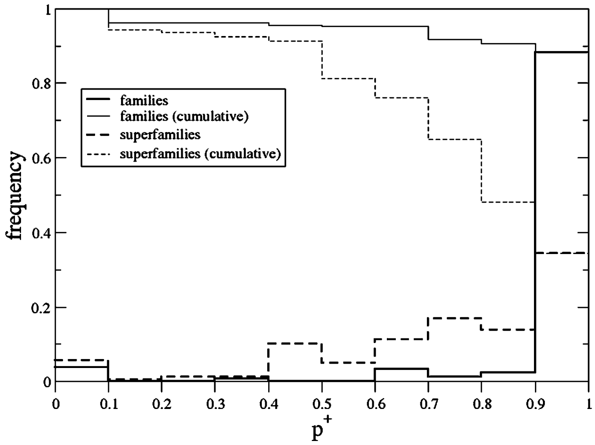

Fig. 3. Conservation of backbone flexibility profiles in families and superfamilies. The histograms show the frequency distributions of the degree of conservation p^+ in families and superfamilies. For each p^+ range we show the normalized number of families (protein pairs which belong to the same family) and superfamilies (protein pairs which belong to the same superfamily but different families) with p^+ values within the range. The figure also shows (thin lines) the total proportion of families and superfamilies with p^+ values larger than the lower limit of the histogram ranges. All families and superfamilies are included.

Flexibility Conservation Beyond Sequence Divergence

It is well known that structure is conserved even between homologous proteins whose sequences have diverged beyond detection of similarity (*Id%* undistinguishable from those of random sequences) (Chothia and Lesk 1986; Rost 1997). Here we investigate whether this is also the case for protein flexibility. To address this issue, we first analysed the distribution of *Id%* for the set of nonhomologous protein pairs. This distribution has a mean of 6.7% and a 99th percentile of 13.7%. Then we divided the whole set of homologous protein pairs into two subsets, pairs with *Id%* > 13.7 (similar sequences) and pairs with $Id\%$ < 13.7 (nonsimilar sequences). We found $p^{+}_{similar} = 0.92$ and $p^{+}_{nonsimilar} = 0.71$ (Table 1, last two rows). A binomial test shows that $p^{+}_{similar} > p^{+}_{nonsimilar} > 0.5$ with $P < < 10^{-2}$, showing that flexibility profiles diverge, but are significantly conserved in both sets. The fact that $p^{+}_{nonsimilar} = 0.71 > 0.5$ means that flexibility profiles are conserved even for protein pairs with sequence identities undistinguishable from those of nonhomologous proteins.

Flexibility Conservation in Relation to Structure and Sequence

Having established the conservation of C_α flexibility profiles, we consider the relationship among the divergences of flexibility, sequence, and structure. In Fig. 2 we plot ρ_B vs. $Id\%$, ρ_B vs. $RMSd$, and $RMSd$ vs. $Id\%$. The Pearson correlation coefficients are $r_{\rho_B,Id\%} = 0.436$, $r_{\rho_B,RMSd} = -0.415$, and $r_{Id\%,RMSd} = -0.872$, all significant at the $P < < 10^{-2}$ level. In an attempt to separate the independent contributions of $Id\%$ and $RMSd$ to the variation of ρ_B, we performed a multivariate correlation analysis. We found that 16.7% of the variance of ρ_B is explained inseparably by $Id\%$ and $RMSd$, the independent contribution of $Id\%$ is 2.3%, the independent contribution of $RMSd$ is 0.5%, and the remaining 80.5% cannot be accounted for by $Id\%$ and $RMSd$.

It is well known that similarity measures, such as $Id\%$ or $RMSd$, obtained after aligning proteins may depend on alignment length (Maiorov and Crippen 1995). Thus, the correlations found before might be, at least partially, the result of a hidden correlation of the three variables with alignment length. In order to eliminate this effect, we calculated the Pearson correlation coefficients of each of these variables with the number of aligned sites. We found r^2 values of 0.004, 2×10^{-6}, and 0.03 for $Id\%$, $RMSd$, and ρ_B, respectively. We then used these values to calculate partial correlations between these variables in which the effect of variation of alignment length was eliminated. These partial correlations have identical values as the correlations reported in the previous paragraph: alignment length does not affect the present analysis.

Methods to predict protein backbone flexibility profiles using either sequences or structures have been developed. Different prediction methods differ slightly but typically the correlation coefficients between predicted and experimental B-factors is about 0.5 for both sequence-based (Yuan et al. 2005) and structure-based methods (Bahar et al. 1997; Halle 2002; Micheletti et al. 2004). Thus, either sequence-based or structure-based methods account for ~25% of the variability of B-factors, which is consistent with our findings. We found that even though there are small independent contributions of sequence ($Id\%$) and structure ($RMSd$) to the variability of ρ_B (flexibility), most of the explainable variation is accounted for inseparably by $Id\%$ and $RMSd$ and cannot be disentangled. Unfortunately, we could find no previous work that considers the contributions of both sequence and structure to B-factors. Further work would be needed to clarify this relationship.

It has been reported that proteins with similar architectures exhibit similar large-scale dynamics (Keskin et al. 2000). Since such slowest large-scale motions determine the backbone flexibility profiles, it is to be expected that proteins with similar structures have similar C_α flexibility profiles. This raises the question whether the observed conservation of backbone flexibility is not just a trivial consequence of structural similarity. A detailed understanding of the relationship between structural similarity and flexibility similarity requires further investigation that goes beyond the scope of the present article. However, we should note that even if structural topology is determined completely from a physicochemical point of view and backbone flexibility, which is not the case (as said before, these prediction methods explain typically ~25% of the variance of B-factors), from an evolutionary point of view, conservation will be determined by natural selection against variations of those aspects of protein physical chemistry most important for function. Thus, it might well be the case that in some cases natural selection acts directly against variations in protein flexibility, which, in turn, would select structures in such a way that flexibility is conserved. Of course, this is speculative at this point, and further work will be required to clarify this issue.

Finally, we would like to comment that regarding the significance of the correlations found in the type of analysis presented here, it is important to keep in mind that the proteins compared result from evolutionary divergence. Therefore, even if the different parameters considered, $Id\%$, $RMSd$, and ρ_B, diverged independently, all would be correlated with divergence time and, therefore, correlated among themselves. The problem with analyzing possible correlations due to covariation of the different variables with divergence time is that this time cannot in general be independently estimated but, rather, it will be inferred, e.g., from measures of sequence dissimilarity. However, despite the difficulty in analyzing this effect, it is important to take into account that the possibility of such an effect implies that a significant correlation does not necessarily imply physicochemical cause-effect relationships. To the best of our knowledge, this issue has not been taken explicitly into account or even commented on in the extended literature on the relationship between sequence similarity and structure similarity. On the contrary, we think most such work is pervaded by the underlying assumption that the correlations found are signals of the degree to which sequence physically determines structure.

Effects of Intermolecule Contacts in Crystals and Multimeric Proteins

Protein motions in the crystal are not necessarily the same as motion in solution. Thus, backbone flexibility profiles might be affected by contacts in the crystal. It has been shown that removing sites involved in crystal contacts may improve (slightly) the agreement between B-factor profiles predicted using simple structure-based methods and experimental profiles (Kundu et al. 2002). To investigate this issue, we repeated our analysis using measures of similarity of C_α flexibility profiles ρ_B calculated after removing those sites involved in crystal contacts. Results (not shown) are almost identical to those obtained without removing these sites, reported above.

Another issue that deserves attention is the effect of intermolecule contacts, such as the intermonomer contacts in multimeric proteins, on B-factor profiles. To investigate this, we first classified the whole set of 11,580 homologous proteins into two sets, one set of 4458 pairs with the same quaternary structure (same number of monomers in both proteins of the pair) and another set of 7122 pairs with different quaternary structures (different number of monomers). We obtained $p^+_{\text{all pairs}} = 0.82$, $p^+_{\text{same quaternary structure}} = 0.88$, and $p^+_{\text{different quaternary structure}} = 0.79$. Therefore, it seems that proteins with the same quaternary structure tend to have more similar C_α B-factor profiles than proteins with different quaternary structures, which would make filtering the database using the "same quaternary structure" criterion worthwhile. However, protein pairs with different quaternary structures tend to be more diverged than those with similar structures, as can be seen by calculating the average *Id%* for these three sets: $\langle Id\% \rangle_{\text{all pairs}} = 22.9$, $\langle Id\% \rangle_{\text{same quaternary structure}} = 28.5$, and $\langle Id\% \rangle_{\text{different quaternary structure}} = 19.4$. Since ρ_B decreases with *Id%* (Fig. 4), p^+ is expected to decrease with $< Id\% >$. To take this into account, we divided each of the three previous datasets (all homologous pairs, pairs with same quaternary structure, and pairs with different quaternary structures) into 10 *Id%* equal bins and calculated p^+ for each case. Plots of p^+ vs. $\langle Id\% \rangle$ for the three datasets (all, same quaternary structure, different quaternary structure) are almost identical (not shown). Thus, the increase in p^+ for the dataset of pairs with the same quaternary structures is just a result of this dataset being less diverged than that of proteins with different quaternary structures. This is consistent with the observation that C_α B-factor profiles result from large-scale slow motions that are not sensitive to quaternary structure (Maguid et al. 2006) and makes it unnecessary to take quaternary structure into account in a coarse-grained analysis such as the one presented here. Of course, quaternary structure and other interprotein contacts and their effects on C_α flexibility should be considered in more detailed case studies.

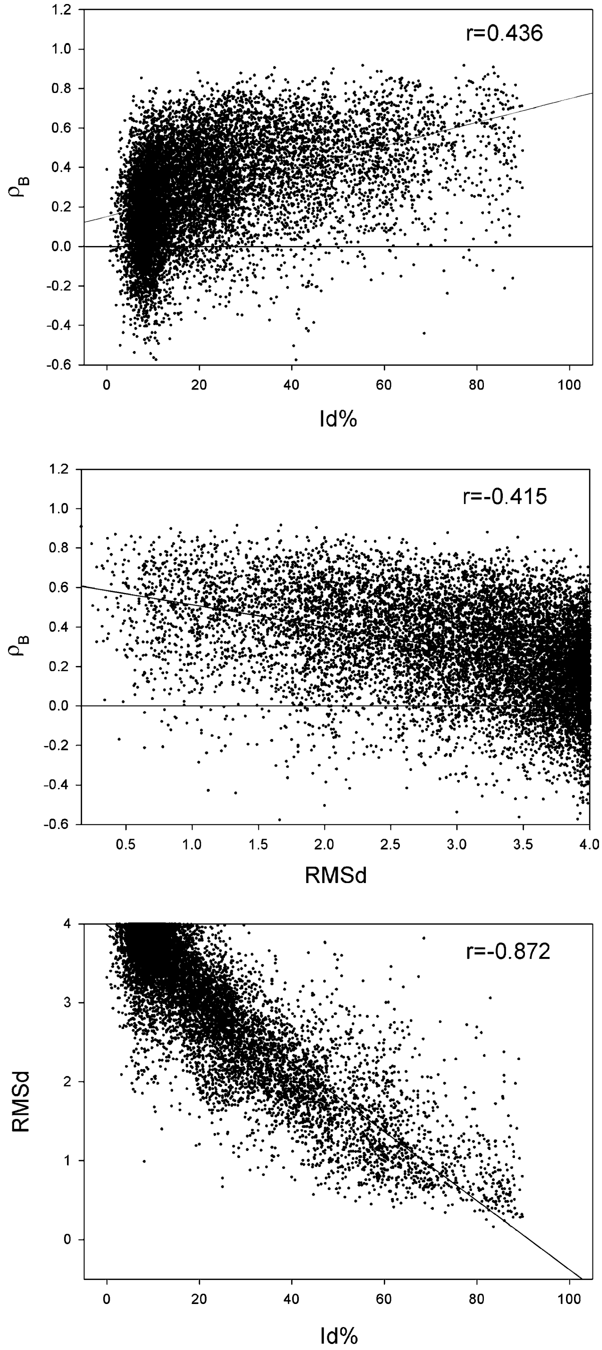

Fig. 4. Relationship among conservations of flexibility, sequence, and structure. **A** Flexibility similarity (ρ_B) vs. sequence similarity (*Id%*). **B** Flexibility similarity (ρ_B) vs. structure dissimilarity (*RMSd*). **C** Structure dissimilarity (*RMSd*) vs. sequence similarity (*Id%*). Linear regression lines are included and linear correlation coefficients are shown in the top right corners.

Unaccounted Variance of Backbone Flexibility Similarity

The unaccounted 80.5% of ρ_B's variance reflects the rather large dispersion of ρ_B found even for constant values of either *Id%* or *RMSd*, which contrasts with the slow decrease in ρ_B with time (Fig. 4, top and middle). This dispersion is due to the sensitivity of B-factor profiles to experimental conditions. To verify this, we calculated ρ_B for 1035 pairs of 46 structures of wild-type sperm whale myoglobin from the SCOP database (Hubbard et al. 1997). These have been determined under different experimental conditions such as temperature, pH, and ligand binding. We found a broad distribution with $\langle \rho_B \rangle = 0.58$ and $\sigma_{\rho_B} = 0.27$, which is similar to the dispersion calculated over the whole dataset of 11,580 protein pairs, $\sigma_{\rho_B} = 0.26$. The high dispersion of ρ_B explains why the correlations between ρ_B and *RMSd* and between ρ_B and *Id%* are lower than the correlation between *RMSd* and *Id%*.

As explained, to avoid redundancy, the HOMSTRAD database includes proteins that represent other similar proteins, so that no two members of a HOMSTRAD family have *Id%* > 90%. The representative included in the database is selected on the basis of several criteria, primarily the resolution of the structure. However, resolution varies between different members of the HOMSTRAD database, which could be one of the reasons behind the large observed unaccounted variation of ρ_B. We investigated the possibility of reducing such variation by filtering our data using resolution of the structures and R-factors. We applied cutoffs to both resolution and R-factors of the proteins included in the analysis to obtain filtered databases and repeated the analysis presented before. We found no difference from the results obtained including all proteins.

Experimental conditions such as pH, ligand binding, temperature, and resolution vary between different proteins of the database, which could result in the large observed unaccounted variation of ρ_B. It is difficult to imagine how these effects could be accounted for in comparisons of different (though homologous) proteins, since these may be different in their ligand-binding properties, optimum pH and temperatures, etc. However, it is significant that despite the large dispersion of ρ_B, there is significant conservation of C_α flexibility profiles, as shown before, which makes our conclusions valid. Of course, it will be interesting to perform more controlled comparisons when more data become available.

Different Measures of Backbone Flexibility Similarity

As explained under Materials and Methods, the Spearman rank-order correlation coefficient is usually preferred to the Pearson correlation coefficient due to its independence on the distribution of B-factors, which is known not to be normal. However, because of its dependence on ranks rather than values, it is possible that the evolutionary divergence of ρ_B is not as well behaved as the value-based Pearson correlation r_B. Besides, a value-based correlation may be sensitive to variations in the C_α flexibility profile that do not affect a rank-based statistic. To investigate this issue, we repeated the whole analysis using r_B. Results are presented as supplementary material.

Supplementary Fig. 1 (analogous to Fig. 1) shows the distributions of r_B for families, superfamilies and nonhomologous protein pairs. The corresponding mean values (standard deviation of the means) are $<r_B> = 0.054$ (0.002), 0.192 (0.003), and 0.376 (0.224) for, respectively, nonhomologous pairs, superfamily pairs, and family pairs. These results are very close to those obtained using ρ_B and confirm that backbone flexibility is conserved at the family and superfamily levels, while it does diverge (more conserved for families than superfamilies).

Supplementary Fig. 2 (analogous to Fig. 3) shows the distribution of p^+, the proportion for sets (families or superfamilies) of protein pairs with r_B larger than the median of the r_B distribution of nonhomologous pairs. As can be seen from this figure, we found that 94% of families and 78% of superfamilies have $p^+ > 0.5$, values that are very close to those obtained previously using the analysis based on ρ_B.

Finally, in Supplementary Fig. 3 we present plots (analogous to Fig. 4) in which we analyze the correlation among r_B, *Id%*, and *RMSd*. These plots are very similar to those shown in Fig. 4. The correlations of r_B with *Id%* and *RMSd* are 0.399 and –0.379, slightly lower than the values found for the correlations of ρ_B. The multivariate correlation analysis produces results similar to those presented before: 1.9% of the variance of r_B is accounted for independently by *Id%*, 0.4% by *RMSd*, and 14% inseparably by *Id%* and *RMSd*.

To summarize, using either the value-based Pearson correlation coefficient r_B or the rank-based Spearman coefficient ρ_B as a measures of similarity of C_α flexibility profiles produces similar results, supporting the conclusions of this work.

Conclusion

We have performed a descriptive statistical analysis of the conservation of backbone flexibility, as characterized by the C_α B-factor, in a large dataset of pair alignments of homologous protein pairs classified into families and superfamilies. A dataset of pair alignments of nonhomologous protein pairs was used as reference. Proteins were structurally aligned, and for each alignment we calculated sequence similarity (*Id%*), structural dissimilarity (*RMSd*), and two measures of backbone flexibility similarity: the Spearman rank-order correlation coefficient between the aligned B-factor profiles, ρ_B, and the Pearson linear correlation coefficient r_B. The correlation among sequence, structure, and flexibility measures of divergence was also analyzed. We found that backbone flexibility is significantly conserved both at family and superfamily levels, with more conservation at the family than the superfamily level. Moreover, we found a correlation among flexibility, sequence, and structure.

The present work has gone beyond the comparison of native structures, by comparing backbone flexibility profiles. These contain information on the equilibrium distribution of protein conformations. Thus, the present results imply the conservation not only of the average structure but also of the dispersion of the equilibrium distribution around the average structure. As discussed in the Introduction, such distribution is determined by the intramolecular dynamics of the protein. Thus, the observed conservation of flexibility profiles provides indirect evidence of the conservation of protein dynamics. However, the backbone flexibility profile contains no information on the time scales of the involved motions. Besides, backbone flexibility is related not only to dynamics but also to other aspects such as protein stability and ligand-induced motions, which could also be subject to natural selection. Therefore, more research will be needed to fully understand the evolutionary variation of protein flexibility and its implications.

Acknowledgments. J.E. acknowledges stimulating discussions with Ugo Bastolla and would like to thank Majed Chergui, Marie Agnès Gassama, and Patricia Esteche for their support. G.P., S.F.A., and J.E. are researchers of CONICET. This work was partially supported by CONICET, ANPCyT, and UNQ.

References

Agresti A, Coul BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 52:119–126

Artymiuk PJ, Blake CCF, Grace DEP, Oatley SJ, Phillips DC, Sternberg MJE (1979) Crystallographic studies of the dynamic properties of Lysozyme. Nature 280:563–568

Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2:173–181

Bayley GV, Hammersley JM (1946) The "effective" number of independent observations in autocorrelated time series. J Roy Stat Soc Suppl 8:184–197

Bourgeois D, Vallone B, Schotte F, Arcovito A, Miele AE, Sciara G, Wulff M, Anfinrud P, Brunori M (2003) Complex landscape of protein structural dynamics unveiled by nanosecond Laue crystallography. Proc Natl Acad Sci USA 100:8704–8709

Case DA, Karplus M (1979) Dynamics of ligand-binding to heme-proteins. J Mol Biol 132:343–368

Chothia C, Gerstein M (1997) Protein evolution—How far can sequences diverge? Nature 385:579–580

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ

Daniel RM, Dunn RV, Finney JL, Smith JC (2003) The role of dynamics in enzyme activity. Annu Rev Biophys Biomol Struct 32:69–92

Debrunner PG, Frauenfelder H (1982) Dynamics of proteins. Annu Rev Phys Chem 33:283–299

Frauenfelder H, Petsko GA, Tsernoglou D (1979) Temperature-dependent x-ray-diffraction as a probe of protein structural dynamics. Nature 280:558–563

Frauenfelder H, Mcmahon BH, Fenimore PW (2003) Myoglobin: the hydrogen atom of biology and a paradigm of complexity. Proc Natl Acad Sci USA 100:8615–8617

Halle B (2002) Flexibility and packing in proteins. Proc Natl Acad Sci USA 99:1274–1279

Hubbard TJP, Murzin AG, Brenner SE, Chothia C (1997) Scop: a structural classification of proteins database. Nucleic Acids Res 25:236–239

Keskin O, Jernigan RL, Bahar I (2000) Proteins with similar architecture exhibit similar large-scale dynamic behaviour. Biophys J 78:2093–2106

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kundu S, Melton JS, Sorensen DC, Phillips GN (2002) Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 83:723–732

Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132

Maguid S, Fernandez Alberti S, Ferrelli L, Echave J (2005) Exploring the common dynamics of homologous proteins. Application to the globin family. Biophys J 89:3–13

Maiorov VN, Crippen GM (1995) Size-independent comparison of protein three-dimensional structures. Proteins 22: 273–283

Micheletti C, Carloni P, Maritan A (2004) Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. Proteins 55:635–645

Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) Homstrad: a database of protein structure alignments for homologous families. Protein Sci 7:2469–2471

Ortiz AR, Strauss CEM, Olmea O (2002) MAMMOTH (Matching Molecular Models Obtained From Theory): an automated method for model comparison. Protein Sci 11:2606–2621

Parak FG (2003) Physical aspects of protein dynamics. Rep Prog Phys 66:103–129

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

Perutz MF, Mathews FS (1966) An x-ray study of Azide Meth-aemoglobin. J Mol Biol 21:199–202

Porto M, Roman HE, Vendruscolo M, Bastolla U (2005) Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. Mol Biol Evol 22:630–638

Ringe D, Petsko GA (1985) Mapping protein dynamics by x-ray-diffraction. Prog Biophys Mol Biol 45:197–235

Rost B (1997) Protein structures sustain evolutionary drift. Fold Des 2:S19–S24

Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269:423–439

Schotte F, Lim M, Jackson TA, Smirnov AV, Soman J, Olson JS, Phillips GN Jr, Wulff M, Anfinrud PA (2003) Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. Science 300:1944–1977

Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G (2003) Improved amino acid flexibility parameters. Protein Sci 12:1060–1072

Sowdhamini R, Burke DF, Huang JF, Mizuguchi K, Nagarajaram HA, Srinivasan N, Steward RE, Blundell TL (1998) Campass: a database of structurally aligned protein superfamilies. Structure 6:1087–1094

Stebbings LA, Mizuguchi K (2004) Homstrad: recent developments of the homologous protein structure alignment database. Nucleic Acids Res 32:D203–D207

Sternberg MJE, Grace DEP, Phillips DC (1979) Dynamic information from protein crystallography—analysis of temperature factors from refinement of the hen egg-white lysozyme structure. J Mol Biol 130:231–252

Vollset SE (1993) Confidence intervals for a binomial proportion. Stat Med 12:809–824

Wampler JE (1997) distribution analysis of the variation of B-factors of x-ray crystal structures: temperature and structural variations in lysozyme. J Chem Inf Comput Sci 37:1171–1180

Wood TC, Pearson WR (1999) Evolution of protein sequences and structures. J Mol Biol 291:977–995

Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B-factor profiles. Proteins 58:905–912