Evolutionary conservation of protein vibrational dynamics

Article *in* Gene · July 2008

DOI: 10.1016/j.gene.2008.06.002 · Source: PubMed

Sebastian Fernandez-Alberti
National University of Quilmes
72 PUBLICATIONS **897** CITATIONS

Julian Echave
National University of General San Martín
113 PUBLICATIONS **1,880** CITATIONS

Gene 422 (2008) 7–13

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Evolutionary conservation of protein vibrational dynamics

Sandra Maguid [a], Sebastian Fernandez-Alberti [a], Julian Echave [b,*]

[a] Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes, Bernal, Argentina
[b] Instituto Nacional de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Universidad Nacional de La Plata, La Plata, Argentina

ARTICLE INFO

Article history:
Available online 6 June 2008

Received by: A. Bernardi

Keywords:
Gaussian Network Model
Normal mode analysis
Flexibility profile

ABSTRACT

The aim of the present work is to study the evolutionary divergence of vibrational protein dynamics. To this end, we used the Gaussian Network Model to perform a systematic analysis of normal mode conservation on a large dataset of proteins classified into homologous sets of family pairs and superfamily pairs. We found that the lowest most collective normal modes are the most conserved ones. More precisely, there is, on average, a linear correlation between normal mode conservation and mode collectivity. These results imply that the previously observed conservation of backbone flexibility (B-factor) profiles is due to the conservation of the most collective modes, which contribute the most to such profiles. We discuss the possible roles of normal mode robustness and natural selection in the determination of the observed behavior. Finally, we draw some practical implications for dynamics-based protein alignment and classification and discuss possible caveats of the present approach.

1. Introduction

Protein flexibility and protein dynamics are widely acknowledged to be crucial for function (Tsai et al., 1999; Daniel et al., 2003; Gunasekaran et al., 2004; Bahar and Rader, 2005; Karplus and Kuriyan, 2005). Therefore it is of fundamental importance to study the evolution of protein dynamics. In a recent study we showed that backbone flexibility profiles diverge slowly, being conserved both at family and superfamily levels, even for homologous protein pairs with seemingly unrelated sequences (Maguid et al., 2006). This has practical implications, such as the use of the similarity between flexibility profiles to detect distant homologues (Pandini et al., 2007).

Since flexibility results from protein motions, the conservation of backbone flexibility profiles provides indirect evidence for the conservation of internal protein dynamics. However, a more detailed analysis is needed to understand the divergence of dynamics. Protein dynamics can be adequately studied by analyzing the vibrational normal modes (Ma, 2005). The slowest and most collective normal modes can be conveniently described by simplified coarse-grained Elastic Network Models, in which the protein is represented as a set of coupled harmonic oscillators (Tirion, 1996; Tirion, 1996; Bahar et al., 1997; Haliloglu et al., 1997; Atilgan et al., 2001; Tama, 2003; Micheletti et al., 2004b; Yang et al., 2005; Tobi and Bahar, 2005; Demirel and Keskin, 2005). Many case studies on single proteins have been performed using Elastic Network Models during the past few years (Bahar and Rader, 2005). In such studies it is usually reported that the lowest, most collective, normal modes are functionally relevant. Only a few studies have addressed the issue of the evolutionary conservation of normal modes (Keskin et al., 2000; Merlino et al., 2003; Maguid et al., 2005). These studies, limited in general to a small set of proteins of the same family or superfamily, have shown that there is a seeming conservation of the lowest collective normal modes. If this was the general case, i.e. if the lowest normal modes were conserved in most protein families, then, it would explain the observed conservation of backbone flexibility profiles, since the lowest modes, being more coherent and of higher amplitude, contribute the most to flexibility profiles. However, case studies are too scarce to generalize, and, as far as we know, no systematic study has been undertaken of the differential evolutionary conservation of normal modes.

The aim of the present work is to investigate what are the general trends of evolutionary divergence of different normal modes. More specifically, we aim to study whether there is any significant relationship between normal mode conservation and collectivity. To address these issues we perform a normal mode analysis based on the Gaussian Network Model (GNM) on a large and diverse dataset of proteins and study how normal mode conservation depends on normal mode index and collectivity. We also explore the conservation of theoretical GNM flexibility profiles, to account for our previously reported evolutionary conservation of experimental flexibility profiles (Maguid et al., 2006). We discuss the possible physical and biological implications of our findings, as well as practical implications for dynamics-based protein alignment, homology detection, and classification. We finish by discussing possible caveats related to the use of the GNM, rather than more detailed methods to analyze protein dynamics.

Abbreviations: GNM, Gaussian Network Model; PDB, Protein Data Base.

* Corresponding author. INIFTA, Suc. 4-C.C. 16, 1900 La Plata, Argentina. Tel.: +54 221 4257430; fax: +54 221 4257291.
E-mail address: jechave@inifta.unlp.edu.ar (J. Echave).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.002

2. Materials and methods

2.1. Normal mode analysis

2.1.1. The Gaussian Network Model (GNM)

The GNM describes the protein as an elastic network of α-carbons linked by springs when they are placed within a cut-off distance r_c (Bahar et al., 1997; Haliloglu et al., 1997). The locations of the α-carbons in the crystallographic structure are considered as the equilibrium positions, about which the atoms fluctuate.

The topology of a network of N nodes (α-carbons) is defined by the $N\times N$ Kirchhoff matrix of contacts $\mathbf{\Gamma}$ with elements:

$$\mathbf{\Gamma}_{ij} = \begin{cases} -1 & i{\neq}j, d_{ij} \leq r_c \\ 0 & i{\neq}j, d_{ij} > r_c \\ -\sum\limits_{k\neq i}\mathbf{\Gamma}_{ik} & i=j \end{cases} \tag{1}$$

where d_{ij} is the distance between the ith and jth α-carbons.

2.1.2. Normal modes

The vibrational normal modes of the protein are the eigenvectors of the Kirchhoff matrix:

$$\mathbf{\Gamma q}_n = \lambda_n \mathbf{q}_n \tag{2}$$

where λ_n is the eigenvalue of normal mode $\mathbf{q}_n$, which is a column vector with N elements q_{in}. The normal modes are normalized so that $||\mathbf{q}_n^2|| = \sum_{i=1}^{N} q_{in}^2 = 1$. Each element q_{in} is the contribution (amplitude) of the ith C_α to normal mode n. The first normal mode corresponds to translation and has eigenvalue $\lambda_0=0$, thus it is left out of the calculations, leaving $N-1$ vibrational normal modes: $q_1, q_2, \ldots q_{N-1}$.

2.1.3. Normal mode collectivity

The degree of collectivity κ_n of normal mode n is a measure of the number of residues which are significantly displaced by this mode. Here we follow (Bruschweiler, 1995) and calculate κ_n as the exponential of the information entropy embedded in $\mathbf{q}_n$:

$$\kappa_n = \frac{1}{N}\exp\left\{-\sum_{i}^{N} q_{in}^2 \log q_{in}^2\right\} \tag{3}$$

where the sum is over the C_α atoms of the protein. It is easy to prove that $\frac{1}{N} \leq \kappa_n \leq 1$. Maximum collectivity, $\kappa=1$, is attained when all the q_{in}^2 are identical, so that all C_α participate in equal proportions. The minimum $\kappa_n = \frac{1}{N}$ is attained when a normal mode displaces only one C_α, in which case q_{in}^2 is 1 for the displaced atom and 0 for the rest.

2.1.4. B-factor profiles

The C_α B-factor profile of the protein can be calculated as a sum of contributions from the $N-1$ internal modes; for the ith C_α:

$$B_i = \frac{8}{3}\pi^2 < \Delta \mathrm{R}_i^2 > = \left(8\pi^2 k_\mathrm{B} T/\gamma\right) \sum_{n=1}^{N-1} 1/\lambda_n q_{in}^2 \tag{4}$$

where k_B is the Boltzmann constant, T is the absolute temperature, and γ is a constant scaling factor. Eq. (4) shows that the Debye–Waller B-factor of an atom (determined in X-ray experiments) is proportional to its square deviation $\Delta\mathbf{R}_i^2$, i.e. to the flexibility of such atom. Therefore we will use equivalently B-factor profile and flexibility profile.

2.1.5. Fitting the GNM parameters

The two GNM parameters are the cut-off r_c (Eq. (1)) and γ (Eq. (4)). Since the latter is constant for different sites it will not affect our dynamical similarity measures, so that we arbitrarily set it to 1. On the other hand, we determine r_c by maximizing the Pearson correlation coefficient r_TE between the theoretical and experimental C_α B-factor profiles. The theoretical profile is obtained using Eq. (4) and the experimental one is available in the PDB file of the protein. For the proteins studied here (see Section 2.4.1), we obtained cut-off values between 5 Å and 13 Å, being about 7 Å for most of the proteins.

To assess the statistical significance of r_TE, for each protein we generated a random set of correlation values $\{r'_\mathrm{TE}\}$ by reshuffling the profiles using an adaptation of the Moving Blocks Bootstrap technique (Kunsch, 1989). We have taken into account the autocorrelation of theoretical and experimental B-factors and we set the autocorrelation length as block size. The normal distribution $\{r'_\mathrm{TE}\}$ was used to calculate a P-value of the correlation coefficient r_TE, to quantify how well the model describes the observed flexibility profile. Only proteins with significant agreement ($P<10^{-2}$) were considered in the final dataset (see Section 2.4.1).

2.2. Comparison of two proteins

Pairs of proteins were structurally aligned and different measures of dynamical similarity were calculated as described next.

2.2.1. Structural alignment

Protein pairs were aligned 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. Only protein pairs with structural *Z-score* above 5, the cut-off recommended by MAMMOTH, were considered in the final dataset (see Section 2.4.1). For further consideration, we used only the “structural core”, as defined by MAMMOTH, which corresponds to aligned sites without gaps within a cut-off RMSd of 4A.

2.2.2. Normal mode similarity

Given two structurally aligned proteins A and B with, with GNM normal modes $\{\mathbf{q}_n^A\}$ and $\{\mathbf{q}_m^B\}$, respectively, we first calculate the overlap matrix $\mathbf{S}^{AB}$ of normal modes:

$$\mathbf{S}_{nm}^{AB} \equiv \frac{\sum\limits_i q_{in}^A q_{im}^B}{\sqrt{\sum\limits_i \left(q_{in}^A\right)^2 \sum\limits_i \left(q_{in}^A\right)^2}}$$

where the sum goes over the set of aligned positions of the structural core. This is the dot product of $\mathbf{q}_n^A$ and $\mathbf{q}_n^B$ projected onto the aligned structural core and renormalized. Then, we reassign the modes of protein B according to their overlaps with the modes of protein A. To this end, we obtain the permutation of columns of $\mathbf{S}^{AB}$ that maximizes its trace. Since the sign of the normal modes is arbitrary, we choose them so that the diagonal elements of the reassigned overlap matrix are all positive.

The diagonal elements of the resorted overlap matrix are a measure of the similarity between corresponding modes of proteins A and B. Thus, after reassignation, S_{nn}^{AB} is a measure of the degree of conservation of mode n. Note that $0 \leq S_{nn}^{AB} \leq 1$.

2.2.3. B-factor profile similarity

The theoretical B-factor profiles, calculated using Eq. (4), of the two proteins of a pair are compared and their similarity is quantified using the Spearman correlation coefficient ρ_T between B-factors of equivalent positions of the aligned structural core. Similarly, we also calculated the Spearman correlation coefficient ρ_E between the experimental profiles obtained from the PDB files. For more details see Maguid et al. (2006).

2.3. Conservation: comparing distributions of similarity measures

To quantify evolutionary conservation, we will compare the distribution of the different similarity measures described in Section

2.2.2 (normal mode overlaps, S_{nn}) and Section 2.2.3 (theoretical, ρ_{T}, and experimental, ρ_{E}, B-factor profile similarities) for three datasets: a set of family pairs, a set of superfamily pairs, and a reference set of random pairs of non-homologous proteins (see Section 2.4). To this end, we use the Kolmogorov–Smirnov statistic (Conover, 1980).

In general, given a similarity measure S (e.g. S_{nn}, ρ_{T}, or ρ_{E}) and two datasets I and II, let $F_{\mathrm{I}}(s)=\frac{N_{\mathrm{I}}(S\geq s)}{N_{\mathrm{I}}}$ and $F_{\mathrm{II}}(s)=\frac{N_{\mathrm{II}}(S\geq s)}{N_{\mathrm{II}}}$ be the empirical complementary cumulative distribution functions of sets I and II, i.e. for any given s the fraction of data points with $S\geq s$. Then to assess whether distribution II leads to (stochastically) larger values of S, we assess $F_{\mathrm{II}}>F_{\mathrm{I}}$ by using the Kolmogorov–Smirnov statistic DKS^{+} = $\max_{s}\left[F_{\mathrm{II}}(s)-F_{\mathrm{I}}(s)\right]$. The significance of the test can be calculated using

$$P=\exp\left[-2N_{\mathrm{eff}}\left(DKS^{+}\right)^{2}\right] \tag{5}$$

where

$$N_{\mathrm{eff}}=\sqrt{\frac{N_{\mathrm{I}}N_{\mathrm{II}}}{N_{\mathrm{I}}+N_{\mathrm{II}}}}.$$

Note that DKS$^{+}\leq 1$. If distribution II is to the right of distribution I (i.e. larger s values) and the distributions do not overlap, we get the maximum DKS$^{+}=1$. If the distributions are identical, on the other hand, we get DKS$^{+}=0$. Thus, the value of DKS^{+} can be seen as a measure the degree in which the property studied is more conserved for dataset II than for dataset I.

2.4. Datasets of protein pairs

2.4.1. Datasets of homologous pairs

We started with the protein dataset used in Maguid et al. (2006), where details can be found. Briefly, we started with the proteins of the HOMSTRAD database of structurally aligned homologous protein families (Mizuguchi et al., 1998; Stebbings and Mizuguchi, 2004), and we used the CAMPASS superfamily classification (Sowdhamini et al., 1998; Bhaduri et al., 2004) to group HOMSTRAD proteins into superfamilies. Then we obtained all the PDB files and removed all proteins that were unsuitable for the analysis of flexibility-profile conservation (e.g. proteins which had no B-factor information, etc).

For each of the previous proteins, we calculated the GNM normal modes and theoretical B-factor profiles and compared them with the experimental B-factor profiles. Then, we removed from the dataset all proteins with theoretical–experimental correlation r_{TE} with $P>10^{-2}$ (see Section 2.1.5).

Next, we built two datasets of homologous protein pairs: (i) *family pairs*: both proteins of the pair belong to the same family; (ii) *superfamily pairs*: both proteins of the pair belong to the same superfamily but to a different family.

Finally, we structurally aligned each protein pair and removed those pairs with structural $Z_{\mathrm{score}}<5$ (see Section 2.2.1). This resulted in retaining 98% of superfamily pairs and 99% of family pairs, which shows the compatibility between HOMSTRAD and CAMPASS classifications with MAMMOTH alignments.

As a result of the whole procedure, we were left with a final protein dataset composed of 1024 proteins classified into 449 families and 272 superfamilies, aligned into 9826 homologous protein pairs classified into *5474 family pairs* and *4352 superfamily pairs*. Since the reassignation procedure of normal modes (Section 2.2.2) is not symmetrical, we included pairs AB and BA into the datasets in order to obtain the frequency distributions of normal mode similarity. However, since these two pairs are not independent, these numbers were divided by two in the calculation of N_{eff} used for the significance calculation of the Kolmogorov–Smirnov statistic (Eq. (5), Section 2.3).

2.4.2. Reference dataset of random non-homologous pairs

For the statistical assessment of whether a certain property is conserved in the previous datasets of homologous pairs, we built a reference dataset of random non-homologous protein pairs. Each pair was obtained by randomly picking 2 proteins from the set of 1024 proteins described in Section 2.4.1, and keeping it only if the proteins were non-homologous (i.e. were not members of the same family or superfamily) and could be structurally aligned. In this way, we obtained a reference dataset of *14,772 random pairs* of structurally aligned non-homologous proteins.

3. Results and discussion

3.1. Normal mode similarity frequency distribution functions

For each protein pair of the datasets considered, we followed the procedure described in Section 2.2.2 to calculate the similarity of each normal mode, measured by S_{nn}^{AB}, the overlap between the normal mode of the first protein of a pair with its corresponding mode of the second protein of the pair. In Fig. 1 we show contour-plots of the frequency distribution functions of such overlaps as a function of normal mode index for the datasets of family, superfamily, and random non-homologous pairs. It can be seen that conservation (i.e. the distribution peaks) decreases with increasing normal mode value for all three datasets. It is also apparent that conservation is higher for the family dataset than for the superfamily dataset, and that it is in both of these cases higher than for the dataset of random pairs. Moreover, the increase of conservation with decreasing normal mode index is more pronounced for families and superfamilies than for the random dataset.

It is important to stress that with respect to a reference distribution of overlaps of random vectors (data not shown), even the dataset of random non-homologous pairs shows a significant conservation of the lowest normal modes. This results from the bias introduced by quantifying normal mode similarity after aligning the structures, projecting the normal modes onto the alignment, and reassigning them in order to maximize their diagonal overlaps (see Sections 2.2.1 and 2.2.2). Therefore, a correct assessment of the conservation of the datasets of homologous proteins requires the *comparison* of their similarity distributions with the similarity distribution of random pairs obtained using the same procedure. It is misleading to use, explicitly or implicitly, the distribution of overlaps of random vectors as a reference. We stress this point, because it is not usually taken into account in studies aimed at studying the similarity of protein dynamics between homologous proteins (Keskin et al., 2000), a common mistake into which have also fell (Maguid et al., 2005).

3.2. Conservation vs. normal mode index

As we said in the last paragraph, the correct assessment of conservation in datasets of family and superfamily pairs must be referred to the set of random non-homologous pairs. Thus, the distributions of Fig. 1 are compared using the Kolmogorov–Smirnov statistic DKS^{+}, which measures the degree of conservation of one distribution with respect to a reference distribution (see Section 2.3). The result of such comparison is shown in Fig. 2, where we plot $\mathrm{DKS}^{+}_{\mathrm{family-random}}$ and $\mathrm{DKS}^{+}_{\mathrm{superfamily-random}}$ as a function of normal mode index. It can be seen that $\mathrm{DKS}^{+}_{\mathrm{family-random}}>\mathrm{DKS}^{+}_{\mathrm{superfamily-random}}>0$ for all normal modes. Moreover, the statistical significance of the displayed DKS^{+} values is $P<10^{-2}$ for all normal modes both for the family and superfamily datasets. Fig. 2 shows that both at family and superfamily levels there is a general trend of decrease of conservation with increasing normal mode index. Since it is usually the case that lowest normal modes are more collective, it is relevant to consider how the observed conservation depends on normal mode collectivity, which is considered next.

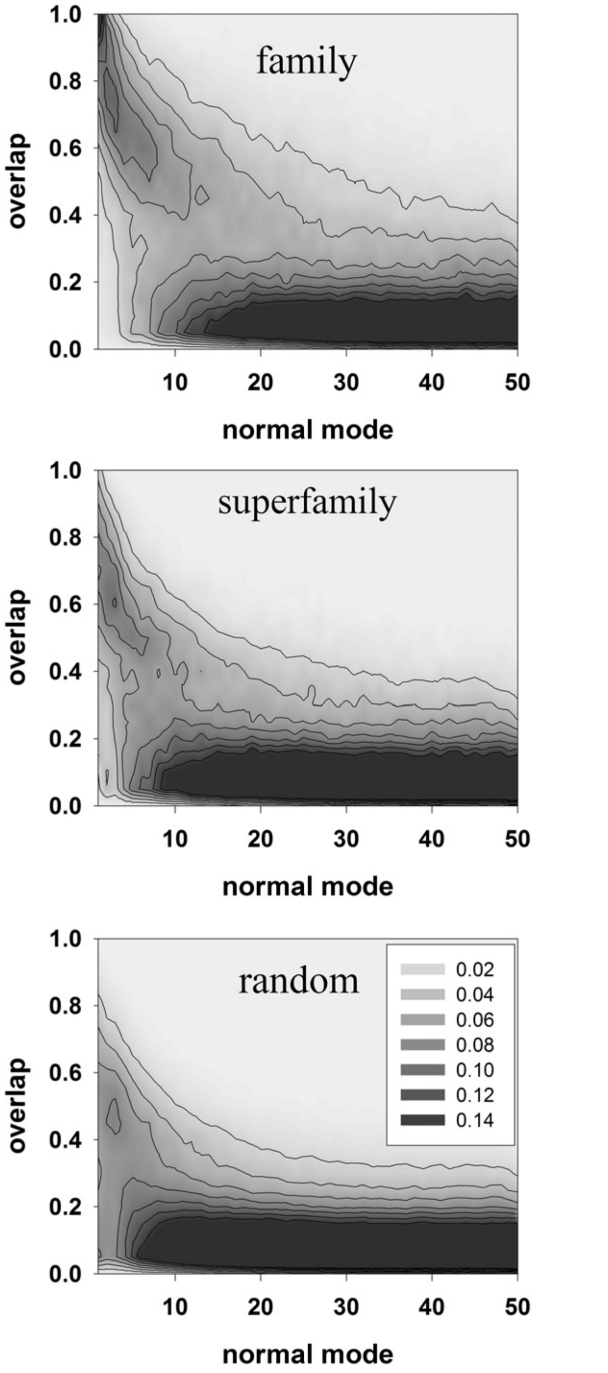

Fig. 1. Distribution of normal mode overlaps vs. normal mode index. Overlaps are calculated as described in Section 2.2.2. Contour plots are color-coded according to the box of the bottom panel. The three panels correspond to family pairs (top), superfamily pairs (middle), and random non-homologous pairs (bottom) (see Section 2.4).

3.3. Conservation vs. collectivity

To study conservation vs. collectivity, we used Eq. (3) to calculate the collectivity κ_n of each normal mode of each of the 1024 proteins of our dataset and averaged them to obtain $\langle\kappa_n\rangle$. Fig. 3 shows this average collectivity vs. normal mode index. It can be seen that collectivity decreases with increasing normal mode. Thus, for mode 1 the average collectivity is $\langle\kappa_1\rangle=0.52$, which means that on average 52% of the protein sites are involved in this mode. Collectivity goes down to less than 25% for the highest normal modes considered ($n=50$).

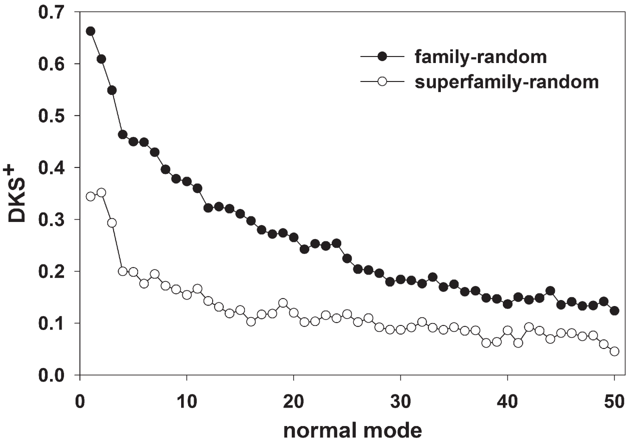

Fig. 2. Conservation vs. normal mode index. Conservation is measured by comparing the distributions of Fig. 1 for families and superfamilies with respect to that of random pairs. The conservation measure is the Kolmogorov–Smirnov statistic DKS$^+$, calculated as described in Section 2.3.

In Fig. 4 we plot the family and superfamily conservation measures, $\mathrm{DKS}^+_{\mathrm{family-random}}$ and $\mathrm{DKS}^+_{\mathrm{superfamily-random}}$, respectively, as a function of average normal mode collectivity, together with linear fits to the data. It can be seen that both at family and superfamily levels, conservation increases linearly with average collectivity, with very significant square correlation coefficients $r^2_{\mathrm{family}}=0.99$ and $r^2_{\mathrm{superfamily}}=0.91$. Thus, on average, normal mode conservation is directly proportional to normal mode collectivity both for families and superfamilies.

3.4. Why does evolutionary conservation depend on mode collectivity?

In the two previous Sections (3.2 and 3.3) we have seen that normal mode conservation decreases with normal mode index (Fig. 2), and increases linearly with normal mode collectivity (Fig. 4) both at family and superfamily levels. This is the most important result of this report. Therefore, it is worthwhile to discuss the possible reasons for such behavior.

We first consider the usual, implicit or explicit, default idea that conservation by itself provides evidence of functional relevance. The underlying assumption is that they are conserved because of selection pressure against the variation of functionally important traits. For our case, such an interpretation would go as follows. The lowest and collective normal modes have frequently been reported in case studies as being functionally important (see (Bahar and Rader, 2005) and

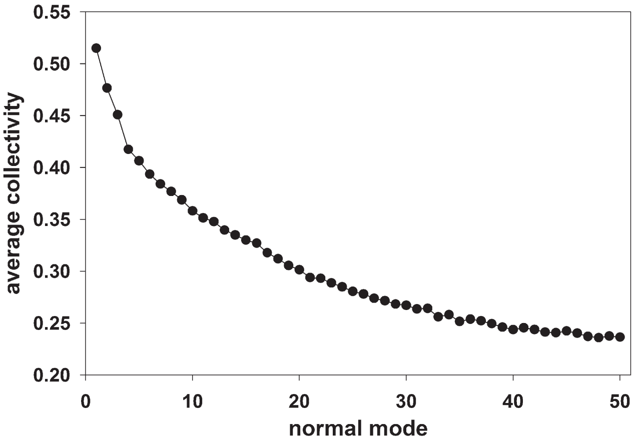

Fig. 3. Average collectivity vs. normal mode index. Average collectivity is calculated as described in Sections 3.3 and 2.1.3.

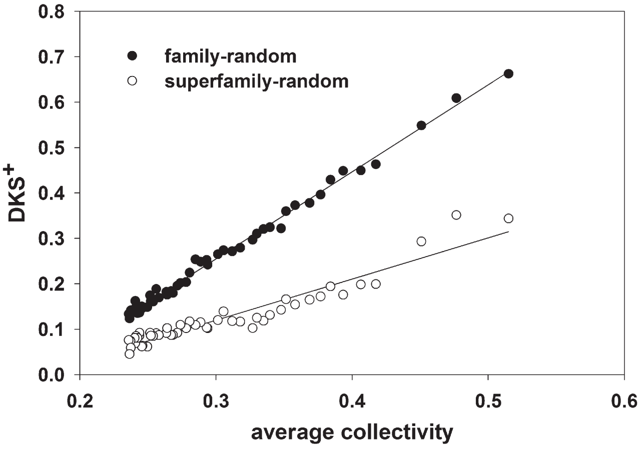

Fig. 4. Conservation vs. average collectivity. Conservation is measured by comparing the distributions of Fig. 1 for families and superfamilies with respect to that of random pairs. The conservation measure is the Kolmogorov–Smirnov statistic DKS$^+$, calculated as described in Section 2.3. The average collectivity is calculated as described in Sections 3.3 and 2.1.3.

references therein). Therefore, if it was the general case that the lowest collective motions were functionally important, according to the neutral theory of evolution (Kimura, 1983), negative selection pressure against functional variation would underlie their higher conservation. One could wonder whether it is likely that the collective modes are the most functionally relevant ones for most proteins. However, this interpretation is tempting since it would conceivably account for our results at family level, where functional diversification is expected to be low. On the other hand, however, at superfamily level we expect higher functional diversification, so that functionally important normal modes should be more divergent, rather than less, due to positive selection. This does not agree with our finding that even at superfamily level the trend is of increased conservation with increased collectivity. Thus, we consider that this scenario is unlikely to account for the general trends reported here.

Another possibility is that the lowest and most collective motions are more conserved just because they are more robust with respect to mutational perturbations. A given normal mode experiences an effective potential energy that is an average over all the sites that contribute to that mode. So, for more collective modes, the potential energy is averaged over more sites, which results in an averaging out of details. Since mutations can be seen as perturbations of the potential parameters, modes that are less sensitive to potential variations will also be more robust with respect to mutational perturbations. In support of this possibility, the robustness of the slowest collective normal modes to variations in the model potential has been shown (Van Wynsberghe and Cui, 2005; Nicolay and Sanejouand, 2006; Zheng et al., 2006). Indeed, it is the very reason behind the adequacy of simplified coarse-grained models, such as the GNM, to describe them. Therefore, we think that the robustness of the most collective normal modes with respect to random perturbations is a more plausible explanation than natural selection of the observed linear relationship between conservation and collectivity both at family and superfamily levels.

We should note, however, that it remains true that the lowest global motions have been proved functionally relevant in several cases. Note, nonetheless, that the scenario proposed in the previous paragraph does not go against the possibility that the lowest modes are functional, as well as robust. Indeed, the robustness of a functional trait may confer selective advantage (Wilke and Adami, 2003; Codoner et al., 2006). Thus, it may well be the case that evolutionary optimization has gone in the direction of selecting proteins for which their function depends on the physically robust normal modes. In other words, since robustness has fitness value, evolution may have allocated functional roles to normal modes which are physically robust.

To summarize, we think that a plausible scenario, consistent with the present findings, is that the more collective normal modes are more conserved because they are more robust, and that the selective advantage of robust traits may have resulted in the evolutionary assignation of functional roles to these modes. We should note, however, that the situation may be even more complex, since once a mode is assigned a function, it will be subject to positive or negative selection pressures. In any case, we must note that these considerations remain speculative, and more work would be needed to fully understand the effects of mutational robustness, natural selection, and their interrelationship on the observed pattern of normal mode conservation.

3.5. Conservation of flexibility profiles

For each protein, we calculated the theoretical flexibility B-factor profile using Eq. (4). Then, as described in Section 2.2.3, for each protein pair we calculated the similarity between theoretical B-factor profiles (ρ_T) and that of the experimental profiles (ρ_E). In Fig. 5 we show the distributions of ρ_T and ρ_E for the family, superfamily, and random datasets. Clearly, the theoretical and experimental distributions are in good qualitative agreement.

The comparison of these distributions using the Kolmogorov–Smirnov statistic is shown in Table 1. All DKS$^+$ values shown have

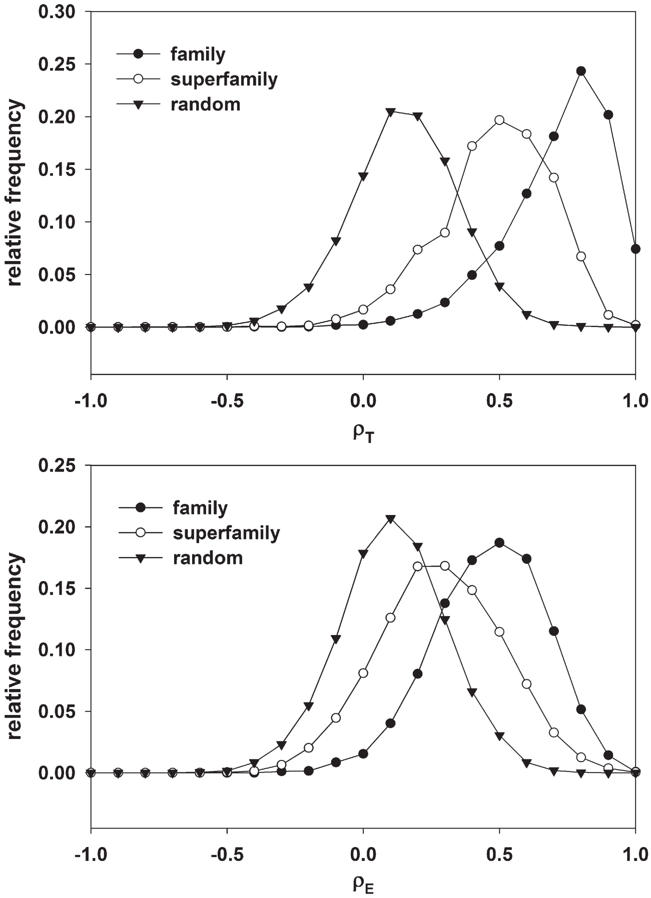

Fig. 5. Distributions of backbone flexibility similarities. Similarity measures are calculated as described in Section 2.2.3. The top panel shows the conservation of theoretical flexibility profiles obtained using the GNM (see Section 2.1.4) and the bottom panel shows the conservation of experimental flexibility profiles. The distributions for family, superfamily, and random pairs are shown.

Table 1
Kolmogorov–Smirnov measures of conservation

Similarity measure	$DKS^{+}_{family\text{-}random}$	$DKS^{+}_{superfamily\text{-}random}$	$DKS^{+}_{family\text{-}superfamily}$
ρ_T	0.85	0.63	0.45
ρ_E	0.63	0.32	0.33
Overlap mode 1	0.66	0.34	0.37
Overlap mode 2	0.61	0.35	0.31
Overlap mode 3	0.55	0.29	0.29

significance $P \ll 10^{-2}$ (Eq. (5)). Table 1 shows that for both ρ_T and ρ_E, $DKS^{+}_{family-random} > DKS^{+}_{superfamily-random} > 0$ and $DKS^{+}_{superfamily-family} > 0$. Thus conservation is significant for both families and superfamilies and it is larger for families than for superfamilies.

It can also be seen that DKS^{+} for the ρ_T distributions are larger than those of the ρ_E distributions. This is due to the fact that the experimental B-factor profiles are very sensitive to experimental conditions (Maguid et al., 2006). In contrast, theoretical profiles depend, via the GNM, on the structure, which is much less sensitive to experimental conditions. In spite of these differences, it is clear that the GNM accounts for our previously reported conservation of flexibility profiles (Maguid et al., 2006). Moreover, it shows that this conservation results from the observed trend of higher conservation of the lowest normal modes (Sections 3.2 and 3.3), which are the ones that contribute the most to the B-factor profile, in accord with Eq. (4), where the contribution of each mode is weighted by the inverse of its eigenvalue. This was not obvious *a priori*, since even a mode-independent conservation could also have resulted in conserved B-factor profiles.

The fact that DKS^{+} for the ρ_T distributions are larger than those of the ρ_E distributions (Table 1) also means that ρ_T is a more reliable measure of dynamical similarity than ρ_E. This has been noted in a recent article that uses the similarity between simulated flexibility profiles rather than the experimental ones for detection of remote homologous proteins (Pandini et al., 2007). The success of this homology detection method would be due to the observed conservation of flexibility profiles reported previously (Maguid et al., 2006), which, as shown here, results from the higher conservation of the lowest normal modes.

Finally, Table 1 also includes the DKS^{+} values for the first three normal modes, which are the most conserved. It can be seen that the DKS^{+} for ρ_T are larger than the DKS^{+} of any of the individual normal modes, so that ρ_T is a better measure of dynamical similarity than the overlap of any single normal mode. The reason for this is that the B-factor profile includes information of the conservation of all the normal modes conveniently weighted (see Eq. (4)). Thus, a dynamics-based protein alignment (or classification) based on comparing theoretical flexibility profiles should lead to better results than alignments based on single normal modes such as suggested in (Keskin et al., 2000).

3.6. Possible caveats of the present approach

3.6.1. Validity of the Gaussian Network Model

The Gaussian Network Model is isotropic, so that it allows the calculation of overall fluctuation patterns and correlation matrices, but doest not take into account the directionality of such fluctuations. More detailed Elastic Network Models take into account the anisotropic character of fluctuations (Doruker et al., 2000; Atilgan et al., 2001; Micheletti et al., 2004a; Zheng and Brooks, 2005; Maragakis and Karplus, 2005; Kondrashov et al., 2006; Jang et al., 2006; Ming and Bruschweiler, 2006). GNM and the Anisotropic Network Model (ANM) were thoroughly investigated with respect to Molecular Dynamics simulations and both proved to be in good agreement giving very good results both for the pattern of fluctuations of individual sites and for the correlation between different sites. Other coarse-grained elastic network models were assessed against MD simulations to find that ENM gives very good results for the lowest normal modes, B-factor profiles, and correlation matrices (Micheletti et al., 2004a; Rueda et al., 2007). Moreover, ENMs have been assessed not only against MD simulations and X-ray data, but also account for backbone flexibility quantified by NMR order parameters (Ming and Bruschweiler, 2006). Despite its relative simplicity the GNM used here also proved to give excellent agreement when compared with NMR data (Yang et al., 2007). This is consistent with the observation that NMR order parameters for the backbone can be very well reproduced by contact matrix and Normal Mode Analysis methods (Zoete et al., 2002; Zhang and Bruschweiler, 2002; Best et al., 2006). Moreover, the agreement between ENMs and MD simulations was found to be independent of protein family, protein length, and of the presence/absence of sulfide bonds or salt bridges (Rueda et al., 2007).

To summarize, ENMs in general, and GNM in particular are found to be in excellent agreement with X-ray and experimental data for backbone motion, as well as with Normal Mode Analysis (NMA) and Molecular Dynamics simulation based on more complex full-atom potentials. Thus, the GNM is an appropriate tool for the purposes of the present study.

3.6.2. Relationship between structure and dynamics

Finally, since the GNM depends exclusively on the protein structure, one could say that the reason behind the observed dynamical similarity (of each normal mode and of the B-factor profiles) is the structural similarity between the aligned structural cores. This rises three issues, discussed next.

The first issue is whether the present results are model-dependent. In response to this, we note that, as discussed in the previous section, ENMs in general, and GNM in particular, are in very good agreement with observed (X-ray or NMR) backbone dynamics and with MD simulations or NMA based on full-atom potentials. This means that the fact that protein structure determines dynamics is not an artifact of the model used, but, rather, is a model-independent result.

The second issue is: since protein structure determines dynamics, is it not tautological to find dynamical similarity on a dataset of proteins classified by their structural similarity? In answer to this, we should note that structural similarity is based on overall measures such as the RMSd, or related, that do not take dynamics explicitly into account. Structure determines dynamics via its topology, as quantified by its contact matrix, in a rather complex way, so that it is *a priori* not obvious that structural similarity should imply similarity of protein dynamics. Moreover, even if structural similarity was a guarantee of dynamical similarity, this does not make studying evolution of protein dynamics trivial. Even if, physically, protein structure determines dynamics, dynamics is necessary for proteins to function. Thus, from a biological point of view, natural selection may act not only directly at structure level (e.g. to conserve the structure of the active site) but also at dynamics level (e.g. to conserve a given functionally important motion, or active site flexibility). Thus, in order to understand structural conservation further studies will be needed of the functional importance of protein dynamics and flexibility, since selection pressures onto these properties may result in constrained structural divergence, just as selection against structural variation has been shown to constrain sequence divergence (Parisi and Echave, 2001; Fornasari et al., 2002; Bastolla et al., 2006). This view that dynamics may be constraining structural divergence, agrees with a view expressed recently: "a hidden flexibility code has been printed by evolution in the structure of biological macromolecules in order to optimize their biological function" (Rueda et al., 2007).

Finally, we should stress that the main point of this paper is to report the differential conservation of normal modes. Thus, even if it was trivial that two structurally similar proteins should be expected to be dynamically similar, as measured by global quantities such as the similarity of flexibility profiles, the findings of this report, that there is a *differential conservation of the lowest normal modes* and that, on average, normal modes are conserved in proportion to their collectivity, are far from being obvious *a priori*. One could oversimplify the problem and argue that for similar structures *all* modes will be conserved, which we found not to be the case.

3.7. Conclusion

We have studied the evolutionary conservation of protein dynamics by calculating the conservation of vibrational normal modes obtained using the Gaussian Network Model for a large dataset of homologous proteins classified into families and superfamilies, compared with a reference dataset of random non-homologous protein pairs.

We found that, on average, normal mode conservation increases linearly with collectivity, so that the slowest most collective modes are the most conserved ones.

We showed that it is this conservation of the collective normal modes, together with their high weights in determining the overall flexibility B-factor profiles, the reason behind the conservation of backbone flexibility.

We speculate that a likely scenario is that most collective normal modes are more conserved because they are more robust with respect to mutational perturbations, rather than because of negative selection against divergence of functionally important modes. On the other hand, positive selection may have a role in assigning functions to these modes, since then their natural robustness may have fitness value. This issue is of utmost importance and it will require much more research.

Then, we briefly discussed how the present results support that for practical purposes such as alignment, homology detection, or classification based on dynamical similarities, it seems better to use the similarity of theoretically predicted B-factor profiles rather than either that of individual normal modes or of experimental B-factor profiles.

Finally, we discussed possible caveats related to using the simplified GNM and the relationship between structural and dynamical conservation.

Acknowledgments

The authors wish to thank G. Parisi for his help in setting up the protein datasets. SFA and JE are researchers of CONICET. This work was partially supported by CONICET, ANPCyT, and UNQ.

References

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., Bahar, I., 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. 80, 505–515.

Bahar, I., Rader, A.J., 2005. Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol. 15, 586–592.

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

Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M., 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the protein data bank. BMC Evol. Biol. 6, 43.

Best, R.B., Lindorff-Larsen, K., Depristo, M.A., Vendruscolo, M., 2006. Relation between native ensembles and experimental structures of proteins. Proc. Natl. Acad. Sci. U. S. A. 103, 10901–10906.

Bhaduri, A., Pugalenthi, G., Sowdhamini, R., 2004. Pass2: an automated database of protein alignments organised as structural superfamilies. BMC Bioinformatics 5, 35.

Bruschweiler, R., 1995. Collective protein dynamics and nuclear-spin relaxation. J. Chem. Phys. 102, 3396–3403.

Codoner, F.M., Daros, J.A., Sole, R.V., Elena, S.F., 2006. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. PLoS Pathog. 2, 1187–1193.

Conover, W.J., 1980. Practical Nonparametric Statistics. Wiley and Sons, New York.

Daniel, R.M., Dunn, R.V., Finney, J.L., Smith, J.C., 2003. The role of dynamics in enzyme activity. Annu. Rev. Biophys. Biomol. Struct. 32, 69–92.

Demirel, M.C., Keskin, O., 2005. Protein interactions and fluctuations in a proteomic network using an elastic network model. J. Biomol. Struct. Dyn. 272, 381–386.

Doruker, P., Atilgan, A.R., Bahar, I., 2000. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins, Struct., Funct., Genet. 40, 512–524.

Fornasari, M.S., Parisi, G., Echave, J., 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol. Biol. Evol. 19, 352–356.

Gunasekaran, K., Ma, B., Nussinov, R., 2004. Is allostery an intrinsic property of all dynamic proteins? Proteins 57, 433–443.

Haliloglu, T., Bahar, I., Erman, B., 1997. Gaussian dynamics of folded proteins. Phys. Rev. Lett. 79, 3090–3093.

Jang, Y.H., Jeong, J.I., Kim, M.K., 2006. Umms: constrained harmonic and anharmonic analyses of macromolecules based on elastic network models. Nucleic Acids Res. 34, W57–W62.

Karplus, M., Kuriyan, J., 2005. Molecular dynamics and protein function. Proc. Natl. Acad. Sci. U. S. A. 102, 6679–6685.

Keskin, O., Jernigan, R.L., Bahar, I., 2000. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys. J. 78, 2093–2106.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kondrashov, D.A., Cui, Q.A., Phillips, G.N., 2006. Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. Biophys. J. 91, 2760–2767.

Kunsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. Ann. Stat. 17, 1217–1241.

Ma, J.P., 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 13, 373–380.

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.

Maguid, S., Fernandez-Alberti, S., Parisi, G., Echave, J., 2006. Evolutionary conservation of protein backbone flexibility. J. Mol. Evol. 63, 448–457.

Maragakis, P., Karplus, M., 2005. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. J. Mol. Biol. 352, 807–822.

Merlino, A., Vitagliano, L., Ceruso, M.A., Mazzarella, L., 2003. Subtle functional collective motions in pancreatic-like ribonucleases: from ribonuclease a to angiogenin. Proteins: Struct., Funct., Genet. 53, 101–110.

Micheletti, C., Carloni, P., Maritan, A., 2004a. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins: Struct., Funct., Bioinformatics 55, 635–645.

Micheletti, C., Carloni, P., Maritan, A., 2004b. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins, Struct., Funct. Bioinformatics 55, 635–645.

Ming, D.M., Bruschweiler, R., 2006. Reorientational contact-weighted elastic network model for the prediction of protein dynamics: comparison with NMR relaxation. Biophys. J. 90, 3382–3388.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Overington, J.P., 1998. Homstrad: a database of protein structure alignments for homologous families. Protein Sci. 7, 2469–2471.

Nicolay, S., Sanejouand, Y.H., 2006. Functional modes of proteins are among the most robust. Phys. Rev. Lett. 96, 078104.

Ortiz, A.R., Strauss, C.E.M., Olmea, O., 2002. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 11, 2606–2621.

Pandini, A., Mauri, G., Bordogna, A., Bonati, L., 2007. Detecting similarities among distant homologous proteins by comparison of domain flexibilities. Protein Eng. Des. Sel. 20, 285–299.

Parisi, G., Echave, J., 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18, 750–756.

Rueda, M., Chacon, P., Orozco, M., 2007. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. Structure 15, 565–575.

Sowdhamini, R., et al., 1998. Campass: a database of structurally aligned protein superfamilies. Structure 6, 1087–1094.

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

Tama, F., 2003. Normal mode analysis with simplified models to investigate the global dynamics of biological systems. Prot. Peptide Letters 10, 119–132.

Tirion, M.M., 1996. Large amplitude elastic motions in proteins from a single-parameter atomic analysis. Phys. Rev. Lett. 77, 1905–1908.

Tobi, D., Bahar, I., 2005. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc. Natl. Acad. Sci. U. S. A. 102, 18908–18913.

Tsai, C.J., Kumar, S., Ma, B.Y., Nussinov, R., 1999. Folding funnels, binding funnels, and protein function. Protein Sci. 8, 1181–1190.

Van Wynsberghe, A.W., Cui, Q., 2005. Comparison of mode analyses at different resolutions applied to nucleic acid systems. Biophys. J. 89, 2939–2949.

Wilke, C.O., Adami, C., 2003. Evolution of mutational robustness. Mutat. Res. 522, 3–11.

Yang, L.W., et al., 2005. Ignm: a database of protein functional motions based on Gaussian network model. Bioinformatics 21, 2978–2987.

Yang, L.W., Eyal, E., Chennubhotla, C., Jee, J., Gronenborn, A.M., Bahar, I., 2007. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. Structure 15, 741–749.

Zhang, F.L., Bruschweiler, R., 2002. Contact model for the prediction of NMR N-H order parameters in globular proteins. J. Am. Chem. Soc. 124, 12654–12655.

Zheng, W.J., Brooks, B.R., 2005. Normal-modes-based prediction of protein conformational changes guided by distance constraints. Biophys. J. 88, 3109–3117.

Zheng, W.J., Brooks, B.R., Thirumalai, D., 2006. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc. Natl. Acad. Sci. U. S. A. 103, 7664–7669.

Zoete, V., Michielin, O., Karplus, M., 2002. Relation between sequence and structure of Hiv-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J. Mol. Biol. 315, 21–52.