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:

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j, d_{ij} \leq r_c \\ 0 & i \neq j, d_{ij} > r_c \\ -\sum_{k \neq i} \Gamma_{ik} & i = j \end{cases} \quad (1)$$

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:

$$\mathbf{\Gamma q}_n = \lambda_n \mathbf{q}_n \quad (2)$$

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 $\mathbf{q}_n$:

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

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:

$$B_i = \frac{8}{3}\pi^2 < \Delta \mathbf{R}_i^2 > = (8\pi^2 k_B T/\gamma) \sum_{n=1}^{N-1} 1/\lambda_n q_{in}^2 \quad (4)$$

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 $\{r'_{TE}\}$ 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 $\{r'_{TE}\}$ 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:

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

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