Pure Appl. Chem., Vol. 84, No. 9, pp. 1931–1937, 2012.
http://dx.doi.org/10.1351/PAC-CON-12-02-15

Why are the low-energy protein normal modes evolutionarily conserved?*

Julian Echave

Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina

Abstract: Proteins fluctuate, and such fluctuations are functionally important. As with any functionally relevant trait, it is interesting to study how fluctuations change during evolution. In contrast with sequence and structure, the study of the evolution of protein motions is much more recent. Yet, it has been shown that the overall fluctuation pattern is evolutionarily conserved. Moreover, the lowest-energy normal modes have been found to be the most conserved. The reasons behind such a differential conservation have not been explicitly studied. There are two limiting explanations. A "biological" explanation is that because such modes are functional, there is natural selection pressure against their variation. An alternative "physical" explanation is that the lowest-energy normal modes may be more conserved because they are just more robust with respect to random mutations. To investigate this issue, I studied a set of globin-like proteins using a perturbed elastic network model (ENM) of the effect of random mutations on normal modes. I show that the conservation predicted by the model is in excellent agreement with observations. These results support the physical explanation: the lowest normal modes are more conserved because they are more robust.

Keywords: chemical physics; computer modeling; molecular dynamics; normal modes; protein dynamics; protein evolution.

INTRODUCTION

Protein motions are important for function. Typical examples are the required flexibility of binding sites and the large conformational transitions necessary for allosteric activation. As with any other functional trait, it is important to study how protein motions change during biological evolution. In contrast with protein sequence and structure, the evolution of dynamics has been much less studied. However, there has been significant recent progress.

The evolution of the overall pattern of protein flexibility is well studied in the case of adaptation to extreme environments (see ref. [1] and references therein). Outside this domain, only recently backbone flexibility, as conveyed by B-factor profiles, has been used to perform systematic studies, which have shown that flexibility diverges slowly so that it is significantly conserved at family and superfamily levels [2,3].

Beyond comparative studies of overall flexibility, it is important to investigate individual motions in more detail. The standard way of analyzing protein motions uses normal modes, the coordinates describing the independent intrinsic vibrations. Each mode has an associated energy and amplitude, which are related (the square amplitude is the inverse of the energy). Normal modes can be obtained in

Pure Appl. Chem.* **84, 1837–1937 (2012). A collection of invited papers based on presentations on the Chemistry of Life theme at the 43rd IUPAC Congress, San Juan, Puerto Rico, 30 July–7 August 2011.

different ways, from diagonalizing the Hessian of the all-atom potentials used in molecular dynamics (MD) simulations, to using coarse-grained elastic network models (ENMs) which model the protein as a network of nodes connected by springs. For our purpose, we highlight that all methods give very similar results, especially for the low-energy large-amplitude motions, which are the most interesting [4,5]. Here we will use an ENM to calculate the normal modes.

In several case studies, low-energy normal modes have been found to have functional value [6]. A remarkable result is that functional transitions between ligand-free and ligand-bound conformations of allosteric proteins can usually be described using one or a few low-energy normal modes. This functional importance prompted studies of evolutionary conservation of normal modes. It has been shown that low-energy normal modes are evolutionarily conserved in several case studies [7–9]. A systematic study of a large dataset of proteins representative of all structural classes and folds shows that this is a general trend: the low-energy large-amplitude normal modes are the most evolutionarily conserved [10].

The present work aims to investigate the reasons behind the observed higher conservation of the lowest-energy normal modes. Are the low-energy normal modes conserved because they are functionally important or is there an alternative explanation? Most case studies mentioned before assume, explicitly or implicitly, the functional interpretation. For example, some studies compare the divergence of sequence or structure with that of motions and connect this to functional aspects [11,12]. However, similarity of low-energy normal modes has been found also for structurally similar but functionally dissimilar proteins, such as for non-homologous proteins with the same architecture [13] or even for completely unrelated proteins [10]. To account for this, an alternative explanation has been proposed: the main reason behind such conservation could be that the low-energy normal modes are just more robust with respect to mutations [10]. Even though one should not discard a role of functional constraints on normal-mode conservation, an adequate null model should take into account the expected variation under random mutations with no selective constraints.

This paper focuses on whether the higher conservation of the low-energy normal modes is due to natural selection against their variation because they are functionally important or to their robustness with respect to unselected random mutations. To investigate this issue, I use the linearly forced elastic network model (LFENM), which models the effect of random mutations on protein structure [14,15], to generate mutant structures and study the variation of their normal modes and compare the model predictions with the observed normal-mode variability for evolved proteins.

MATERIALS AND METHODS

Myoglobin and relatives

I will use the three datasets of proteins used to study the role of normal modes on evolutionary structural divergence in refs. [14,15]. All proteins are related to the sperm-whale myoglobin with Protein Data Bank (PDB) code 1a6m. The three sets are:

1) The "globin-like" dataset, which consists of 23 members of the globin-like superfamily of evolutionarily related proteins, according to the SCOP classification [16].
2) The "myoglobin variants" dataset, composed by 1a6m and other 185 PDB files that correspond to sperm-whale myoglobins including engineered mutants and alternative structures of the wild-type determined under different experimental conditions: different ligand, pH, and/or temperature.
3) The "LFENM" dataset, composed by 1500 structures generated using the linearly forced elastic network model (see below).

A detailed list of the proteins included in the first two datasets can be found in [15].

Elastic network model

The ENM represents the protein as a set of nodes centered at the alpha carbons connected by springs. The potential energy is of the form

$$V(\mathbf{r}) = \frac{1}{2}\left(\mathbf{r} - \mathbf{r}_0\right)^T \mathbf{K}\left(\mathbf{r} - \mathbf{r}_0\right) \tag{1}$$

where $\mathbf{r}$ is the position vector of a given conformation of the protein's C_α, $\mathbf{r}_0$ is the equilibrium (native) conformation, and $\mathbf{K}$ is the stiffness matrix of the network of oscillators, which depends on the force constants of the springs connecting nodes. There are a number of ENMs. Here we use the beta Gaussian model [17].

Normal-mode analysis

Normal modes are the set of coordinates that uncouple the potential(1). They are obtained by solving the eigenvalue problem

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

where the eigenvectors $\mathbf{q}_n$ are the normal modes and the eigenvalues λ_n represent the energies needed to deform the protein along the normal mode directions. Thus, the lowest eigenvalues correspond to the low-energy directions of deformation. The frequency of oscillations along mode n is proportional to λ_n, and its amplitude is proportional to $1/\sqrt{\lambda_n}$. Therefore, the lowest normal modes represent the slow, large-amplitude, low-energy motions of the protein.

The lowest 6 normal modes have 0 eigenvalue. They represent translations and rotations of the whole protein. The modes with non-zero eigenvalue represent internal motions and are the ones that are going to be considered here. They are numbered $n = 1,2,\ldots 3N - 6$, N being the number of nodes of the elastic network.

Normal-mode conservation

In order to study the evolutionary conservation of normal modes, we need to define and quantify normal-mode similarity. Let us assume that we have two proteins A and B that are aligned, so that there is a one-to-one correspondence between a subset of sites of protein A with a subset of sites of protein B. Let us number the sites included in these subsets from 1 to L, so that the *i*th site of the subset of protein A corresponds to the *i*th site of the subset of protein B. Let K be the stiffness matrix of protein A (or B). Then sites can be sorted so that K can be written in block form:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{PP} & \mathbf{K}_{PQ} \\ \mathbf{K}_{QP} & \mathbf{K}_{QQ} \end{pmatrix} \tag{3}$$

where P corresponds to the subset of aligned sites and Q to that of nonaligned sites. Following [8,18,19] we can find an effective matrix:

$$\tilde{K}_{PP} = K_{PP} - K_{PQ}K_{QQ}^{-1}K_{QP} \tag{4}$$

The eigenvectors of this matrix are the effective normal modes that describe the motion of the aligned part of the given protein. Since the aligned subsets of proteins A and B correspond to each other, their effective normal modes are comparable in the sense that they are vectors in the same space.

Since the effective matrices $\mathbf{K}^{\mathrm{A}}$ and $\mathbf{K}^{\mathrm{B}}$ are symmetric, their eigenvectors form complete basis sets of the space spanned by the coordinates of the aligned sites. Therefore, any vector within such

space can be spanned in terms of either $\{\mathbf{q}_n^{\mathrm{A}}\}$ or $\{\mathbf{q}_n^{\mathrm{B}}\}$. In particular, any normal mode of protein A can be written as a linear combination of normal modes of protein B:

$$\mathbf{q}_n^{\mathrm{A}} = \sum_m S_{nm}^{\mathrm{AB}} \mathbf{q}_m^{\mathrm{B}} \tag{5}$$

where the overlap is given by the inner product

$$S_{nm}^{\mathrm{AB}} = \mathbf{q}_n^{\mathrm{A}} \cdot \mathbf{q}_n^{\mathrm{B}} \tag{6}$$

Let

$$P_{nm}^{\mathrm{AB}} = \left(S_{nm}^{\mathrm{AB}}\right)^2 \tag{7}$$

If the normal modes are normalized, it follows that

$$\sum_n P_{nm}^{\mathrm{AB}} = 1 \tag{8}$$

so that P_{nm}^{AB} can be interpreted as the relative contribution of the mth mode of protein B to the nth mode of protein A. The set $\{P_{nm}^{\mathrm{AB}}\}$ has the properties of a probability. Therefore, we can define an entropy

$$H_n^{\mathrm{AB}} = -\sum_m P_{nm}^{\mathrm{AB}} \ln P_{nm}^{\mathrm{AB}} \tag{9}$$

This entropy is a measure of the variability of mode $\mathbf{q}_n^{\mathrm{A}}$ when compared with protein B: $0 \le H_n^{\mathrm{AB}} \le \ln(3N - 6)$. An alternative measure of variability is

$$\kappa_n^{\mathrm{AB}} \equiv e^{H_{nm}^{\mathrm{AB}}} \tag{10}$$

It can be interpreted as the effective number of modes of B contained in the nth mode of A: $1 \le \kappa_n^{\mathrm{AB}} \le 3N - 6$.

To summarize, given two proteins A and B, they are aligned and superimposed, the effective K matrices and corresponding normal modes are calculated, and the variability of each mode is calculated using eq. 10. To quantify the variability of a mode of A with respect to a set of proteins, the variability is averaged over all members of the dataset.

Since the variability will depend on the difference between the two proteins compared, and I shall compare very divergent globin-like proteins with low sequence identities, with cases in which there are only a one or a few mutations, I will normalize the variabilities using Z-scores:

$$Z_n = \frac{\kappa_n - \langle \kappa \rangle}{\sqrt{\langle \kappa^2 \rangle - \langle \kappa \rangle^2}} \tag{11}$$

where averages are obtained over all normal modes.

Linearly forced elastic network model

Let (1) be the ENM potential for a given reference "wild-type" protein. Then, according to the LFENM, the potential for a protein that results from introducing a perturbation into the reference protein is given by

$$V' = \frac{1}{2}\left(\mathbf{r} - \mathbf{r}_0\right)^T \mathbf{K}\left(\mathbf{r} - \mathbf{r}_0\right) - \mathbf{f}^T\left(\mathbf{r} - \mathbf{r}_0\right) \tag{12}$$

where **f** is a "force" vector that models the effect of the perturbation. To model random mutations at a given site i we use

$$\mathbf{f}^i = \begin{pmatrix} \mathbf{f}_1^i \\ \mathbf{f}_2^i \\ \cdots \\ \mathbf{f}_N^i \end{pmatrix} \tag{13}$$

where $\mathbf{f}_k^i$ is a three-dimensional force vector onto site k due to a random mutation at i and we use a vector directed along contact i-k for all sites in contact with i (within a 7.5A cutoff) and a reaction force $\mathbf{f}_i^i = -\sum_{k \neq i} \mathbf{f}_k^i$. The magnitude of each force is picked randomly within the interval $[-f,f]$. Here we used $f = 2$. The equilibrium structure of the perturbed proteins, which minimizes its potential, is

$$\mathbf{r}_0' = \mathbf{r}_0 + \mathbf{K}^{-1}\mathbf{f} \tag{14}$$

For a detailed description and derivation of the LFENM see refs. [14,15].

Once the perturbed equilibrium structure is obtained, we calculate the stiffness matrix of the perturbed protein **K'**, which, within the ENM approximation depends only on structure. Then, we calculate the perturbed normal modes $\{\mathbf{q}_n'\}$ and compare with the reference ones $\{\mathbf{q}_n\}$ to obtain the degree of variation of each mode. To study the effect of random mutations, we introduce several random mutations at randomly picked sites and average over them.

RESULTS AND DISCUSSION

To study the origin of the higher degree of conservation of low-energy normal modes, I compared the mode-dependent conservation of three sets of proteins: using as reference the sperm-whale myoglobin with PDB code 1a6m, (1) the "globin-like" dataset consists of 22 members of the homologous "globin-like" superfamily, (2) the "myoglobin variants" dataset that includes 185 mutants and variants of 1a6m in different experimental conditions, and (3) the "LFENM" dataset that consists of simulated mutants obtained from 1a6m by applying the LFENM model. For more details, see Methods and ref. [15].

For a given dataset, each protein was aligned and superimposed with the reference 1a6m, I calculated the normal modes, compared them with those of 1a6m, and calculated the mode-dependent degree of variation κ_n, then I averaged these over all the proteins of each dataset. Figure 1 shows the average degree of variation as a function of normal mode for the dataset of globin-like evolved proteins and the dataset of simulated LFENM proteins. As expected from previous reports, for the evolutionarily related proteins, normal-mode variability tends to increase with normal-mode number: the lowest normal modes are the most conserved. The lines that fit the points for the experimental and simulated results are indistinguishable, showing that the agreement is excellent. More quantitatively, the correlation coefficient between the globin-like and LFENM results is 0.95. Finally, inspection of the dataset of myoglobin variants shows that even though it is noisier, there is a clear tendency of increasing variability with normal-mode number (data not shown). Results are noisy because despite the rather large number of variants (185) the spectrum of mutations and experimental conditions studied is very biased: most variation is due to either different ligands at the active site or engineered mutations at sites related to the active site. Despite this, the agreement between the variants dataset and the other two datasets is very good: the correlation coefficient between the variants dataset and the globin-like dataset is 0.75, and with the LFENM dataset is 0.70.

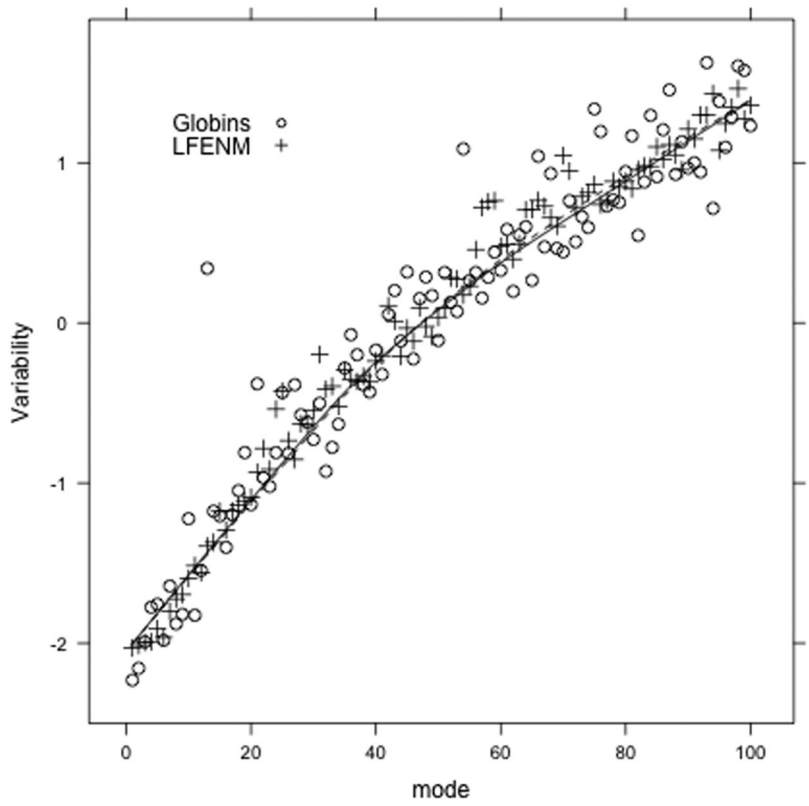

Fig. 1 Evolutionary variability of normal modes. Variability, quantified using the Z scores of eq. 11 vs. normal-mode index. Results for the experimental globin-like dataset of evolutionarily related proteins (open circles) and for random mutants simulated using the LFENM model (crosses) are shown. The smooth lines are Loess fit to the points. They are difficult to tell apart using visual inspection due to the high similarity between experimental and simulated results.

The excellent agreement between the mode-dependent evolutionary variability observed in globin-like proteins with that due to random mutations as obtained using the LFENM supports the idea that the observed variability of normal modes reflects their robustness with respect to mutations, rather than the effect of natural selection to conserve "functional" modes. This is not to say that the lowest normal modes are nonfunctional. There is evidence for the functional relevance of the lowest normal modes in many case studies (see ref. [6] and references therein). For the case of globin-like proteins, specifically, the lowest two normal modes have been linked to function [20]. However, the present results imply that one cannot use as evidence of the functional importance of a given normal mode its higher degree of conservation. Selection acts onto the material produced by random mutations, therefore, to demonstrate the effect of natural selection one should take first into account, as null hypothesis, the effect of random mutations. LFENM is such a null model and the present results show that the relative variabilities of different normal modes for the benchmark case of globin-like proteins do not depart from the expectations of the null model.

To summarize, I have shown that a very simple model of the effect of random mutations on protein normal modes accounts for the observed higher conservation of the lowest normal modes in evolutionary related proteins. This strongly supports the notion that the observed evolutionary conservation of the lowest normal modes is due to their higher robustness with respect to random mutations rather than natural selection against the variation of functional traits.

ACKNOWLEDGMENTS

This work has been supported by ANPCyT. The author is researcher of CONICET.

REFERENCES

1. E. Papaleo, L. Riccardi, C. Villa, P. Fantucci, L. De Gioia. *Biochim. Biophys. Acta* **1764**, 1397 (2006).
2. S. Maguid, S. Fernández-Alberti, G. Parisi, J. Echave. *J. Mol. Evol.* **63**, 448 (2006).
3. A. Pandini, G. Mauri, A. Bordogna, L. Bonati. *Protein Eng. Des. Sel*. **20**, 285 (2007).
4. M. Rueda, P. Chacón, M. Orozco. *Structure* **15**, 565 (2007).
5. A. Ahmed, S. Villinger, H. Gohlke. *Proteins: Struct., Funct., Bioinform.* **78**, n/a (2010).
6. I. Bahar, T. R. Lezon, L.-W. Yang, E. Eyal. *Ann. Rev. Biophys*. **39**, 23 (2010).
7. A. Pang, Y. Arinaminpathy, M. S. P. Sansom, P. C. Biggin. *Proteins: Struct., Funct., Bioinform.* **61**, 809 (2005).
8. V. Carnevale, S. Raugei, C. Micheletti, P. Carloni. *J. Am. Chem. Soc.* **128**, 9766 (2006).
9. E. Marcos, R. Crehuet, I. Bahar. *PLoS Comput. Biol.* **6**, e1000738 (2010).
10. S. Maguid, S. Fernandez-Alberti, J. Echave. *Gene* **422**, 7 (2008).
11. F. Raimondi, M. Orozco, F. Fanelli. *Structure* **18**, 402 (2010).
12. M. Münz, R. Lyngsø, J. Hein, P. C. Biggin. *BMC Bioinformatics* **11**, 188 (2010).
13. S. M. Hollup, E. Fuglebakk, W. R. Taylor, N. Reuter. *Protein Sci.* **20**, 197 (2011).
14. J. Echave. *Chem. Phys. Lett.* **457**, 413 (2008).
15. J. Echave, F. Fernández. *Proteins* 173 (2010).
16. L. Conte, B. Ailey, T. J. P. Hubbard, S. E. Brenner, A. G. Murzin, C. Chothia. *Nucleic Acids Res*. **28**, 257 (2000).
17. C. Micheletti, P. Carloni, A. Maritan. *Proteins* **55**, 635 (2004).
18. K. Hinsen, A. J. Petrescu, S. Dellerue, M. C. Bellisent-Funel, G. Kneller. *Chem. Phys.* **261**, 25 (2000).
19. A. Zen, V. Carnevale, A. M. Lesk, C. Micheletti. *Protein Sci.* **17**, 918 (2008).
20. S. Maguid, S. Fernandez-Alberti, L. Ferrelli, J. Echave. *Biophys. J.* **89**, 3 (2005).

