ELSEVIER

Biophysical Chemistry 115 (2005) 125–128

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Core deformations in protein families: a physical perspective

Alejandra Leo-Macias[a], Pedro Lopez-Romero[a], Dmitry Lupyan[b], Daniel Zerbino[a], Angel R. Ortiz[a,*]

[a]*Bioinformatics Unit, Centro de Biologia Molecular "Severo Ochoa", CSIC-UAM, Universidad Autonoma de Madrid, Cantoblanco 28049, Madrid, Spain*
[b]*Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY, USA*

Received 30 June 2004; received in revised form 10 November 2004; accepted 10 December 2004
Available online 18 January 2005

Abstract

An analysis is presented on how structural cores change shape within protein families, and whether or not there is a relationship between these structural changes and the vibrational modes that proteins experiment due to topological constraints. A set of 13 representative and well-populated protein families are studied. The evolutionary directions of deformation are obtained by applying a new multiple structural alignment technique to superimpose the structures and extract a conserved core, together with Principal Components Analysis (PCA) to extract the main deformation modes. A low-resolution Normal Mode Analysis (NMA) technique is used in parallel to study the properties of the mechanical core plasticity of the same proteins. We find that the evolutionary deformations span a low dimensional space. A statistically significant correspondence exists between these principal deformations and the vibrational modes accessible to a particular topology. We conclude that, to a significant extent, the structures of evolving proteins seem to respond to sequence changes by collective deformations along combinations of low-frequency modes. The findings have implications in structure prediction by homology modeling.

Keywords: Homology modeling; Normal mode analysis; Protein evolution

1. Introduction

One stagnant area in structure prediction by homology modelling is model refinement [1]. Results from every CASP edition consistently highlight that no model has resulted to be closer to the target structure than the template to any significant extent, i.e., the maximum improvement is never larger than ~0.4 Å. By contrast, the average RMSD in the structural core among remote homologues (those below 40% sequence identity) is about ~2.0 Å (*vide infra*). This indicates that there is a great need for a more accurate modeling of distortions and rigid body shifts imposed by sequence changes among protein homologues. Here, we study the main directions of protein structural change among multiple homologous proteins using multiple structural alignments and Principal Components Analysis (PCA) [2], and study their relationship with the low-frequency modes of vibration imposed by the protein topology [3]. We evaluate quantitatively the relationship between both spaces as well as the statistical significance of the overlap. We show that in fact a statistically significant relationship can be found. The implications in homology modeling are briefly discussed.

2. Methods

2.1. Data set

The data set (Table 1) was selected from the ASTRAL40 structural database [4]. A sample of thirteen large, well-studied superfamilies, classified according to the SCOP classification [5], has been selected. The maximum percentage of identity between structures is 40%, while the average sequence identity in the core among the pairs is of ~25%.

* Corresponding author. Tel.: +34 91 497 2376; fax: +34 91 497 4799. *E-mail address:* aro@cbm.uam.es (A.R. Ortiz).

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.12.016

Table 1
Protein families analyzed and summary of PCA results

Family	#Struct	#Core	%Core	<rms>±σ	# PC's
GLOBINS	23	75	69	1.89±0.63	5
KINASES	22	166	64	2.03±0.47	6
IMMUNOGLOBULINES	23	50	58	1.92±0.54	6
GLUTATION *S*-TRANSF	22	67	59	1.90±0.51	6
INTERLEUKIN 8-LIKE CHEM.	11	51	83	1.63±0.71	4
RNA-BINDING DOMAIN	21	51	68	2.70±0.59	5
FIBRONECTIN	46	38	45	2.34±0.89	9
CYTOCHROME C	16	36	46	1.64±0.43	3
THIOREDOXIN-LIKE	35	39	53	2.08±0.84	3
SH3	24	34	60	1.87±0.55	5
CUPREDOXINS	22	48	49	2.00±0.56	4
SNAKE TOXIN-LIKE	11	36	60	1.49±0.41	4
ALDOLASES	19	84	40	2.07±0.45	5

#Struct, number of structures for each family; #Core, the core size to the subjected to PCA, as obtained from the multiple structural alignment; #PC's, the number of principal components required to explain 70% of the variance in the core; <rms>±σ, average RMS and standard deviation of the proteins in that set; %Core, percentage of core with respect to the smallest protein in the set.

Set of ASTRAL domains used: GLOBINS (d3sdha, d1b0b, d1h97a, d1jl7a, d1a6m, d1mba, d1eco, d2gdm, d1irda, d1gcva, d1hjb, d1cg5b, d1gcvb, d1it2a, d1ash, d1itha, d1hlb, d1cqxa1, d1ew6a d1dlwa, d1dlya, d1idra d1kr7a); KINASES (d1jvpp, d1apme, d1a06, d1kwpa, d1o6la, d1a8a, d1phk, d1gnga, d1kia, d1koba, d1pme, d1csn, d1lpua, d1b6cb, d1f3mc, d1howa, d1jksa, d1o6ya, d1qpca, d1fgka, d1ir3a, d1m14a); IMMUNOGLOBULINS (d1c5ch2, d1c5cl2, d1dn0b2, d1dr9a2, d1fnga1, d1fngb1, d1fp5a1, d1fp5a2, d1gzqa1, d1hdma1, d1hdmb1, d1hxma2, d1hxmb2, d1hyrc1, d1iam_1, d1k5na1, d1k5nb, d1kgce2, d1l6xa1, d1o0va1, d1vcaa1, d1zxq_1, d2fbjh2); GLUTATION-*S*-TRASNFERASES (d1glqa1, d2gsta1, d1k3ya1, d1duga1, d1oe8a1, d1ljra1, d1iyha1, d1m0ua1, d2gsq_1, d1eema1, d1e6ba1, d1gwca1, d1oyja1, d1jlva1, d1gnwa1, d1aw9_1, d1a0fa1, d1f2ea1, d1g7oa1, d1k0da1, d1nhya1, d1k0ma1), INTERLEUKIN 8-LIKE CHEMOKINES (d1o80a, d1m8aa, d1cm9a, d1b3aa, d1doka, d1el0a, d1eiha, d1g2ta, d1j9oa, d1f2la, d1tvxa), RNA-BINDING DOMAIN (d1l3ka1, d1l3ka2, d1nu4a, d2u1a, d2u2fa, d1o0pa, d1u2fa, d1fxla1, d1fxla2, d2msta, d1cvja1, d1cvja2, d1qm9a1, d1qm9a2, d1fj7a, d1fjeb2, d1h6kx, d1oo0b, d1owxa,d1koha2,d1jmta), FIBRONECTIN (d2hft_1, d2hft_2, d1fna, d1fnf_1, d1fnf_2, d1fnf_3, d1fnha1, d1fnha2, d2fnba, d1j8ka, d1qr4a1, d1qr4a2, d1cfb_1, d1cfb_2, d1lwra, d1k85a, d1qg3a1, d1qg3a2, d1axib1, d1axib2, d1eerb1, d1eerb2, d1f6fb1, d1f6fb2, d1iarb1, d1iarb2, d1gh7a1, d1gh7a2, d1gh7a3, d1egja, d1cd9b1, d1cd9b2, d1fyhb1, d1fyhb2, d1bqua1, d1bqua2, d1i1ra1, d1lqsr1, d1lqsr2, d1bpv, d1f42a2, d1f42a3, d1n26a2, d1n26a3, d1n6va1, d1n6va2), CYTOCHROME C (d1h1oa1, d1fcdc1, d1fcdc2, detpa2, d1h1oa2, d1h32b, d1c53, d1cnoa, d1c52, d451c, d1ql3a, d1ycc, d1i8oa, d1cot, d1kb0a1, d1kv9a1); THIOREDOXIN-LIKE (d1a8y_1, d1a8y_3, d1a8y_2, d1mek, d1bjx, d1a8l_1, d1hyua3, d1a8l_2, d1hyua4, d1qgva, d1g7ea, d1erv, d1fo5a, d1iloa, d1aba, d1qfna, d1kte, d1nm3a1, d1h75a, d1k0ma2, d1a0fa2, d1g7oa2, d1ljra2, d1glqa2, d1eema2, d1oyja2, d1jlva2, d1e6ba2, d1gnwa2, d1k0da2, d2gsq_2, d2gsta2, d1k3ya2, d1iyha2, d1nhya2), SH3 (d1i07a, d1ng2a1, d1kjwa1, d1pht, d1ckaa, d1awj, d2hsp, d1sema, d1fmk_1, d1gl5a, d1bbza, d1pwt, d1gbra, d1k4us, d1ng2a2, d1oeba, d1bb9, d1i1ja, d1cska, d1neb, d1jqqa, d1ycsb2, d1gcqc, d1jo8a), CUPREDOXINS (d1kcw_2, d1oe1a2, d1kbva2, d1hfua2, d1aoza2, d1gw0a2, d1kv7a2, d1gska2, d1kv7a3, d1gska3, d1aoza3, d1hfua3, d1gw0a3, d1kcw_5, d1kcw_1, d1kcw_3, d1bqk, d1aac, d1kdj, d1plc, d1bawa, d1ibya), SNAKE TOXIN-LIKE (d2ctx, d1f94a, d3ebx, d1ff4a, d1jgka, d1fas, d1tgxa, d1es7b, d1btea, d1m9za, d1erh), ALDOLASES (d1epxa, d1f74a, d1dhpa, d1hl2a, d1euaa, d1n7ka, d1jcla, d1ub3a, d1qfea, d1i2oa, d1l6wa, d1dosa, d1gvfa, d1l6sa, d1gzga, d1ohla, d1jcxa, d1n8fa, d1nvma2).

2.2. Multiple structural alignment

The structural set corresponding to each one of the 13 families was subjected to multiple structural alignment using MAMMOTH-mult (Lupyan et al., in preparation), a multiple alignment version of the structural alignment program MAMMOTH [6]. From the alignment, the evolutionary core of the protein family is selected. This is defined as the set of gapless positions for which the Cα atoms of all members are within 4 Å from the family average. This way, a matrix $\mathbf{X}_{\mathrm{nxp}}$ is obtained containing the Cartesian coordinates of the Cα core positions in the family, with n being the number of structures and p is 3 times the number of core positions (each position is defined by its corresponding x,y,z Cartesian coordinates).

2.3. Principal Components Analysis (PCA)

PCA [2] was used to extract the set of main modes of motion in the alignment that best describe the deformations experienced by the core. Starting from $\mathbf{X}_{\mathrm{nxp}}$, the covariance matrix $\mathbf{C}_{\mathrm{pxp}}$ is computed, with elements $c_{ij}=<(x_i-<x_i>)(x_j-<x_j>)>$, where averages <> are over the n structures. Then, $\mathbf{C}$ is subjected to spectral decomposition as $\mathbf{C}=\mathbf{V}\mathbf{\Delta}\mathbf{V}^{\mathbf{T}}$, where $\mathbf{V}$ is an orthogonal matrix containing the set of eigenvectors and $\mathbf{\Delta}$ is a diagonal matrix containing the set of eigenvalues. The eigenvector matrix $\mathbf{V}$ is then used in the comparison with NMA.

2.4. Vibrational modes: the Anisotropic Network Model (ANM)

For the simulation of the vibrational modes we used ANM [3]. ANM is a special type of NMA method. It is a coarse-grained model, which assumes that the protein in the folded state is equivalent to a three-dimensional elastic network. The junctions of the network (the Cα atoms) undergo Gaussian-distributed fluctuations under the potential of the pendant chains. The potential energy of the protein (V) as a function of the displacement ($\mathbf{R}$) from the native conformation (in Cartesian coordinates) is thus $V=\mathbf{R}\mathbf{H}\mathbf{R}^{\mathbf{T}}$, where $\mathbf{H}$ is the Hessian matrix containing the second derivatives of the energy function, assumed to be harmonic. Diagonalization of $\mathbf{H}$ as $\mathbf{H}=\mathbf{U}\mathbf{\Lambda}\boldsymbol{U}^{\mathbf{T}}$ yields $3N$-6 intrinsic normal modes (N being the number of residues), contained in eigenvector matrix **U**, to be compared with the PCA directions.

2.5. Relating both spaces: the RMSIP calculation

We compared the vibrational modes obtained by ANM with the structural fluctuations detected by PCA. The overlap between both spaces is calculated from the root-mean-square inner product (RMSIP) [7] of the PCA eigenvectors with the vibrational ones:

$$\mathrm{RMSIP} = \left(\frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{k} \left(\boldsymbol{\eta}_i \cdot \boldsymbol{\nu}_j \right)^2 \right)^{1/2} \qquad (1)$$

Here $\boldsymbol{\eta}_i$ and $\boldsymbol{\nu}_j$ are, respectively, the set of eigenvectors of the evolutionary and ANM spaces, respectively; D is the dimensionality of the evolutionary space, whereas k is the dimensionality of the ANM space. The evolutionary space is restricted to the number of components required to explain 70% of the variance, 5 components on average (see below). The normal mode space is restricted to its 50 lowest frequency modes. For each family, the structure closest to the family average, as determined by MAMMOTH-mult, is used for the computation of the normal modes. The statistical significance of the observed RMSIP was tested using a randomization distribution of RMSIP values under the null hypothesis of no relationship between both spaces. The randomization distribution of RMSIP values was based on the generation of ten thousand orthogonal **Q** matrices following the Stewart algorithm [8]. The empirical distribution of RMSIP under the null hypothesis allows computing the Z-score of the observed RMSIP, as follows:

$$Z - \mathrm{score} = \frac{\mathrm{RMSIP(obs)} - <\mathrm{RMSIP(ran)}>}{\sigma(\mathrm{ran})} \qquad (2)$$

3. Results and discussion

The number of proteins used in the alignments ranges from 11 to 46. The structural core detected from the alignments and used in the PCA studies comprises on average 58% of the structure (Table 1). Thus both the number of structures used and the core size detected seem to be large enough to ensure that the deformations detected approximate the true deformations experienced by the protein family. The average RMSD among structures in the studied protein families is ~2.0 Å. This is substantially larger than the ~0.4 Å of improvement that in favorable cases can be expected from current homology modeling tools. Given that this corresponds to the situation most frequently found nowadays in homology modeling, the result points to the necessity of improving the tools to consider collective deformations.

Structural deformations span a space of low dimensionality; 5 components explain 70% of the variance in most cases (Table 1). The behavior of all families in PCA is rather similar, independently of the structural class, size, or number of structures in the family. We used the number of evolutionary components that explains 70% of the total variance and compared the directions with ANM, where we considered up to 50 modes. Results are in Fig. 1. A significant overlap quickly accumulates within the first ~15 modes, reaching an average Z-score of ~12. Thus, it seems that there is a statistically significant overlap between the deformations observed in the core of homologous proteins and the lowest frequency modes imposed by the protein topology.

The use of PCA directions appears as a promising technique to model structural plasticity in homology modeling problems, due to its low dimensionality [9]. This low dimensional subspace overlaps significantly with the subspace spanned by the low frequency modes imposed by the topology, suggesting that a subset of ~15 low frequency modes can be used in model refinement in those cases where the number of structures in the family does not allow for the use of PCA. Monte Carlo sampling along these low-frequency modes could be used to model backbone flexibility in the core, coupled to side chain repacking algorithms [10] on fixed backbones to consider their side chain degrees of freedom. On a more fundamental level, our results suggest that the evolutionary pathways of structural adaptation make use, to some extent, of combinations of low-frequency modes imposed by the topology, i.e., the protein topology could be an important factor determining the evolutionary history of proteins at the structural level.

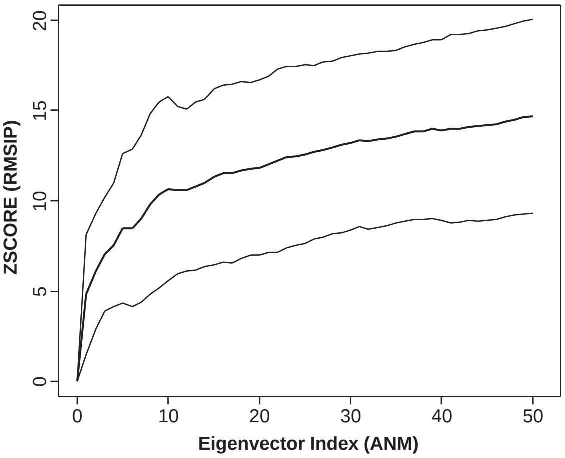

Fig. 1. Z-scores of the RMSIP values for the overlap between the PCA and ANM spaces as a function of number of ANM modes employed. Up to 50 modes have been considered. The thick line indicates the average value over the 13 families. The thin lines correspond to one standard deviation from the mean.

Acknowledgments

Funded by grant BIO2001-3745 from Spanish MCYT. ALM is a FPI predoctoral fellow. DZ thanks a visiting fellowship from the Ecole Polytechnique (France). Research at CBM-SO is facilitated by an institutional grant from *Fundación Ramón Areces*.

References

[1] A. Tramontano, V. Morea, Assessment of homology-based predictions in Casp5, Proteins 53 (Suppl. 6) (2003) 352–368.

[2] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Upper Saddle City, NJ, 1998.

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

[4] S.E. Brenner, P. Koehl, M. Levitt, The astral compendium for protein structure and sequence analysis, Nucleic Acids Res. 28 (2000) 254–256.

[5] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, Scope: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

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

[7] A. Amadei, M.A. Ceruso, A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations, Proteins 36 (1999) 419–424.

[8] G.W. Stewart, The efficient generation of random orthogonal matrices with an application to condition estimation, SIAM J. Numer. Anal. 17 (1980) 403–409.

[9] B. Qian, A.R. Ortiz, D. Baker, Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 15346–15351.

[10] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, J. Tsai, Protein folding: the endgame, Annu. Rev. Biochem. 66 (1997) 549–579.