BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 19 2012, pages 2431–2440
doi:10.1093/bioinformatics/bts445

Structural bioinformatics

Advance Access publication July 12, 2012

Measuring and comparing structural fluctuation patterns in large protein datasets

Edvin Fuglebakk[1,2], Julián Echave[3] and Nathalie Reuter[2,4,*]

[1]Department of Informatics, University of Bergen, Pb. 7803, N-5020 Bergen, Norway, [2]Computational Biology Unit, Uni Computing, Uni Research, Pb. 7810, N-5020 Bergen, Norway, [3]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, 1650 San Martín, Argentina and [4]Department of Molecular Biology, University of Bergen, Pb. 7803, N-5020 Bergen, Norway

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The function of a protein depends not only on its structure but also on its dynamics. This is at the basis of a large body of experimental and theoretical work on protein dynamics. Further insight into the dynamics–function relationship can be gained by studying the evolutionary divergence of protein motions. To investigate this, we need appropriate comparative dynamics methods. The most used dynamical similarity score is the correlation between the root mean square fluctuations (RMSF) of aligned residues. Despite its usefulness, RMSF is in general less evolutionarily conserved than the native structure. A fundamental issue is whether RMSF is not as conserved as structure because dynamics is less conserved or because RMSF is not the best property to use to study its conservation.

Results: We performed a systematic assessment of several scores that quantify the (dis)similarity between protein fluctuation patterns. We show that the best scores perform as well as or better than structural dissimilarity, as assessed by their consistency with the SCOP classification. We conclude that to uncover the full extent of the evolutionary conservation of protein fluctuation patterns, it is important to measure the directions of fluctuations and their correlations between sites.

Contact: Nathalie.Reuter@mbi.uib.no

Supplementary Information: Supplementary data are available at *Bioinformatics* Online.

Received on April 11, 2012; revised on June 6, 2012; accepted on July 6, 2012

1 INTRODUCTION

Proteins like all molecules undergo structural fluctuations in physiological conditions. Small and large structural changes directly related to protein functions occur at a wide range of time scales. Relatively fast local fluctuations are for example involved in induced-fit ligand binding. Slower and more global fluctuations are important in complex coherent transitions of allosteric proteins and molecular machines. Such dynamics–function relationships are at the basis of a large body of theoretical and experimental work on protein dynamics (Henzler-Wildman and Kern, 2007).

Concerted motions in protein structures are conditioned by the lowest energy modes of fluctuations around their equilibrium conformation. These modes can be obtained from simulation techniques, such as molecular dynamics (MD) simulations, where the modes are extracted by principal component analysis of the covariance matrix. Alternatively, and to avoid high computational costs, the normal modes can be calculated from an equilibrium structure (typically an experimental structure determined by X-ray scattering or nuclear magnetic resonance spectroscopy (NMR)) using coarse-grained representations and elastic network models (ENMs), implying a harmonic approximation of the potential around the equilibrium structure. Several independent studies have shown that ENM modes and principal components obtained from MD simulations are in very good agreement, see e.g. Rueda *et al.* (2007), Skjaerven *et al.* (2011), and Yang *et al.* (2008). For these reasons, ENM normal modes are commonly used to characterize protein dynamics and are convenient for studying large protein datasets. In addition, allosterism has been successfully studied using ENMs and it has been shown that functional conformational changes in general usually involve only a few normal modes, see for instance Bahar *et al.* (2010). This has for example been exploited to investigate allosteric mechanisms and predict effector binding sites (Mitternacht and Berezovsky, 2011a,b; Zen *et al.*, 2009)

In contrast with sequence or structure, using dynamics in comparative computational studies has only recently started to emerge. Yet this is a field that has already contributed significantly to increasing our understanding of protein dynamics, and in particular its conservation through evolution and relation to function. In general, it has been shown that backbone fluctuations are evolutionarily conserved (Maguid *et al.*, 2006, 2008). More detailed studies of particular folds or protein families have found that the lowest energy normal modes, or equivalently, the highest principal components of the fluctuation covariance matrix, are conserved for homologous proteins or even for non-homologous proteins with similar folds (Carnevale *et al.*, 2006; Hollup *et al.*, 2011; Keskin *et al.*, 2000; Maguid *et al.*, 2005, 2008; Pang *et al.*, 2005). One obvious field of application is the adaptation of proteins to extreme temperatures and the associated changes in flexibility—see Papaleo *et al.* (2006) and references therein. Among other applications reported are dynamics-based alignments and detection of distant homologs

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press on behalf of The Society for Financial Studies. All rights reserved.
For Permissions, please e-mail: journals.permissions.com.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

(Keskin *et al*., 2000; Münz *et al*., 2010; Potestio *et al*., 2010; Zen *et al*., 2008) and analysis of evolutionary dynamics (Echave and Fernández, 2010; Leo-Macias *et al*., 2005; Raimondi *et al*., 2010).

Comparing protein dynamics implies the choice of one or more properties that characterize the dynamics of the proteins and measures of the (dis)similarity of such properties. A number of properties and (dis)similarity measures have so far been reported to capture the conservation of protein dynamics. The simplest and most used property is the atomic root mean square fluctuations (RMSF) (Keskin *et al*., 2000; Maguid *et al*., 2006; Papaleo *et al*., 2006), which is comparable to X-ray B-factors. Commonly, interpretations are restricted to only the fluctuations of the proteins α-carbons (C_α). RMSFs can be readily calculated from any ensemble of protein conformations, such as those obtained from MD simulations or Monte Carlo (MC) simulations. MD and MC simulations are computationally expensive and hard to parameterize. It is therefore convenient that multiple studies have shown that protein flexibility is well captured by the lowest energy normal modes, even when coarse-grained approaches are used (Rueda *et al*., 2007; Skjaerven *et al*., 2011; Yang *et al*., 2008).

Several studies on comparative protein dynamics have taken advantage of this and are based on comparing protein flexibility as it is described by the proteins' normal modes (Keskin *et al*., 2000; Maguid *et al*., 2005, 2008). Subspaces spanned by a few important normal modes or, equivalently, principal components of the fluctuation covariance matrix, have been compared using the root mean square inner product [RMSIP see for instance Carnevale *et al*. (2006)] or the related RWSIP (Carnevale *et al*., 2007). A dynamic fingerprint matrix has been used to compare the backbone dynamics between PDZ domains (Münz *et al*., 2010). A dynamic fingerprint matrix is analogous to a distance matrix for a single protein conformation, measuring variability of inter-residue distances in an ensemble of conformations.

Actually all the methods described for comparing protein dynamics can be seen as ways of comparing the conformational ensembles that characterise protein fluctuations. Each method focuses on some property, such as the RMSF profile or the covariance matrix, and uses some measure of similarity or dissimilarity, such as the Spearman correlation or the RMSIP, for comparing this property between different proteins.

Because of the variation in the properties and (dis)similarity measures that have been used on different sets of proteins, it is difficult to compare the results of these studies. Yet there are fundamental differences in the dynamical properties and (dis)similarity measures reported, and there is a need for knowing which one(s) should be used. For example, systematic studies of the evolutionary conservation of the fluctuation patterns are based largely on the similarity between RMSF profiles (Maguid *et al*., 2006, 2008). These studies have demonstrated the evolutionary conservation of backbone fluctuations. However, RMSF profiles do not utilize the full extent of information available from computational analysis of protein flexibility, either obtained by simulations or normal mode analysis. In addition, RMSF profiles have not been shown to be more conserved than protein structure. Rather, previous studies thoroughly discuss the possibility that they are less conserved (Maguid *et al*., 2006, 2008), which we indeed find to be the case. This may be a practical issue of RMSF not being the property that most adequately describes the fluctuation patterns or a fundamental one of protein dynamics being less evolutionarily conserved than protein structure. The purpose of this work is also to address this issue.

To tackle the question of whether the rather low conservation of RMSF profiles is a sign of low conservation of protein dynamics, we performed an assessment of different ways of comparing protein fluctuations. We consider protein domains with well-defined native structures, for which the conformational ensembles that result from fluctuations are adequately described by multidimensional Gaussian distributions (Hinsen *et al*., 2000). Such distributions around an equilibrium are completely characterized by the covariance matrices. The two most frequently used measures of dynamical similarity are the correlation between RMSF profiles and the RMSIP between the 10 principal components of the full covariance matrix. These can be thought of as two limiting cases: RMSF considers only the overall magnitude of the fluctuation of each C_α-atom, whereas the full covariance matrix includes information on the anisotropy (the directionality) of such fluctuations and their correlations between sites. To disentangle the different contributions to dynamical similarity, we also study the correlation matrix, the isotropic covariance matrix and the isotropic correlation matrix. These matrices are compared using different measures of (dis)similarity. RMSIP and RWSIP have previously been used in the literature, and although they are well motivated by physical arguments they are not rigorously derived for the comparison of multidimensional Gaussian distributions. Therefore, we also introduce a new measure, ND_B, based on the Bhattacharyya distance between probability density functions (Bhattacharyya, 1943). The performances of the properties and the (dis)similarity measures are evaluated by investigating their consistency with the SCOP classification at the superfamily level.

2 METHODS

2.1 Datasets

We choose four datasets from the ASTRAL compendium (Chandonia *et al*., 2004) (version 1.75), one for each of the four main SCOP classes (Murzin *et al*., 1995). All domains were chosen from the subset of the ASTRAL compendium that has at most 95% sequence identity between domains. Each dataset is composed of protein domains that belong to two different superfamilies of the same fold. In addition, for each superfamily we included domains from two different families. We made sure to make the selection so that all families are represented by at least 6 domains, for a total set of 189 domains. The selection of protein domains was also partly guided by the need for a structural alignment procedure to terminate with a successful alignment. The selection contains structures determined by either X-ray crystallography or NMR spectrometry. For the 40 protein domains determined by NMR spectroscopy, we verified that all structures were representative of a well-defined equilibrium in the core positions. Throughout, we will refer to specific folds and superfamilies by the relevant part of the SCOP concise classification strings, introduced in Lo Conte *et al*. (2002). Full listings of the datasets are available as Supplementary Tables S1–S4.

2.2 Definition of the structural core

We generated multiple structural alignments of all the domains of each dataset using STAMP (Russell and Barton, 1992). We define the structural core as the aligned sites for which there are no gaps and perform all comparisons on this structural core.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

Comparing only the sites conserved over a whole dataset, we make sure that the (dis)similarity scores are directly compared only when they are obtained for corresponding sites. This more conservative way of aligning the protein domains allows us to properly remove information about evolutionary relations that are not common to all the compared domains in a dataset, avoiding the problem of normalizing scores to alignment length, which is necessary when pairwise alignments are used. Excluding loops and loosely associated secondary structure elements also ensures that we are considering the part of the proteins for which the ENM approach is best motivated.

2.3 Properties compared

All (dis)similarity scores between two proteins involve some property compared and a measure of (dis)similarity of such a property. To compare structures, we used the equilibrium coordinates of the C_α-atoms in the structural core. To compare the structural fluctuations, we used different statistics that characterize the multidimensional distribution of conformations that the proteins may adapt through fluctuations around the native structure.

We will characterize a protein conformation using the column vector:

$$\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2 \ldots)^T,$$

where T denotes the transpose and $\mathbf{r}_i = (x_i, y_i, z_i)^T$ is the position vector of the C_α-atom of site i. Let $\mu = <\mathbf{r}>$ be the mean conformation, which we shall refer to interchangeably as the 'native structure' or 'equilibrium conformation'. Then a 'fluctuation' is characterized by $\delta\mathbf{r} = \mathbf{r} - \boldsymbol{\mu}$ and is composed of C_α fluctuations $\delta\mathbf{r}_i = \mathbf{r}_i - \boldsymbol{\mu}_i$.

A full characterization of fluctuations is given by a probability density function $\rho(\delta\mathbf{r})$. Given this, different statistics can be calculated to describe the fluctuation patterns. The ones used here are

$$\mathbf{C}_{i\mu,j\nu} = <\delta\mathbf{r}_{i\mu}\delta\mathbf{r}_{j\nu}>, \tag{1}$$

where $\mathbf{C}$ is the covariance matrix, i and j are two C_α-atoms and μ and ν identify the cartesian components.

$$\mathbf{P}_{i\mu,j\nu} = \mathbf{C}_{i\mu,j\nu}/\left(\mathbf{C}_{i\mu,i\mu}\mathbf{C}_{j\nu,j\nu}\right)^{1/2}, \tag{2}$$

where $\mathbf{P}$ is the correlation matrix.

$$\mathbf{C}_{ij}^{iso} = <\delta\mathbf{r}_i \cdot \delta\mathbf{r}_j> = \mathbf{C}_{ix,jx} + \mathbf{C}_{iy,jy} + \mathbf{C}_{iz,jz}, \tag{3}$$

where $\mathbf{C}^{iso}$ is the isotropic covariance matrix.

$$\mathbf{P}_{ij}^{iso} = \mathbf{C}_{ij}^{iso}/\left(\mathbf{C}_{ii}^{iso}\mathbf{C}_{jj}^{iso}\right)^{1/2}, \tag{4}$$

where $\mathbf{P}^{iso}$ is the isotropic correlation matrix.

$$\mathbf{RMSF}_i = <||\delta\mathbf{r}_i||^2> = \mathbf{C}_{ii}^{iso}, \tag{5}$$

where RMSF is the root mean square fluctuation profile.

Note that $\mathbf{C}^{iso}$, $\mathbf{P}^{iso}$ and RMSF are invariant with respect to rotations of the whole protein: they are isotropic. RMSF contains information on the amplitudes of fluctuation of each C_α-atom, but not on the correlation of fluctuations between sites. $\mathbf{P}^{iso}$ and $\mathbf{P}$ quantify such correlations, the latter taking into account their anisotropic nature. $\mathbf{C}^{iso}$ and $\mathbf{C}$ combine amplitudes and correlations. Comparison of these different properties allows the disentangling of different aspects of the fluctuation patterns.

2.3.1 Calculation of the covariance matrix From the previous definitions it follows that all properties that will be used here to characterize the fluctuation pattern can be obtained from the covariance matrix. To calculate the covariance matrix we used the coarse-grained ENM developed by Hinsen (1998). Thorough assessments show that for calculation of covariance matrices ENMs are as good as or better than MD simulations (Ahmed *et al.*, 2010). Each residue is represented as a node located at the equilibrium position of its C_α-atom and each pair of nodes are connected by a Hookean spring. In this work we set the equilibrium positions to the experimental atomic coordinates deposited in the Protein Data Bank (Berman *et al.*, 2000). The potential energy of the spring connecting nodes i and j due to departures from the equilibrium conformation is given by

$$U_{ij}(\mathbf{r}) = k\left(\left|\left|\mathbf{r}_{ij}^0\right|\right|\right)\left(\left|\left|\mathbf{r}_{ij}\right|\right| - \left|\left|\mathbf{r}_{ij}^0\right|\right|\right)^2,$$

where $\mathbf{r}$ is a protein conformation, $\mathbf{r}^0$ is the equilibrium conformation and $\mathbf{r}_{ij}$ and $\mathbf{r}_{ij}^0$ are distance vectors between nodes i and j in the respective conformations. The force constant is given by

$$k(r) = \begin{cases} ar - b, & \text{for } r < d \\ cr^{-6}, & \text{for } r \geq d \end{cases}.$$

The following parameters for this equation have been determined by fitting the ENM to MD simulation results (Hinsen *et al.*, 2000): $a = 8.6 \times 10^5$; kJ mol^{-1} nm^{-3}, b $= 2.39 \times 10^5$; kJ mol^{-1} nm^{-2}; c $= 128$ kJ nm^4 mol^{-1} $d = 0.4$ nm.

Adding the potential energy terms for all site pairs and expanding the result in Taylor series around $\mathbf{r}^0$ up to second order we obtain

$$U(\mathbf{r}) = \left(\mathbf{r} - \mathbf{r}^0\right)^T\mathbf{K}\left(\mathbf{r} - \mathbf{r}^0\right),$$

where the force-constant matrix, $\mathbf{K}$, is the matrix of second derivatives of the ENM potential.

To obtain an effective force-constant matrix for only the atoms in the aligned core, we follow Carnevale *et al.* (2006) and Hinsen *et al.* (2000)

$$\tilde{\mathbf{K}} = \mathbf{K}_{AA} - \mathbf{K}_{AQ}\mathbf{K}_{QQ}^{-1}\mathbf{K}_{QA},$$

where $\mathbf{K}_{AA}$ is the sub-matrix of $\mathbf{K}$ corresponding to the aligned sites $\mathbf{K}_{QQ}$ is the sub-matrix corresponding to the non-aligned sites and $\mathbf{K}_{QA}$ and $\mathbf{K}_{AQ}$ are the sub-matrices that couple coaligned and non-aligned sites.

Given $\tilde{\mathbf{K}}$, we solve the eigenvalue equation

$$\tilde{\mathbf{K}}\mathbf{q}_n = \lambda_n\mathbf{q}_n$$

to obtain the normal modes $\mathbf{q}_n$, which are vectors of size $3p$ describing independent deformations, p being the number of aligned sites. The eigenvalues λ_n represent the energetic cost of these deformations. The six lowest eigenvalues are 0. They correspond to rotations and translations of the whole molecule. These trivial modes will not be considered in our analyses.

The covariance matrix is given by

$$\mathbf{C} = \tilde{\mathbf{K}}^{-1}.$$

Since some eigenvalues of $\tilde{\mathbf{K}}$ are zero, we use the pseudo-inverse

$$\mathbf{C} = \sum_{n=1}^{3p-6}\frac{1}{\lambda_n}\mathbf{q}_n\mathbf{q}_n^T,$$

where the sum runs over the $3p - 6$ non-trivial modes. The normal modes are also the principal components of the distribution describing protein fluctuations. They are the eigenvectors of $\mathbf{C}$ identical to those of $\tilde{\mathbf{K}}$, with eigenvalues $\sigma_n^2 = 1/\lambda_n$ that represent the positional variance or fluctuations in the direction of these deformations.

2.3.2 Comparing properties of the structural core When comparing pairs of covariance matrices, they were first transformed to a common frame of reference in which the RMSD over the C_α positions in the structural core of the two protein domains is minimized.

2.4 (Dis)similarity measures

To calculate a score quantifying the (dis)similarity between pairs of proteins, we need a property to compare, described in the previous section, and a measure of (dis)similarity, discussed here.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

2.4.1 The root mean square deviation The similarity of native structure coordinates is quantified using the optimal RMSD over the aligned sites. RMSD is a dissimilarity measure that varies between 0 and ∞.

2.4.2 The Spearman correlation coefficient (ρ) The similarity of backbone fluctuation profiles is quantified using the Spearman correlation coefficient between the aligned RMSF profiles. ρ is a similarity measure that varies between -1 and 1.

2.4.3 The root mean square inner product The RMSIP is a measure of similarity of the N principal components of the covariance/correlation matrices

$$\mathrm{RMSIP} = \left[\frac{\sum_{n=1}^{N}\sum_{m=1}^{N}(\mathbf{U}_n \cdot \mathbf{V}_m)^2}{N}\right]^{1/2}, \tag{6}$$

where the columns of matrices $\mathbf{U}$ and $\mathbf{V}$ are the principal components of the covariance or correlation matrices compared. Following common practice, N is arbitrarily chosen to be 10. RMSIP is a similarity measure that varies between 0 and 1.

2.4.4 The root weighted square inner product The RWSIP between two sets of principal components, introduced by Carnevale *et al.* (2007), is as follows:

$$\mathrm{RWSIP} = \left[\frac{\sum_{n=1}^{N}\sum_{m=1}^{N} u_n v_m (\mathbf{U}_n \cdot \mathbf{V}_m)^2}{\sum_{n=1}^{N} u_n v_m}\right]^{1/2}, \tag{7}$$

where u_n and v_m are the eigenvalues of the covariance/correlation matrix corresponding to principal components $\mathbf{U}_n$ and $\mathbf{V}_m$, respectively. $N = |U| = |V|$, the number of non-trivial principal components in either set. RWSIP is a similarity measure that varies between 0 and 1.

2.4.5 The normalized Bhattacharyya distance ($\mathrm{ND_B}$) The Bhattacharyya distance (Bhattacharyya, 1943), D_B measures the overlap between two, distributions. Within the ENM model, the probability density function that characterizes the distribution of fluctuations around the equilibrium conformation is a multivariate normal distribution with covariance matrix $\mathbf{C}$ and mean $<\delta\mathbf{r}> = \mathbf{0}$. The D_B between two such distributions is given by

$$D_\mathrm{B} = \frac{1}{2}\ln\left[\frac{|\mathbf{D}|}{(|\mathbf{C}_\mathrm{A}||\mathbf{C}_\mathrm{B}|)^{\frac{1}{2}}}\right],$$

where $\mathrm{C_A}$ and $\mathrm{C_B}$ are the covariance matrices of the distributions compared, $\mathbf{D} = (\mathbf{C}_\mathrm{A} + \mathbf{C}_\mathrm{B})/2$ and $|\mathbf{X}|$ denotes the determinant of $\mathbf{X}$. D_B is a dissimilarity measure that varies between 0 and ∞.

Before calculating D_B we normalize the covariance matrices for the proteins to be compared, by dividing them by their trace: $C \rightarrow \mathbf{C}/tr(\mathbf{C})$.

The derivation of D_B assumes that the covariance matrices are positive-definite matrices. Since we have six trivial modes (eigenvectors with eigenvalue 0), we projected the covariance matrices onto a common lower-dimensional representation using

$$\tilde{\mathbf{C}} = \mathbf{Q}^T\mathbf{C}\mathbf{Q},$$

where $\mathbf{C}$ is $\mathbf{C}_\mathrm{A}$ or $\mathbf{C}_\mathrm{B}$, and the columns of $\mathbf{Q}$ are the s eigenvectors of $(\mathbf{C}_\mathrm{A} + \mathbf{C}_\mathrm{B})/2$ with highest variances. Here we choose s to be the lowest number that includes enough eigenvectors to explain 95% of the total variance. We define a rank-normalized version of D_B that eliminates the dependency on s:

$$\mathrm{ND_B} = \left[\frac{D_\mathrm{B}}{s}\right]^{1/2} \tag{8}$$

Note that, as the ENM is not parameterized to reproduce the fluctuations to scale, all the measures that compare covariance matrices or RMSF-profiles apply amplitude normalization as specified above, so that only relative amplitudes are compared.

2.5 Assessment

We assessed the performance of all the (dis)similarity scores studied, by evaluating their consistency with the SCOP classification at the superfamily level. SCOP is an expert classification-based mainly on visual inspection of protein structures (Murzin *et al.*, 1995). It is a hierarchical system with four levels: class, fold, superfamily and family. Fold-related proteins are structurally similar but not homologous, whereas superfamily and family-related proteins are, respectively, probably homologous and clearly homologous, as suggested by sequence or functional similarity.

To quantify the performance of a given (dis)similarity score, we calculate the proportion of cases for which it ranks domains in agreement with SCOP. Consider a triplet of protein domains (d_1, d_2 and d_3) with the first two being members of the same superfamily and the third of a different superfamily within the same fold. A given measure is consistent with SCOP for this triplet if the (dis)similarity between d_1 and d_2 is (smaller) larger than between d_1 and d_3. The number of correctly ranked triplets is the Mann–Whitney statistic comparing the distributions of (dis)similarity scores with respect to d_1 in the two superfamilies. Expressed as a proportion this statistic is the area under the curve (AUC) of a receiver-operating characteristic (ROC) curve obtained using the studied measure to classify domains according to their (dis)similarity with d_1 (Hanley and McNeil, 1982). Therefore, we will designate the consistency measure AUC:

$$\mathrm{AUC} = \frac{1}{|S||S'|}\sum_{d_2 \in S}\sum_{d_3 \in S'} H(m(d_1, d_2) - m(d_1, d_3)),$$

where S is the set of domains in a dataset that are in the same superfamily as d_1 (except d_1 itself), S' is the set of domains not belonging to the same superfamily as d_1, m is a similarity measure, $|X|$ denotes the cardinality of X and H evaluates to 1 if its argument is positive, 0 if it is negative and 1/2 if it is 0. For a dissimilarity measure the logic of H is reversed. We thus calculate an AUC for each domain, similar to how sequence matches are evaluated in Gribskov and Robinson (1996). This is referred to as an element-wise ranking scenario in Sonego *et al.* (2008). We estimate the consistency in a superfamily by averaging the AUC values for the domains in the superfamily. An overall consistency value is obtained by averaging over the superfamily-specific values.

The AUC values for each domain d_1 in a superfamily are approximately normally distributed, with means and variances derived from the distribution of the Mann–Whitney statistic (Mann and Whitney, 1947). The approximation to normality is further justified when we consider the averages of these nearly identically distributed AUC values. When calculating statistical significance, we therefore approximate the distribution of the average AUC values by a normal distribution with mean and variance estimated from the sample.

3 RESULTS

We performed a comparative assessment of different scores of (dis)similarity of protein fluctuation patterns. To this end, and

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

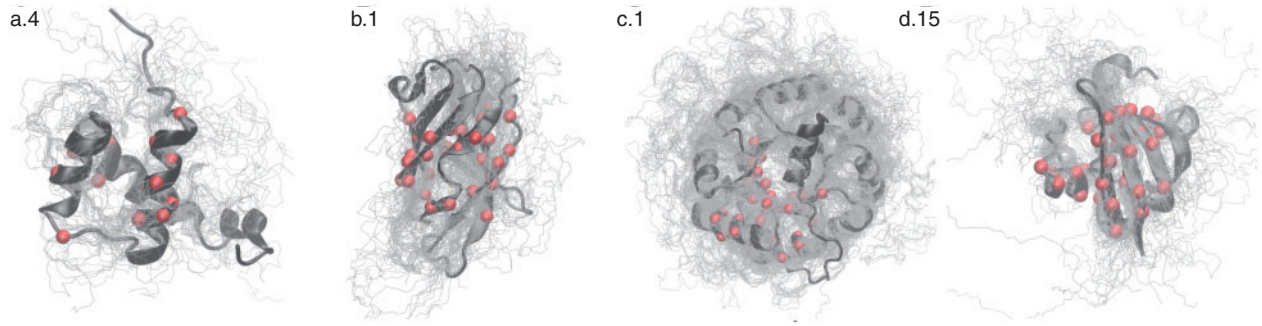

Fig. 1. Aligned structures. Superimposed protein domains for each dataset (a.4 DNA/RNA-binding 3-helical bundle, b.1 Immunoglobulin-like beta-sandwich, c.1 TIM beta/alpha-barrel, d.15 beta Grasp (ubiquitin-like). The backbone of each domain is shown with lines. One representative from each dataset is represented using cartoons and red spheres for C_α-atoms included in the structural core. Number of sites in alignment (smallest domain) is as follows: a.4: 14 (49), b.1: 30 (72), c.1: 33 (184) and d.15: 34 (73).

as described in Section 2, we

(1) chose four datasets, each of them including protein domains from 1-fold represented by two different SCOP superfamilies;
(2) obtained a multiple structural alignment of all proteins in each dataset from which we extract the conserved structural core;
(3) produced pairwise superimpositions of all proteins within a dataset;
(4) calculated properties that characterise the fluctuations of the aligned core of each protein;
(5) quantified the similarity of such properties with different (dis)similarity measures and
(6) calculated the consistency of each measure with the SCOP classification.

The aligned and superimposed structures of each of the four datasets are shown in Figure 1. STAMP captures well representative residues in the conserved core of the structures for almost all superfamilies. One exception is the a.4.5 superfamily that is classified by SCOP in the all-alpha class, although it has several β-sheets in the regions corresponding to unstructured loops in domains of a.4.1. This affects the arrangement of the adjacent α-helices, causing some counterintuitive matching of α-helices in a.4.1 with parts of loops and β-sheets in a.4.5, even if a central helix is consistently captured across the fold. The core for a.4 covers 29% of the smallest domain in the alignment. For the barrel structures (c.1), the obtained conserved core is skewed to one side of the characteristic barrel, but several spatially adjacent secondary structure elements are well captured by the alignment procedure.The core for c.1 covers only 18% of the smallest domain. The core for b.1 and d.15 covers in excess of 40% of the smallest domain. There can certainly be relevant dynamics in the more peripheral regions of the structures that can possibly be described by the ENM approach, but for comparing the conservation of (dis)similarity measures we regard this very restrictive definition of the conserved core as suited.

For each pair of proteins in each dataset, we calculated several similarity and dissimilarity scores. Thus, structural dissimilarity was quantified using the RMSD of the coordinates of aligned C_α-atoms in the native structure. The properties defined by Equations (1) through (5) were used to characterize the pattern of fluctuations. All these properties were obtained from the covariance matrix, which was calculated using a coarse-grained ENM. To quantify the (dis)similarity between the covariance and correlation matrices of different proteins, we used the three measures defined in Equations 6–8, while the similarity of RMSF profiles was quantified using the Spearman's correlation coefficient ρ. Each (dis)similarity score is a combination of a (dis)similarity measure and a property compared, so that the notation 'Measure(Property)' will be used where needed (e.g. RMSIP(**C**)). All (dis)similarity scores were calculated using only the C_α-atoms of sites that are aligned for all proteins of the same fold.

3.1 Conservation of the covariance matrix

First, we studied whether the covariance matrix, which is *a priori* the most informative characterization of the fluctuation pattern, is evolutionarily conserved. Figure 2 shows that RMSF profiles (left panel) and **C** (right panel) are more conserved for homologous (red and green bars) than for non-homologous proteins (black bars). For homologous proteins, they are more conserved at the family level (red bars) than at the superfamily level (green bars). Through analysing the distributions of scores using the Kolmogorov–Smirnov (KS) tests, we found that the differences are highly statistically significant for all the (dis)similarity measures (P-values $\ll 10^{-3}$), both when comparing the distribution of family-related domains with that of the superfamily-related ones and when comparing superfamily-related domains with fold-related domains. Visual inspection of Figure 2 shows that the dissimilarity measure $ND_B(\mathbf{C})$ outperforms the similarity measure ρ in discriminating between the different SCOP levels for all folds considered here. Thus, both RMSF profiles and covariance matrices can be used to characterize the divergence of protein fluctuation patterns, but covariance matrices do a significantly better job.

3.2 Comparison of different (dis)similarity scores

Next, we compared different fluctuation (dis)similarity scores. The covariance matrix adds to the RMSF profile information on the directionality of fluctuations and their covariation

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

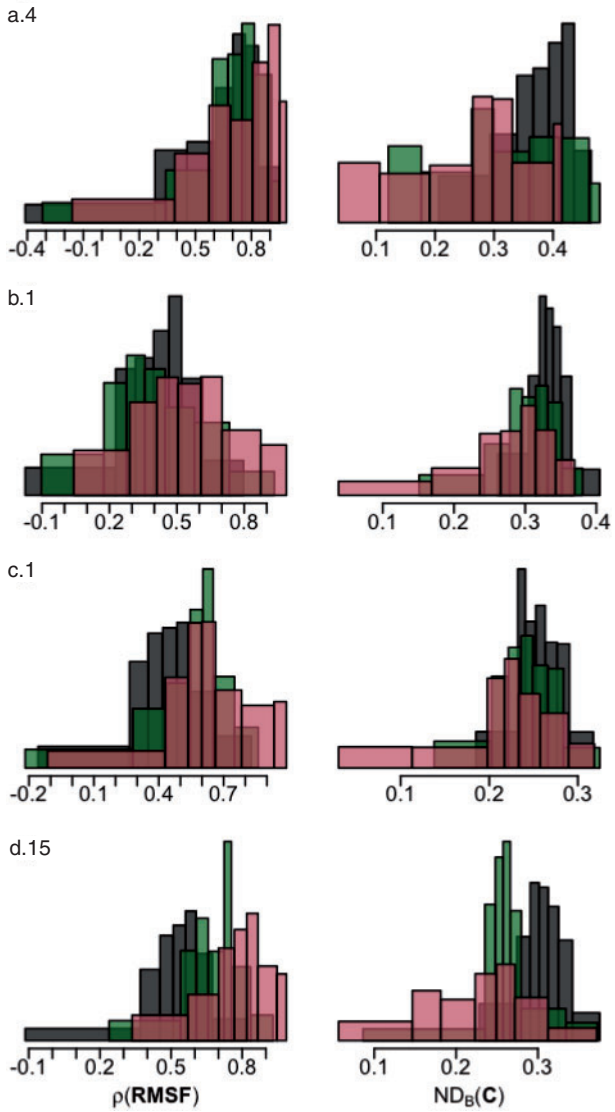

Fig. 2. Conservation of RMSF profiles and covariance matrices. Distributions of the similarity measure ρ(**RMSF**) and the dissimilarity measure ND_B(**C**) for all pairs of protein domains in the four datasets. The histograms are coloured according to the common SCOP classification in the distribution. Red histogram: family-related domains; green histogram: superfamily-related domains; black histogram: fold-related domains. Each bar contains the same number of domain comparisons. The height of the bars corresponds to the probability density, so that the histograms for each level of similarity have a total area of 1.

between sites. To disentangle the contributions of overall fluctuation magnitude from that of covariation, we compared covariance matrices with correlation, matrices. To study the effect of fluctuation directionality, we compared full covariance and correlation matrices with their isotropic counterparts. To compare the different measures, we evaluated their consistency with SCOP at the superfamily level as explained in Section 2.5. Consistency is given as AUC values ranging from 0 to 1, 1 being perfect agreement and 0.5 being the agreement expected for a random ranking procedure.

For comparing covariance and correlation matrices we apply several measures. The RMSIP is widely used for comparing the consistency of a chosen range of principal components. Ideally, the exact range should be motivated for each case, which is generally difficult when dealing with large datasets and more common than not an arbitrary choice of principal components is made. This choice is typically 10, as in this study [see for instance Amadei *et al.* (1999)]. The RWSIP uses a well-motivated weighting and normalization scheme to avoid the arbitrary choice of principal components. However, both these measures lack a rigorous foundation in statistics. For example in the somewhat contrived example of comparing any distribution with a distribution with exactly equal variance along each principal component, they have some unsettling properties. The RMSIP will not be defined in this case since the principal components cannot be ordered in a meaningful way. The RWSIP, on the other hand, will evaluate to exactly 1 which is supposed to represent identity. In contrast, the ND_B is well founded in the theory of multivariate statistics. As explained above, the complication of comparing distributions with no variance along the roto-translational components introduces some arbitrary choices of which principal components to retain, but the effect of this is supposedly much less severe than with the RMSIP as most of the information from the covariance matrix can be included. Looking past the application to elastic network models, although the formulation of the Bhattacharyya distance that is used here is adapted to compare Gaussian distributions, the more general formulation does not require that assumption. In principle, it can even be used to compare distributions that do not have a well-defined equilibrium, such as intrinsically disordered regions of a protein.

3.2.1 Overall performance Table 1 shows the overall performance of all the (dis)similarity scores. Consistency values were accumulated over all superfamilies except a.4.5 and c.1.2 for which fluctuation patterns seem to be only mildly conserved, or poorly captured, as will be shown below. As Table 1 accumulates AUC values obtained for each domain in six superfamilies, the differences between the measures are best analysed in terms of paired statistics. We therefore report also the mean difference between AUC values for the domains when either the similarity measure or the property compared is varied (Tables 2 and 3). From theoretical considerations, we expect ND_B(**C**) to perform best when assessing similarity of fluctuations, so we report differences to ND_B and **C** with significance values for a paired *t*-test with the null hypothesis that this difference is non-positive. We start by noticing that the assumption of ND_B consistently being best is not supported by our analysis, although it is the best for the anisotropic covariance matrices, which we preferred *a priori* (Table 2). For the isotropic covariances, the difference to RWSIP is negligible and for both the correlation matrices, the RWSIP somewhat surprisingly turns out to be the best. Interestingly, ND_B is generally better than the widely used RMSIP, again with the exception of the anisotropic correlation matrix where the difference is negligible. No single (dis)similarity measure is consistently the best, and the differences between RWSIP and ND_B are generally small. As can be seen in Tables 1 and 3, the property measured is more important for obtaining high consistency with SCOP than the measure used.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

Table 1. AUC values for (dis)similarity measures (rows) applied to properties (columns) accumulated over superfamilies with well-conserved fluctuation patterns (all except a.4.5 and c.2.1)

	r	RMSF	C	P	C^{iso}	P^{iso}
RMSD	0.86					
ρ		0.72				
ND_B			0.86	0.86	0.80	0.82
RWSIP			0.84	0.88	0.81	0.83
RMSIP			0.83	0.87	0.76	0.78

Table 2. Mean differences (p values) between AUC values for ND_B to other measures calculated with the same property

	RWSIP	RMSIP
C	0.021 ($<10^{-5}$)	0.035 ($<10^{-5}$)
P	−0.010 (0.99)	−0.002 (0.63)
$\mathbf{C}^{iso}$	−0.005 (0.93)	0.031 ($<10^{-3}$)
$\mathbf{P}^{iso}$	−0.008 (0.93)	0.036 ($<10^{-6}$)

p-values are calculated with a paired t-test under the null hypothesis that the difference is non-positive.

Table 3. Mean differences (p values) between AUC values for measures calculated with **C** to the same measure calculated with $\mathbf{C}^{iso}$ and ρ(**RMSF**)

	$\mathbf{C}^{iso}$	ρ(**RMSF**)
ND_B	0.064 (10^{-8})	0.137 (10^{-14})
RWSIP	0.038 (10^{-5})	0.116 (10^{-11})
RMSIP	0.060 (10^{-7})	0.102 (10^{-7})

p-values are calculated with a paired t-test under the null hypothesis that the difference is non-positive.

From Table 1, we see that the RMSF profile similarity ρ is much worse than the structure-based measure RMSD. In contrast, the other fluctuation-based measures largely outperform ρ, showing performances similar to or even higher than RMSD. Scores based on correlation matrices perform somewhat better than those based on covariance matrices. This suggests that the differences observed between comparisons of RMSF profiles and comparisons of covariance matrices are largely due to the introduction of the correlation between sites. Scores based on **C** and **P** perform better than those based on $\mathbf{C}^{iso}$ and $\mathbf{P}^{iso}$, respectively, which means that it is also important to consider the anisotropy of fluctuations.

In Table 3, we see that considering both inter-residue correlations and anisotropy improve the expected fraction of correctly ranked protein domains by 10–14%, with the addition of inter-residue correlations alone contributing 4–6% of this.

To summarize, the directionality of fluctuations and their correlation between sites have significant effects on overall performance.

3.2.2 Performance for different superfamilies To gain further insight into the different methods, we analysed the consistency of the different scores with SCOP for different superfamilies (Table 4). In agreement with the previous section, ρ is much less consistent with SCOP than the structural dissimilarity RMSD for all superfamilies except a.4.5 and c.1.8. The other fluctuation-based scores, in contrast, are rivalling structural similarity in their agreement with SCOP. For superfamily c.1.8, the best fluctuation-based scores show a remarkable improvement over RMSD.

Further inspection of Table 4 shows that there is no single score that clearly outperforms all the others for all superfamilies. In general, using **C** and **P** is better than using their isotropic counterparts, and the correlation matrices perform slightly better than the covariances, at least in the case of RWSIP. However, there are exceptions. For a.4.5, for example, the best scores are those that do not take into account any cross-correlations (RMSD and ρ). This is in agreement with the observation that some domains in this superfamily has been counter-intuitively aligned, with the effect that the similarity of the central helix is reflected in RMSD and ρ, while this local similarity is not reflected when the cross-correlations with the misaligned environment is considered, leading to consistencies with SCOP of the order of 50%, which is what would be obtained randomly. Another exception is c.1.2 for which both the *RMSF* profile and scores that include the anisotropy of fluctuations perform very poorly, but significant results are obtained using the isotropic correlation matrix. Here again, the alignment produced by STAMP might not be optimal, although significantly better than for a.4.5. As this alignment is localized to only a part of the barrel it is plausible that the frame of reference defined by minimising the RMSD over aligned sites is suboptimal, possibly explaining that the rotational invariant properties (C^{iso} and P^{iso}) in this case perform consistently better than their anisotropic counterparts. Interestingly, reviewing the enzyme commission classifications (http://www.chem.qmul.ac.uk/iubmb/enzyme/) of the structures compared in c.1 reveals that our c.1.2 selection is more heterogeneous in terms of enzyme function than our c.1.8 selection. This can possibly explain both the low consistencies observed in c.1.2 and the very high consistencies observed in c.1.8.

The (dis)similarity measure used (ND_B, RWSIP or RMSIP) has a smaller effect than the matrices compared and depends on them, which makes it difficult to establish a general trend. However, it can be seen that ND_B and RWSIP are somewhat better than RMSIP. Keeping in mind these difficulties to generalize, we can say that the best fluctuation-based scores are those that take into account fluctuation anisotropy and correlation between sites. These scores have performances rivalling that of the RMSD, despite the fact that the analysis is slightly biased in favour of the RMSD, both by the superimposition procedure and the structural alignment procedure.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

Table 4. Superfamily-specific AUC values for measures (rows) applied to properties (columns)

	a.4.1						a.4.5					
	r	RMSF	C	P	C^{iso}	P^{iso}	r	RMSF	C	P	C^{iso}	P^{iso}
RMSD	0.96						0.65					
ρ		0.59						0.65				
ND_B			0.89	0.88	0.87	0.89			0.48	0.49	0.49	0.45
RWSIP			0.90	0.89	0.90	0.91			0.39	0.47	0.44	0.41
RMSIP			0.88	0.90	0.70	0.79			0.49	0.52	0.65	0.38
	b.1.1						b.1.18					
	r	RMSF	C	P	C^{iso}	P^{iso}	r	RMSF	C	P	C^{iso}	P^{iso}
RMSD	0.91						0.68					
ρ		0.58						0.61				
ND_B			0.92	0.91	0.78	0.89			0.67	0.72	0.56	0.66
RWSIP			0.90	0.93	0.78	0.85			0.62	0.68	0.55	0.66
RMSIP			0.92	0.93	0.84	0.90			0.67	0.68	0.56	0.56
	c.1.2						c.1.8					
	r	RMSF	C	P	C^{iso}	P^{iso}	r	RMSF	C	P	C^{iso}	P^{iso}
RMSD	0.81						0.79					
ρ		0.55						0.85				
ND_B			0.51	0.51	0.61	0.69			0.90	0.91	0.89	0.93
RWSIP			0.52	0.59	0.61	0.71			0.93	0.90	0.93	0.91
RMSIP			0.44	0.59	0.55	0.79			0.94	0.88	0.92	0.91
	d.15.1						d.15.4					
	r	RMSF	C	P	C^{iso}	P^{iso}	r	RMSF	C	P	C^{iso}	P^{iso}
RMSD	0.86						0.95					
ρ		0.84						0.87				
ND_B			0.82	0.81	0.85	0.76			0.98	0.96	0.85	0.79
RWSIP			0.79	0.87	0.81	0.83			0.92	0.98	0.86	0.82
RMSIP			0.70	0.85	0.76	0.71			0.85	0.96	0.81	0.83

Except for some entries in a.4.5 and c.1.2, all results are highly statistically significant (P-values $\ll 10^{-3}$) as assessed with a t-test against the null hypothesis that the mean AUC is ≤ 0.5.

3.3 Complementarity to *RMSD*

Finally, we investigated the relationship between fluctuation-based scores and *RMSD*. Figure 3 shows $ND_B(\mathbf{C})$ versus RMSD for the 4-folds considered here. It can be seen that the two scores are correlated. The mean Spearman's correlation coefficient over all datasets is $r = 0.76$. Even though this correlation is significant, there is about 42% of the variance of $ND_B(\mathbf{C})$ rankings that is not accounted for by variations in RMSD rankings. If the unaccounted variance was just noise, we would expect the performance of $ND_B(\mathbf{C})$ to be much lower than that of RMSD. This, however, is not the case. Inspection of Tables 1 and 4 shows that $ND_B(\mathbf{C})$ is as consistent with SCOP as RMSD or even more. The analysis yields similar results for other (dis)similarity measures, which demonstrates that the performance of fluctuation-based scores is not just a trivial consequence of their correlation with RMSD. In fact, Table 4 shows that in half of the superfamilies a fluctuation-based score performs better than RMSD. Moreover, a detailed analysis shows that for 10% of the cases for which the $ND_B(\mathbf{C})$ ranking agrees with SCOP, the RMSD ranking does not. Therefore, even though fluctuation-based scores and RMSD are correlated, they contain complementary information.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

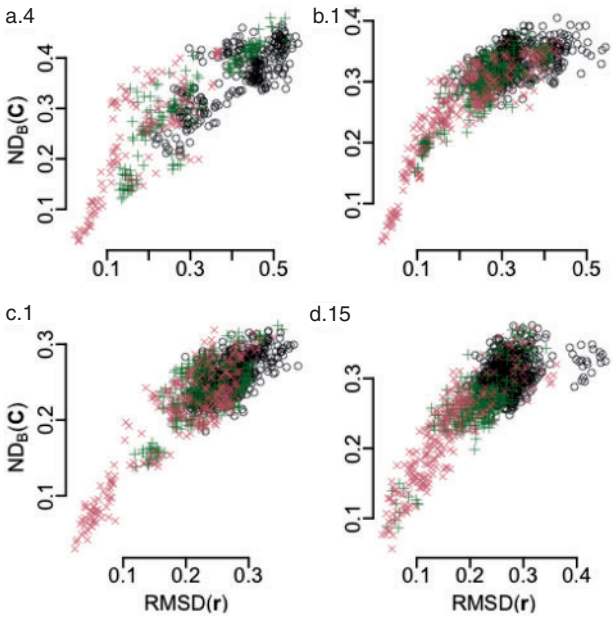

Fig. 3. Complementarity between $ND_B(\mathbf{C})$ and RMSD. Bivariate distribution of RMSD (in nanometers) and $ND_B(\mathbf{C})$ for all pairs of protein domains in each of the four data sets. The symbols used for each pair of domains are coloured according to their common SCOP classification. Red ×: family-related domains; green +: superfamily-related domains; black O: fold-related domains.

4 CONCLUSIONS

If we quantify the similarity of protein dynamics using ρ(RMSF) of the structural core, we find that it is less conserved than structure (Tables 1 and 4). This has been previously discussed in studies using ρ(RMSF) to investigate the evolution of protein dynamics (Maguid *et al.*, 2006, 2008). However, better fluctuation-based (dis)similarity scores can be used, which perform as well as or even better than the structure-based RMSD. Using the covariance matrix results in much better separation of homologous and non-homologous proteins than using RMSF profiles (Fig. 2). The same can also be seen more quantitatively when measuring the consistency between (dis)similarity measures and SCOP. The scores accounting for anisotropy and correlations between residues (comparing **C** and **P** matrices) have much better performances than the comparison of RMSF profiles, even rivalling that obtained with RMSD (Table 1).

Moreover, the good performance of such fluctuation-based measures is not a trivial consequence of structural similarity. Even though there is a high correlation between the fluctuation-based measures and RMSD, there are many cases in which a chosen fluctuation-based measure agrees with SCOP, when RMSD does not. The reverse is also true, supporting the idea that fluctuation-based measures and structure-based measures are complementary, in agreement with previous studies (Maguid *et al.*, 2006, 2008; Pandini *et al.*, 2007; Zen *et al.*, 2008).

The key to a good fluctuation-based (dis)similarity score is to take into consideration the correlation of fluctuations between sites and their directionality. This is in agreement with work in which dynamical similarity measures are based on the conservation of the lowest normal modes, which are the principal components of the covariance matrix (Maguid *et al.*, 2005, 2008; Zen *et al.*, 2008). In addition to expressing a property of the fluctuations that takes into account cross-correlations and directionality, the measure of (dis)similarity used (ND_B, *RWSIP* or *RMSIP*) also has an effect, but this is smaller and less consistent.

We also learn from the folds for which we found the alignment to be counter-intuitive or not capturing representative residues from all parts of the core, that the measures that quantify similarity of cross-correlations are sensitive to the quality of the alignments. Measures that take into account the directionality of fluctuations will also necessarily be dependent on a good definition of a common frame of reference for compared domains.

To summarize, the best scores are those based on properties that inform on the anisotropy of fluctuations and their correlation between sites. The similarity score most used in evolutionary studies of conservation of protein fluctuations has so far been ρ(RMSF) (Maguid *et al.*, 2006, 2008; Pandini *et al.*, 2007), followed by RMSIP(**C**) (e.g. Zen *et al.* (2008)). From the present study it follows that better scores are $ND_B(\mathbf{C})$ and RWSIP(**P**). In particular, ND_B is founded in the theory of multivariate statistics and works as well on the better motivated covariance matrices as on the correlation matrices. In their consistency with the SCOP classification, these measures are as good as, or even better than RMSD. This supports and complements previous reports on the evolutionary conservation of protein fluctuations (Maguid *et al.*, 2006, 2008; Pandini *et al.*, 2007; Zen *et al.*, 2008).

Although the use of ρ(RMSF) would indicate that protein structural fluctuations are less conserved than the native structure, the strategy we propose demonstrates a higher conservation of protein dynamics, notably by taking into account the covariance or correlation of movements within domains.

ACKNOWLEDGEMENT

J.E. is a researcher of CONICET. Parallab (High Performance Computing Laboratory at the University of Bergen) is thankfully acknowledged for provision of CPU time to this project.

Funding: This work was supported by grants from Agencia Nacional de Promoción Científica y Tecnológica (PICT 1459) to J.E.; and the Bergen Research Foundation to N.R.

Conflict of Interest: none declared.

REFERENCES

Ahmed,A. *et al.* (2010) Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins*, **78**, 3341–3352.

Amadei,A. *et al.* (1999) On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins*, **36**, 419–424.

Bahar,I. *et al.* (2010) Global dynamics of proteins: bridging between structure and function. *Ann. Rev. Biophys.*, **39**, 23–42.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhattacharyya,A. (1943) On a measure of divergence between two multinomial populations. *Sankhya: Indian J. Statist.*, **7**, 401–406.

Carnevale,V. *et al.* (2006) Convergent dynamics in the protease enzymatic superfamily. *J. Am. Chem. Soc.*, **128**, 9766–9772.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012

Carnevale,V. *et al.* (2007) Structural and dynamical alignment of enzymes with partial structural similarity. In *Journal of Physics-Condensed Matter*, SISSA, CNR, INFM Democritos, I-34014 Trieste, Italy.

Chandonia,J.-M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Echave,J. and Fernández,F.M. (2010) A perturbative view of protein structural variation. *Proteins*, **78**, 173–180.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Henzler-Wildman,K. and Kern,D. (2007) Dynamic personalities of proteins. *Nature*, **450**, 964–972.

Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.

Hinsen,K. *et al.* (2000) Harmonicity in slow protein dynamics. *Chem. Phys.*, **261**, 25–37.

Hollup,S.M. *et al.* (2011) Exploring the factors determining the dynamics of different protein folds. *Protein Science*, **20**, 197–209.

Keskin,O. *et al.* (2000) Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.*, **78**, 2093–2106.

Leo-Macias,A. *et al.* (2005) An analysis of core deformations in protein superfamilies. *Biophys. J.*, **88**, 1291–1299.

Lo Conte,L. *et al.* (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.

Maguid,S. *et al.* (2005) Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophys. J.*, **89**, 3–13.

Maguid,S. *et al.* (2006) Evolutionary conservation of protein backbone flexibility. *J. Mol. Evol.*, **63**, 448–457.

Maguid,S. *et al.* (2008) Evolutionary conservation of protein vibrational dynamics. *Gene*, **422**, 7–13.

Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statis.*, **18**, 50–60.

Mitternacht,S. and Berezovsky,I.N. (2011a) Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput. Biol.*, **7**, e1002148.

Mitternacht,S. and Berezovsky,I.N. (2011b) Coherent conformational degrees of freedom as a structural basis for allosteric communication. *PLoS Comput. Biol.*, **7**, e1002301.

Münz,M. *et al.* (2010) Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. *BMC Bioinformatics*, **11**, 188.

Murzin,A.G. *et al.* (1995) SCOP—a structural classification of proteins. Database for the investigation of sequences and structures. *J. Mole. Biol.*, **247**, 536–540.

Pandini,A. *et al.* (2007) Detecting similarities among distant homologous proteins by comparison of domain flexibilities. *Prot. Eng. Design Select.*, **20**, 285–299.

Pang,A. *et al.* (2005) Comparative molecular dynamics—similar folds and similar motions? *Proteins*, **61**, 809–822.

Papaleo,E. *et al.* (2006) Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family. *BBA-Prot. Proteom.*, **1764**, 1397–1406.

Potestio,R. *et al.* (2010) ALADYN: a web server for aligning proteins by matching their large-scale motion. *Nucleic Acids Res.*, **38**(Web Server), W41–W45.

Raimondi,F. *et al.* (2010) Deciphering the deformation modes associated with function retention and specialization in members of the Ras superfamily. *Structure*, **18**, 402–414.

Rueda,M. *et al.* (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**, 565–575.

Russell,R. and Barton,G. (1992) Multiple protein-sequence alignment from tertiary structure comparison—assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.

Skjaerven,L. *et al.* (2011) Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins*, **79**, 232–243.

Sonego,P. *et al.* (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinformatics*, **9**, 198–209.

Yang,L. *et al.* (2008) Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure*, **16**, 321–330.

Zen,A. *et al.* (2008) Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. *Prot. Sci.*, **17**, 918–929.

Zen,A. *et al.* (2009) Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. *Bioinformatics*, **25**, 1876–1883.

Downloaded from http://bioinformatics.oxfordjournals.org/ at FundaÃ§Ã£o CoordenaÃ§Ã£o de AperfeiÃ§oamento de Pessoal de NÃ-vel Superior on October 14, 2012